

EISSN 1305-3612

DIR

Diagnostic and Interventional
Radiology

Editor in Chief

 Şükrü Mehmet Ertürk, MD

Department of Radiology, Istanbul University Faculty of Medicine, Istanbul, Türkiye

E-mail: smerturk@gmail.com

Section Editors and Scientific Editorial Board

Abdominal Imaging

 Sonay Aydın, MD

Department of Radiology, Erzincan Binali Yıldırım University Mengücek Gazi Training and Research Hospital, Erzincan, Türkiye

E-mail: sonay.aydin@erzincan.edu.tr

 Nurullah Dağ, MD

Department of Radiology, İnönü University Faculty of Medicine, Malatya, Türkiye

E-mail: nurullah.dag@inonu.edu.tr,
drndag@icloud.com

 Oğuz Dicle, MD

Department of Radiology, Dokuz Eylül University Faculty of Medicine, İzmir, Türkiye

E-mail: odicle59@gmail.com

 İlkey S. İdilman, MD

Department of Radiology, Hacettepe University Faculty of Medicine, Ankara, Türkiye

E-mail: ipolater@yahoo.com

Artificial Intelligence and Informatics

 Tuğba Akıncı D'Antonoli, MD

Division of Diagnostic and Interventional Neuroradiology, Department of Radiology, University Hospital Basel, Basel, Switzerland

Department of Pediatric Radiology, University Children's Hospital Basel, Basel, Switzerland

E-mail: tugba.akincidantonoli@unibas.ch

 Ulaş Bağcı, PhD

Department of Radiology, Northwestern University Faculty of Medicine, Chicago, United States

E-mail: ulas.bagci@northwestern.edu

 Görkem Durak, MD

Department of Radiology, Northwestern University Faculty of Medicine, Chicago, United States

E-mail: gorkem.durak@northwestern.edu

 İsmail Meşe, MD

Department of Radiology, Üsküdar State Hospital, Istanbul, Türkiye

E-mail: ismail_mese@yahoo.com

Breast Imaging

 Işıl Başara Akın, MD

Department of Radiology, Dokuz Eylül University Faculty of Medicine, İzmir, Türkiye

E-mail: isilbasara@gmail.com

 Serap Gültekin, MD

Department of Radiology, Gazi University Faculty of Medicine, Ankara, Türkiye

E-mail: sergultekin@yahoo.com

Chest and Cardiovascular Imaging

 Serçin Özkök, MD, PhD

Department of Radiology, University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Istanbul, Türkiye

E-mail: sercinbas2005@gmail.com

 Gamze Durhan, MD

Department of Radiology, Hacettepe University Faculty of Medicine, Ankara, Türkiye

E-mail: gamzedurhan@gmail.com

 Furkan Ufuk, MD

Department of Radiology, The University of Chicago, Chicago, United States of America

E-mail: furkan.ufuk@hotmail.com

Emergency Radiology

 Francesca Iacobellis, MD

Department of General and Emergency Radiology, Antonio Cardarelli Hospital, Naples, Italy

E-mail: iacobellisf@gmail.com

Hybrid Imaging and Nuclear Medicine

 Zeynep Gözde Özkan, MD

Department of Internal Medicine, Istanbul University Faculty of Medicine, Division of Nuclear Medicine, Istanbul, Türkiye

E-mail: zeynep.ozkan@istanbul.edu.tr,
zgodeozkan@yahoo.com

Interventional Radiology

 Barbaros Çil, MD, FCIRSE

Department of Radiology, Koç University School of Medicine, Istanbul, Türkiye

E-mail: bcil@ku.edu.tr

 Osman Öcal, MD

Department of Diagnostic and Interventional Radiology, Heidelberg University Hospital, Heidelberg, Germany

E-mail: osman.oecal@med.uni-heidelberg.de

 Onur Taydaş, MD

Department of Radiology, Sakarya University Faculty of Medicine, Sakarya, Türkiye

E-mail: onurtaydas@sakarya.edu.tr,
taydasonur@gmail.com

 Bahri Üstünsöz, MD

Department of Radiology, LSUHSC (Louisiana State University Health Science Center) School of Medicine, New Orleans, United States

E-mail: bustunsoz2000@yahoo.com

Musculoskeletal Imaging

 Zeynep Maraş Özdemir, MD

Department of Radiology, Güven Hospital, Ankara, Türkiye

E-mail: zynpmaras@yahoo.com

 Hayri Oğul, MD

Department of Radiology, Istanbul Medipol University Faculty of Medicine, Istanbul, Türkiye

E-mail: drhogul@gmail.com

 Adalet Elçin Yıldız, MD

Department of Radiology, Hacettepe University Faculty of Medicine, Ankara, Türkiye

E-mail: aelcindr@gmail.com

Neuroradiology

 Gökçen Çoban Çiftçi, MD

Department of Radiology, Hacettepe University Faculty of Medicine, Ankara, Türkiye

E-mail: drgokcencoban@gmail.com

 Erkan Gökçe, MD

Department of Radiology, Tokat Gaziosmanpaşa University Faculty of Medicine, Tokat, Türkiye

E-mail: drerkangokce@gmail.com

 Elif Günay Bulut, MD

Department of Radiology/Neuroradiology, Hacettepe University Faculty of Medicine, Ankara, Türkiye

E-mail: drelifbulut@yahoo.com

 Suzan Şaylısoy, MD

Department of Radiology, Eskişehir Osmangazi University Faculty of Medicine, Eskişehir, Türkiye

E-mail: ssaylisoy@ogu.edu.tr

Oncologic and General Radiology

 Afak Durur Karakaya, MD, PhD

Department of Radiology, Koç University Faculty of Medicine, Istanbul, Türkiye

E-mail: afakdurur@yahoo.com

 İnci Kızıldağ Yırgın, MD

Department of Radiology, Istanbul University Faculty of Medicine, Istanbul, Türkiye

E-mail: inci.kizildag@gmail.com

Pediatric Radiology

 Evrim Özmen, MD

Department of Radiology, Koç University Hospital, Istanbul, Türkiye

E-mail: evrimkilicdr@gmail.com

Ultrasound and Interventional Radiology

 Halit Nahit Şendur, MD

Department of Radiology, Gazi University Faculty of
Medicine, Ankara, Türkiye

E-mail: hsendur@gmail.com

Editorial Consultant to Editor-in-Chief

 Mehmet Ruhi Onur, MD

Department of Radiology, Hacettepe University Faculty
of Medicine, Ankara, Türkiye

E-mail: ruhionur@yahoo.com

Publication Coordinator

 Nermin Tunçbilek, MD

Department of Radiology, Trakya University Faculty of
Medicine, Edirne, Türkiye

E-mail: drtuncbilek@hotmail.com

Biostatistical Consultant

 İlker Ercan, PhD

Department of Biostatistics, Uludağ University School
of Medicine, Bursa, Türkiye

E-mail: iercan@msn.com

Social Media Leads

Sıtkı Safa Taflan, MD

Istanbul University Faculty of Medicine, Department of
Radiology, Istanbul, Türkiye

Sıla Atacan, MD

Istanbul University Faculty of Medicine, Department of
Radiology, Istanbul, Türkiye

Gursel Karacaoglu, MD

Istanbul Medeniyet University, Department of
Radiology, Istanbul, Türkiye

Mahmut Bilal Doğan, MD

Marmara University Faculty of Medicine, Department
of Radiology, Istanbul, Türkiye

Publication Services

Galenos Publishing, İstanbul, Türkiye

Past Editors

Editors in Chief

Mehmet Ruhi Onur, MD (2023-2026)

Mustafa Seçil, MD (2016-2023)

Nevzat Karabulut, MD (2011-2016)

Üstün Aydingöz, MD (2010-2011)

Okan Akhan, MD (2001-2010)

Ferhun Balkancı, MD (1999-2001)

Aytekin Besim, MD (1994-1999)*

* Dr. Aytekin Besim actually served as the
General Coordinator. His work in this capacity,
however, was in effect that of an Editor in Chief.

Editors

Ayşenur Cila, MD (2001-2002)

Suat Kemal Aytac, MD (1997-2001)

Erhan Ilgıt, MD (1994-2001)

Okan Akhan, MD (1994-2001)

Ferhun Balkancı, MD (1994-2000)

Serdar Akyar, MD (1994-1997)

Section Editors

Section Editorship was established in 2002 at
the tenure of Dr. Okan Akhan, Editor in Chief.

Abdominal Imaging

Bengi Gürses, MD (2020-2023)

Mehmet Ruhi Onur, MD (2016-2023)

Barış Türkbey, MD (2014-2020)

Mustafa N. Özmen, MD (2012-2018)

Murat Acar, MD (2015-2016)

Mustafa Seçil, MD (2011-2016)

Ahmet Tuncay Turgut, MD (2011)

Deniz Akata, MD (2007-2011)

Ayşe Erden, MD (2002-2011)

Okan Akhan, MD (2002-2010)

Hakan Özdemir, MD (2002-2010)

Artificial Intelligence and Informatics

Burak Koçak, MD (2023-2026)

Barış Türkbey, MD (2020-2023)

Breast Imaging

Mustafa Erkin Arıbal, MD (2016-2023)

Sibel Kul, MD (2015-2018)

Ayşenur Oktay, MD (2009-2014)

Ayşegül Özdemir, MD (2004-2009)

Cardiovascular Imaging

Uğur Bozlar, MD (2016-2023)

Muşturay Karçaaltıncaba, MD (2007-2010)

Mecit Kantarcı, MD (2010-2016)

Chest Imaging

Nevzat Karabulut, MD (2010-2014)

Çetin Atasoy, MD (2007-2010)

Macit Arıyürek, MD (2002-2007)

Figen Demirkazık, MD (2014-2018)

General Radiology

Ersin Öztürk, MD (2014-2017)

Utku Şenol, MD (2010-2013)

Oğuz Dicle, MD (2007-2010)

Interventional Radiology

James Milburn, MD (2023-2026)

Cüneyt Aytekin, MD (2016-2023)

Bora Peynircioğlu, MD (2012-2015)

Levent Oğuzkurt, MD (2011-2014)

Fatih Boyvat, MD (2007-2010)

İsmail Oran, MD (2015-2019)

Musculoskeletal Imaging

Hatice Tuba Sanal, MD (2016-2023)

Fatih Kantarcı, MD (2014-2016)

Ayşenur Oktay, MD (2011-2013)

Üstün Aydingöz, MD (2002-2011)

Berna Dirim Mete (2016-2017)

Neuroradiology and Head & Neck Imaging

Gülğün Yılmaz Ovalı, MD (2023-2026)

Kubilay Aydın, MD (2016-2023)

Nafi Aygün, MD (2016-2023)

Kader Karlı Oğuz, MD (2011-2015)

Süleyman Men, MD (2007-2013)

Muhteşem Ağıldere, MD (2002-2011)

Nuclear Medicine

Evrin Bengi Türkbey, MD (2023-2026)

A. Cahid Civelek, MD (2016-2023)

Oktay Sarı, MD (2015)

Akın Yıldız, MD (2011-2014)

Pediatric Radiology

Meltem Ceyhan Bilgici, MD (2023-2026)

Korgün Koral, MD (2016-2023)

Murat Kocaoğlu, MD (2016-2023)

Ensar Yekeler, MD (2014-2016)

Suat Fitöz, MD (2007-2013)

Diagnostic and Interventional Radiology (Diagn Interv Radiol) is a bimonthly periodical of the Turkish Society of Radiology and the content of the journal is available at <https://www.dirjournal.org/>. It is peer-reviewed and adheres to the highest ethical and editorial standards. The editors of the journal endorse the Editorial Policy Statements Approved by the Council of Science Editors Board of Directors (<https://cse.memberclicks.net/>). The journal is in compliance with the Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals published by the International Committee of Medical Journal Editors (updated May 2022, www.icmje.org).

First ten volumes of Diagnostic and Interventional Radiology have been published in Turkish under the name of Tanısal ve Girişimsel Radyoloji (MEDLINE abbreviation: Tani Girişim Radyol), the current title's exact Turkish translation.

Diagnostic and Interventional Radiology is an open access publication, and the journal's publication model is based on Budapest Open Access Initiative (BOAI) declaration. All published content is available online, free of charge at <https://www.dirjournal.org/>. Authors retain the copyright of their published work in Diagnostic and Interventional Radiology. The journal's content is licensed under a Creative Commons Attribution-NonCommercial (CC BY-NC) 4.0 International License which permits third parties to share and adapt the content for non-commercial purposes by giving the appropriate credit to the original work.

Please refer to the journal's webpage (<https://dirjournal.org/>) for "Aims and Scope", "Instructions to Authors" and "Instructions to Reviewers".

The editorial and publication processes of the journal are shaped in accordance with the guidelines of the ICMJE, WAME, CSE, COPE, EASE, and NISO.

Diagnostic and Interventional Radiology is indexed in **SCI-Expanded, Pubmed/Medline, Pubmed Central, TUBITAK ULAKBİM TR Index, Scopus, DOAJ, EBSCO, HINARI, EMBASE, CINAHL, ProQuest, Gale and CNKI**.

The journal is published online.

Owner: Turkish Society of Radiology

Responsible Manager: Şükrü Mehmet Ertürk

Contact Information

Diagnostic and Interventional Radiology Turkish Society of Radiology

Hoşdere Cad., Güzelkent Sok., Çankaya Evleri, F/2, 06540

Ankara, Türkiye

E-mail: info@dirjournal.org

Phone: +90 (312) 442 36 53 **Fax:** +90 (312) 442 36 54

Publisher Contact

Address: Molla Gürani Mah. Kaçamak Sk.

No: 21/1 34093 İstanbul, Türkiye

Phone: +90 (530) 177 30 97

E-mail: info@galenos.com.tr/yayin@galenos.com.tr

Web: www.galenos.com.tr **Publisher Certificate Number:** 14521

Online Publication Date: September 2026

EISSN 1305-3612

International scientific journal published bimonthly.



Contents

ABDOMINAL IMAGING

521 Original Article Dual-layer spectral detector computed tomography of peritumoral fat for preoperative prediction of tumor budding in colorectal adenocarcinoma. *Jian-Bo Xu, Wei-Jin Zhong, Yuan-Hui Li, Xiang Zhou, Su-E Cao, Xiao-Min Liu, Ruo-Mi Guo*

531 Letter to the Editor Letter to the Editor: comments on the diagnostic value of ADC texture analysis in PI-RADS 5 lesions. *Esat Kaba*

533 Letter to the Editor Reply: comments on the diagnostic value of ADC texture analysis in PI-RADS 5 lesions. *Yan Bai, Chen Jiang Wu*

ARTIFICIAL INTELLIGENCE AND INFORMATICS

535 Review Human–AI interaction and collaboration in radiology: from conceptual frameworks to responsible implementation. *Burak Kocak, Renato Cuocolo*

548 Original Article Artificial intelligence in radiology examinations: a psychometric comparison of question generation methods. *Emre Emekli, Betül Nalan Karahan*

555 Original Article Diagnostic accuracy and consistency of ChatGPT-4o in radiology: influence of image, clinical data, and answer options on performance. *Kadir Atakır, Kaan Işın, Abdullah Taş, Hakan Önder*

565 Original Article Potential radiation dose reduction in computed tomography with deep learning reconstruction: a retrospective monocentric study. *Lucas Graber, Melike Zeynep Akış, François Séverac, Luc Mertz, Serhat Akış, Catherine Roy, Mickaël Ohana*

574 Original Article Multi-parametric magnetic resonance imaging-based radiomics for differentiation of skull base osteomyelitis from locally advanced nasopharyngeal carcinoma: a multi-center external validation study. *Fırat Atak, Hanife Avcı, Yeliz Pekçevik, Ayça Karaosmanoğlu*

585 Original Article Large-scale evaluation of multimodal large language models for pneumothorax detection. *Hamza Eren Güzel, Cemre Özenbaş, Ali Murat Koç*

591 Original Article Prospective quantitative analysis of hyperparameter and input optimization in GPT-5: comparative contribution to radiologist performance in abdominal radiology. *Eren Çamur, Turay Cesur, Yasin Celal Güneş, Muhammed Batuhan Gökhan, Rıza Sarper Ökten*

600 Original Article Keros-Net: a convolutional block attention module–squeeze-and-excitation-integrated hybrid learning framework for olfactory fossa depth classification. *Zülküf Akdemir, Murat Canayaz, Abdulaziz Yalınkılıç, Suat Orak*

608 Original Article Performance of multimodal large language models for the detection and characterization of bone lesions on radiographs. *Raşit Eren Büyüktoka, Ali Salbas, Atilla Hikmet Cilengir, Asli Dilara Buyuktoka, Murat Sürücü, Zehra Hilal Adibelli*

616 Invited Commentary Quality of reporting of studies on artificial intelligence in radiology: what is the current state of the field? *Ahmed Maiter-*

BREAST IMAGING

618 Original Article Radiomics-based differentiation of benign and malignant breast masses on contrast-enhanced mammography: a reproducible workflow using open-source tools. *Aykut Teymur, Sibel Kul, Ramazan Özgür Doğan, Hayati Türe, Temel Kayıkçıoğlu, Onur Bektaş, Sinan Erdem, Cemal Aydoğan*

CARDIOVASCULAR IMAGING

626 Invited Review A scoping review of photon-counting detector computed tomography in cardiovascular imaging. *Deniz Alis, Mehmet Onur Önal, Müjgan Orman, Mustafa Ege Seker, Ahmet Akyol, Cem Alhan, Ercan Karaarslan*

CHEST IMAGING

643 Pictorial Essay Radiologic patterns of interstitial lung disease: a pictorial review based on the 2025 European Respiratory Society/American Thoracic Society Classification. *Furkan Ufuk*

655 Original Article The relationship between preoperative pulmonary hypertension risk in liver transplant and vascular and biliary complications in the postoperative period. *Burak Yağdıran, Can Ugochkwu Uchehara, Adem Şafak, Nedim Çekmen, Mehmet Haberal*

GENERAL RADIOLOGY

662 Original Article Transperineal ultrasonography with a high-frequency probe and a three-dimensional end-fire probe for imaging perianal abscesses and locating the internal openings of anal fistulas. *Hongyun Wu, Juan Chang, Xiuhong Zhang, Yunfei Tao*

Contents

669 Original Article Microvascular dysfunction in patients with transthyretin cardiac amyloidosis evaluated by ¹³N-ammonia positron emission tomography–computed tomography: is it an early marker of the disease? *Aristóteles Neto, Caio Cafezeiro, Bruno Bueno, Cristhian Espinoza Romero, Vitor Emer Rosa, Viviane Hotta, Camila de Godoi Carneiro, Carlos Rochitte, Vagner Madrini, Jose Soares, William Chalela, Marcos Lima, Carlos Buchpiguel, Edward Miller, Fabio Fernandes*

INTERVENTIONAL RADIOLOGY

675 Review Ultrasound-guided microwave ablation for breast tumors: current status and future perspectives. *Jin Wang, Hongjuan Wei, Liming Zhou, Xin Chang, Jing Zhao, Hui Chen*

692 Original Article Establishment of an inferior vena cava filter retrieval program: the effect on trauma and non-trauma patient populations. *Katelyn Gill, Hector Ferral, Alexandra H. Fairchild, Bahri Üstünsöz, Dan F. Laney IV, John Hunt, Jonathan E. Schoen, Alan Marr, Patrick Greiffenstein, Lance Stuke, Alison A. Smith*

700 Original Article Single-center retrospective comparative study of 40–120 µm versus 100–300 µm trisacryl gelatin microspheres as the main embolic agents for prostatic artery embolization. *Hippocrates Moschouris, Çağın Şentürk, Konstantinos Stamatiou, Apostolos Mouroukis, Chrysovalantis Stylianou, Dimitrios Kontothanasis*

709 Pictorial Essay Post-cryoablation magnetic resonance imaging features of desmoid tumors: a pictorial essay. *Leonor G. Savarese, Nicolas Papalexis, Marco Miceli, Michela Carta, Giancarlo Facchini*

MUSCULOSKELETAL IMAGING

715 Original Article How may complementary magnetic resonance imaging findings facilitate the diagnosis of inflammatory hand arthritis involving the distal interphalangeal joint? A prospective cohort study. *Yasin Yaraşır, Gizem Ayan, Hanife Avcı, Levent Kılıç, Üstün Aydingöz, Umut Kalyoncu, Adalet Elçin Yıldız*

NEURORADIOLOGY

725 Original Article The hematoma measurement accuracy of dual-energy computed tomography angiography in spontaneous intracerebral hemorrhage. *Baoteng Zhang, Qingjing Huang, Feijian Wu, Suying Wu, Jianfang Huang, Zhihua Lu*



Dual-layer spectral detector computed tomography of peritumoral fat for preoperative prediction of tumor budding in colorectal adenocarcinoma

Jian-Bo Xu^{1,#}
 Wei-Jin Zhong^{1,#}
 Yuan-Hui Li^{1,#}
 Xiang Zhou¹
 Su-E Cao¹
 Xiao-Min Liu²
 Ruo-Mi Guo¹

¹The Third Affiliated Hospital, Sun Yat-Sen University, Department of Radiology, Guangzhou, China

²Philips Healthcare, Clinical and Technical Support, Guangzhou, China

#Contributed equally to this work as the first authors.

Handling editor: İlkay İdilman

Corresponding author: Ruo-Mi Guo

E-mail: guoruomi86@mail.sysu.edu.cn

Received 25 October 2025; revision requested 01 December 2025; accepted 25 January 2026.



Epub: 06.03.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263712

PURPOSE

To explore the value of dual-layer spectral detector computed tomography (DLSTCT)-derived parameters from peritumoral fat for preoperative tumor budding (TB) grade assessment in patients with colorectal adenocarcinoma (CRAC).

METHODS

This retrospective study enrolled 90 patients with pathologically confirmed CRAC who underwent preoperative DLSTCT between February 2024 and July 2025. Patients were stratified into low-grade and intermediate- to high-grade TB groups based on histopathology. The effective atomic number (Z_{eff}), iodine concentration (IC), normalized IC (NIC), virtual monoenergetic images (VMIs) at 40, 50, 60, and 70 keV in arterial phase (AP) and venous phase (VP), and the arterial enhancement fraction of intratumor and peritumoral fat were measured. Independent predictors were identified via multivariable logistic regression, and diagnostic performance was evaluated using receiver operating characteristic analysis.

RESULTS

Multiple DLSTCT-derived parameters of the peritumoral fat, including VMIs (40–70 keV) and Z_{eff} in both AP and VP, as well as NIC in VP, were significantly higher in the intermediate- to high-TB group than in the low-TB group (all $P < 0.05$; AP-VMI 40 keV, $P = 0.001$; AP-VMI 50 keV, $P = 0.002$; AP-VMI 60 keV, $P = 0.001$; AP-VMI 70 keV, $P = 0.001$; AP- Z_{eff} , $P = 0.002$; VP-VMI 40 keV, $P = 0.004$; VP-VMI 50 keV, $P = 0.001$; VP-VMI 60 keV, $P = 0.001$; VP-VMI 70 keV, $P < 0.001$; VP-NIC, $P = 0.024$; VP- Z_{eff} , $P = 0.019$), whereas other parameters derived from DLSTCT were not significantly different between these two groups ($P > 0.05$). Multivariable analysis revealed that VMIs (40–70 keV) and Z_{eff} in both AP and VP, as well as NIC in VP, remained independent predictors of TB grade. Among these, VMI at 70 keV in VP exhibited the strongest discriminatory ability [area under the curve (AUC): 0.742, 95% confidence interval (CI): 0.623–0.860]. The combined model integrating these parameters yielded the best predictive performance (AUC: 0.809, 95% CI: 0.713–0.906).

CONCLUSION

DLSTCT-derived quantitative parameters of peritumoral fat, particularly VMI and Z_{eff} values from both AP and VP, serve as independent predictors of TB grade in CRAC.

CLINICAL SIGNIFICANCE

DLSTCT-derived quantitative parameters of peritumoral fat can serve as a complementary, noninvasive quantitative imaging biomarker for preoperatively predicting TB grade in CRAC.

KEYWORDS

Colorectal adenocarcinoma, tumor budding, peritumoral fat, dual-layer spectral detector computed tomography

Colorectal cancer (CRC) ranks as the third most common cancer worldwide and the second leading cause of cancer-related deaths.^{1,2} Tumor budding (TB) is recognized as a significant prognostic factor in colorectal adenocarcinoma (CRAC) by the National Comprehensive Cancer Network and European Society for Medical Oncology guidelines.³ Defined by single cells or small clusters of up to four cells at the invasive tumor front, TB is a robust independent prognostic marker.⁴ It stratifies patients into risk categories for recurrence and mortality,⁵ making its preoperative identification crucial for treatment planning. Although histopathological biopsy remains the gold standard for TB grading, the technique is invasive and carries potential complications.⁴ Contrast-enhanced computed tomography (CECT) is a cornerstone in CRAC management, providing standardized protocols for initial staging and therapy monitoring.^{6,7} However, CECT is often insufficient for the comprehensive preoperative prediction of CRC pathological types, necessitating the development of more advanced quantitative imaging techniques.⁸

Dual-layer spectral detector CT (DLSCT) is an advanced modality that generates material decomposition images, such as iodine- or water-based maps. With more quantitative parameters than CECT, DLSCT has proven useful in diagnosing and managing various tumors.^{6,9-11} It demonstrates advantages in T-staging, evaluating tumor differentiation, perineural invasion (PNI), and lymphovascular invasion in CRC,¹²⁻¹⁷ and has been applied to assess TB grade.¹⁷ Additionally, DLSCT-based radiomics and machine learning have been used to evaluate distant metastasis and microsatellite instability in CRC.^{18,19}

In precision oncology, the peritumoral fat microenvironment (PFME) is increasingly recognized as a prognostic marker. Quanti-

tative PFME analysis can predict histopathological markers and molecular subtypes. The mesenchymal transition of peritumoral adipose tissue is characterized by the generation of dedifferentiated adipocytes, which constitute the major stromal cell type within tumors.²⁰⁻²² A dynamic interaction exists between tumors and adipose tissue, where cancer cells induce metabolic reprogramming in adjacent adipocytes to promote progression.²³ CRC remodels the peritumoral fat through an initial phase of structural disruption followed by reorganization, ultimately leading to a brown fat-like phenotype.²⁴ These alterations directly modulate X-ray attenuation properties, which are quantified by DLSCT, providing a direct mechanistic link between spectral parameters of peritumoral fat and TB, a histopathological marker of local tumor invasiveness.¹⁷

However, the quantitative application of DLSCT for peritumoral fat assessment in CRAC TB grading has not been previously explored. Therefore, this study evaluates the utility of DLSCT quantitative parameters for the preoperative diagnosis of TB grade in patients with CRAC.

Methods

Patients

This study was approved by the Institutional Research Ethics Committee of the Third Affiliated Hospital of Sun Yat-Sen University (clinical trial number: II2025-229-01, date: 06.17.2025), and the requirement for informed consent was waived by the ethics committee due to the retrospective nature of the study. Between February 2024 and July 2025, 577 consecutive patients with pathologically confirmed CRAC who underwent

preoperative spectral CT were initially enrolled. The inclusion criteria were (1) spectral CT performed within 2 weeks before surgery and (2) no prior treatment. The exclusion criteria included (1) preoperative radiotherapy or chemotherapy ($n = 446$), (2) CT-surgery interval exceeding 2 weeks ($n = 20$), (3) poor image quality due to artifacts ($n = 8$), and (4) mucinous or signet ring cell carcinoma ($n = 13$). The final cohort comprised 90 patients (24 with intermediate- to high-grade TB; 66 with low-grade TB). The entire analytical workflow of the study and the patient selection process are detailed in Figures 1 and 2, respectively.

Image acquisition

All scans were performed using a spectral CT scanner (IQon Spectral CT, Philips Healthcare, Best, The Netherlands). Patients were scanned in the supine position from the diaphragm to the pubic symphysis. The key parameters were as follows: tube voltage: 120 kV, automatic tube current modulation, pitch: 0.99, rotation time: 0.75 s, detector configuration: 64×0.625 mm, and slice thickness: 1.25 mm. The arterial phase (AP) was triggered at 150 Hounsfield units (HU) in the abdominal aorta, followed by a venous phase (VP) 60 seconds later.

Non-ionic contrast (iodixanol, Visipaque™, 300 mg I/mL) was administered intravenously at 1.2 mL/kg via a high-pressure syringe (3.0 mL/s), followed by a 30-mL saline flush (3.0 mL/s).

Image evaluation

Images were analyzed on an IntelliSpace Portal (Philips Healthcare) using Spectral CT Viewer software. Two radiologists (10 and 5 years of experience, respectively), blinded to

Main points

- Dual-layer spectral detector computed tomography-derived quantitative parameters of intratumor and peritumoral fat are used for assessing tumor budding (TB) grade in colorectal adenocarcinoma (CRAC).
- The virtual monoenergetic images (VMIs) (40–70 keV) and effective atomic number (Z_{eff}) in both arterial phase (AP) and venous phase (VP), as well as normalized iodine concentration in the VP of peritumoral fat, were higher in the intermediate- to high-TB group than in the low-TB group.
- The VMI and Z_{eff} values from both the AP and the VP of peritumoral fat serve as independent predictors of TB grade in CRAC.

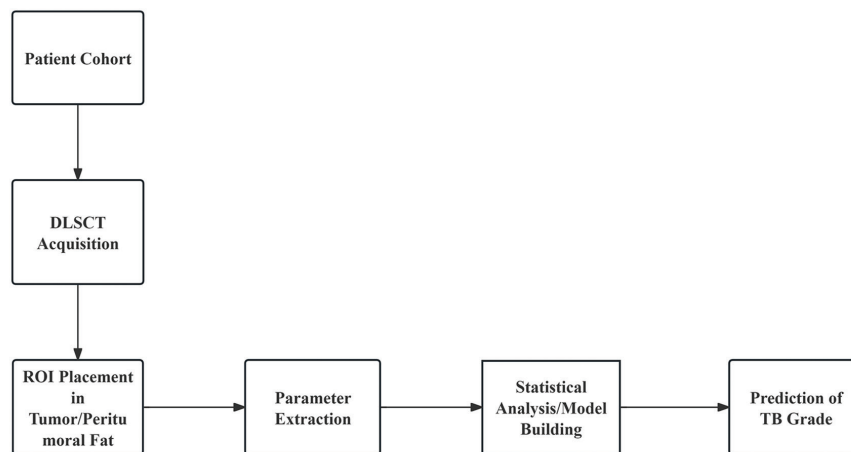


Figure 1. The entire analytical workflow of the study. DLSCT, dual-layer spectral detector computed tomography; TB, tumor budding; ROI, region of interest.

the pathological results, adhered to the same standardized protocol and independently established regions of interest (ROIs). Intra-tumor ROIs were drawn on 70-keV virtual monoenergetic image (VMI) slices showing the largest tumor area, avoiding necrosis, vessels, and cysts. Peritumoral fat ROIs were established, with a width of ≤ 5 and a 1-mm margin from the tumor (mean: 5.5 mm², range: 2.0–12.5 mm²).²⁵

The quantitative parameters, including iodine concentration (IC) for both the tumor and the abdominal aorta or iliac vessels, effective atomic number (Z_{eff}), and VMIs at 40, 50, 60, and 70 keV, were generated. All measurements were performed independently by two blinded radiologists, and the mean values were used for final analysis. Normalized IC (NIC) was calculated to standardize measurements.

Pathological evaluation

Grading of TB was conducted using hematoxylin–eosin staining, assessed by two gastrointestinal pathologists (10 and 20 years’ experience, respectively) following International Tumor Budding Consensus Conference 2016 guidelines.²⁶ Disagreements were resolved by consensus. Buds were counted at 20 × magnification and normalized to buds/0.785 mm².²⁶ The grading was as follows: Bd1 (low, ≤ 4 buds), Bd2 (intermediate, 5–9 buds), and Bd3 (high, ≥ 10 buds). For the analysis, Bd1 was classified as low-grade TB, and Bd2 and Bd3 as intermediate-to-high-grade TB.

Statistical analysis

All statistical analyses were performed using SPSS (version 26.0; IBM Corp., Chicago, IL, USA) and GraphPad Prism (version 9.5.1; GraphPad Software, San Diego, CA, USA). Data normality was assessed using the Shapiro–Wilk test. Normally distributed variables are presented as mean ± standard deviation, and non-normally distributed variables are presented as median and interquartile range. Group comparisons were conducted using the Mann–Whitney U test. Univariable and multivariable logistic regression analyses were performed to calculate odds ratios (ORs) with corresponding 95% confidence intervals (CIs). Predictive performance was evaluated using the area under the receiver operating characteristic curve (AUC), and AUCs were compared using the DeLong test. Multicollinearity was assessed using variance inflation factors (VIFs); variables with VIF > 10 were excluded. A backward stepwise

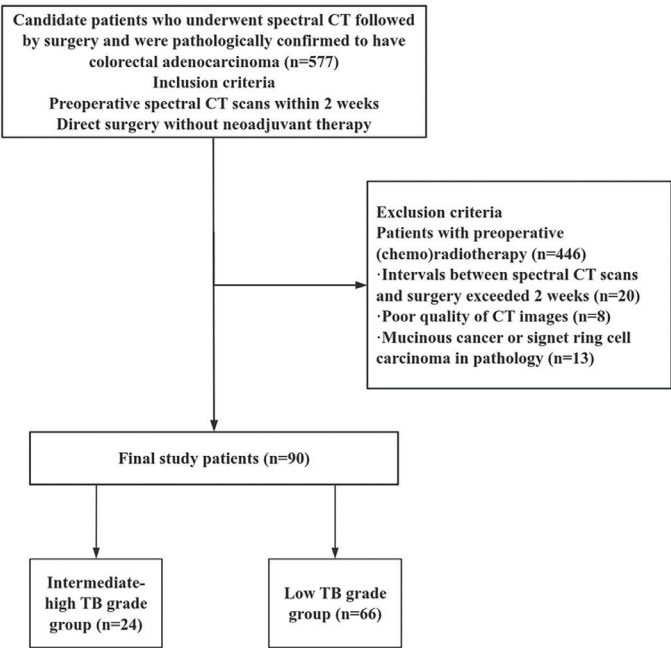


Figure 2. Patient screening flowchart. CT, computed tomography; CRAC, colorectal adenocarcinoma; TB, tumor budding.

elimination method (likelihood ratio) with a removal criterion of $P > 0.05$ was used to derive the final model. Interobserver agreement was assessed using the intraclass correlation coefficient (ICC), which was classified as poor (0–0.20), fair (0.21–0.40), moderate (0.41–0.60), good (0.61–0.80), or excellent (0.81–1.00). A two-sided P value < 0.05 was considered statistically significant.

Results

Demographic, clinical, and pathologic features

As shown in Table 1, 90 patients (53 men and 37 women; median age: 62 years) were enrolled in this study. All clinical characteristics (age, $P = 0.51$; gender, $P = 0.22$; location, $P = 0.75$; carcinoembryonic antigen, $P = 0.51$; carbohydrate antigen 199, $P = 0.26$) and

Table 1. Clinical characteristics of the patients with colorectal adenocarcinoma				
Characteristics	Overall n = 90	Tumor budding		P value
		Intermediate–high n = 24	Low n = 66	
Age (mean ± SD)	61.61 ± 11.34	59.38 ± 10.45	62.42 ± 11.61	0.51
Gender				
Female	37 (41.11%)	6 (25.00%)	31 (46.97%)	0.22
Male	53 (58.89%)	18 (75.00%)	35 (53.03%)	
Location				
Right	29 (32.22%)	9 (37.50%)	20 (30.30%)	0.75
Left	61 (67.78%)	15 (62.50%)	46 (69.70%)	
cT stage				
T1–2	17 (18.89%)	2 (8.33%)	15 (22.73%)	0.82
T3	50 (55.56%)	18 (75.00%)	32 (48.48%)	
T4	23 (25.55%)	4 (16.67%)	19 (28.79%)	
cN stage				
N0	47 (52.22%)	10 (41.67%)	37 (56.06%)	0.51
N1	28 (31.11%)	10 (41.67%)	18 (27.27%)	
N2	15 (16.67%)	4 (16.66%)	11 (16.67%)	

qualitative parameters of CECT (cT stage, $P = 0.82$; cN stage, $P = 0.51$; c-extramural vascular invasion, $P = 0.24$) were not significantly different between low-grade TB and intermediate- to high-grade TB of CRAC ($P > 0.05$).

Among the spectral parameters, VMIs (40–70 keV) and Z_{eff} in both AP and VP, as well as NIC in the VP of peritumoral fat, were significantly different between low-grade TB and intermediate- to high-grade TB of CRAC (all $P < 0.05$; AP-VMI 40 keV, $P = 0.001$; AP-VMI 50 keV, $P = 0.002$; AP-VMI 60 keV, $P = 0.001$; AP-VMI 70 keV, $P = 0.001$; AP- Z_{eff} , $P = 0.002$; VP-VMI 40 keV, $P = 0.004$; VP-VMI 50 keV, $P = 0.001$; VP-VMI 60 keV, $P = 0.001$; VP-VMI 70 keV, $P < 0.001$; VP-NIC, $P = 0.024$; VP- Z_{eff} , $P = 0.019$) (Table 2). Examples of CRAC cases with low-grade TB and intermediate- to high-grade TB are shown in Figures 3 and 4.

Spectral computed tomography parameters, predictors, and logistic models for tumor budding prediction

As shown in Table 3, univariate analysis showed that VMIs at 40, 50, 60, and 70 keV, Z_{eff} in both AP and VP, and NIC in the VP of peritumoral fat were associated with the TB of CRAC ($P < 0.05$; AP-VMI 40 keV, $P = 0.001$; AP-VMI 50 keV, $P = 0.002$; AP-VMI 60 keV, $P = 0.002$; AP-VMI 70 keV, $P = 0.001$; AP- Z_{eff} , $P = 0.002$; VP-VMI 40 keV, $P = 0.004$; VP-VMI 50 keV, $P = 0.001$; VP-VMI 60 keV, $P = 0.001$; VP-VMI 70 keV, $P < 0.001$; VP-NIC, $P = 0.024$; VP- Z_{eff} , $P = 0.019$). In the multivariate analysis, VMI at 40, 50, 60, and 70 keV, and Z_{eff} in both AP and VP remained independent predictors of TB grade (OR: 1.025, 95% CI: 0.610–0.833, $P = 0.001$; OR: 0.987, 95% CI: 0.601–0.827, $P = 0.002$; OR: 0.845, 95% CI: 0.606–0.830, $P = 0.002$; OR: 1.166, 95% CI: 0.609–0.832, $P = 0.002$; OR: 8.649, 95% CI: 0.594–0.826, $P = 0.002$; OR: 0.910, 95% CI: 0.575–0.820, $P = 0.004$; OR: 1.091, 95% CI: 0.608–0.841, $P = 0.001$; OR: 1.069, 95% CI: 0.604–0.841, $P = 0.001$; OR: 1.015, 95% CI: 0.623–0.860, $P = 0.001$; OR: 0.433, 95% CI: 0.533–0.792, $P = 0.019$). The AUCs of the spectral CT parameters are presented in Table 4, ranging from 0.662 to 0.742. Among these, VP-VMI at 70 keV of peritumoral fat yielded the highest AUC of 0.742, with sensitivities of 66.67% and specificities of 81.82% for differentiating between low-grade and intermediate- to high-grade TB. Combining all spectral parameters and clinical characteristics further improved predictive ability, with an AUC of 0.809, a sensitivity of 70.80%, and a specificity of 81.80% (Figures 5 and 6, Table 4). Figure 7 presents the results from the final fitted multivariable logistic regression model.

Table 1. Continued

Characteristics	Overall n = 90	Tumor budding		P value
		Intermediate–high n = 24	Low n = 66	
cEMVI				
Positive (+)	26 (28.89%)	10 (41.67%)	16 (24.24%)	0.24
Negative (–)	64 (71.11%)	14 (58.33%)	50 (75.76%)	
CEA				
< 5 ng/mL	62 (68.89%)	15 (62.50%)	47 (71.21%)	0.51
≥ 5 ng/mL	28 (31.11%)	9 (37.50%)	19 (28.79%)	
CA199				
< 35 U/mL	78 (86.67%)	19 (79.17%)	59 (89.39%)	0.26
≥ 35 U/mL	12 (13.33%)	5 (20.83%)	7 (10.61%)	

SD, standard deviation; cT stage, clinical T stage based on computed tomography; cN stage, clinical N stage based on computed tomography; cEMVI, extramural vascular invasion based on computed tomography; CEA, carcinoembryonic antigen; CA199, carbohydrate antigen 199.

Table 2. Quantitative parameters of DLSCT in the evaluation of tumor budding grade in patients with colorectal adenocarcinoma

Parameters	Overall n = 90	Tumor budding		P value
		Intermediate– high n = 24	Low n = 66	
Intratumor				
AP-VMI 40 keV	171.70 ± 51.51	162.80 ± 50.94	175.00 ± 51.73	0.194
AP-VMI 50 keV	121.10 ± 32.68	117.10 ± 29.53	122.60 ± 25.16	0.297
AP-VMI 60 keV	96.79 ± 23.36	94.17 ± 20.91	97.74 ± 24.26	0.387
AP-VMI 70 keV	85.03 ± 18.00	83.47 ± 17.72	85.59 ± 18.21	0.423
AP-IC	1.45 ± 0.57	1.32 ± 0.58	1.50 ± 0.57	0.135
AP-NIC	0.15 ± 0.06	0.14 ± 0.07	0.15 ± 0.62	0.124
AP- Z_{eff}	8.09 ± 0.29	8.02 ± 0.30	8.12 ± 0.29	0.151
VP-VMI 40 keV	169.30 ± 35.87	167.80 ± 37.87	169.90 ± 35.40	0.777
VP-VMI 50 keV	121.40 ± 24.52	122.10 ± 23.15	121.20 ± 25.16	0.991
VP-VMI 60 keV	96.05 ± 17.66	96.83 ± 16.29	95.77 ± 18.25	0.944
VP-VMI 70 keV	83.70 ± 13.77	83.72 ± 13.69	83.70 ± 13.90	0.995
VP-IC	1.43 ± 0.37	1.41 ± 0.41	1.44 ± 0.36	0.801
VP-NIC	0.34 ± 0.09	0.35 ± 0.09	0.33 ± 0.09	0.588
VP- Z_{eff}	8.10 ± 0.22	8.06 ± 0.28	8.11 ± 0.19	0.767
AEF	1.22 ± 1.32	1.18 ± 0.35	1.23 ± 1.53	0.107
Peritumoral fat				
AP-VMI 40 keV	–43.39 ± 56.42	–11.00 ± 46.51	–55.17 ± 55.36	0.001*
AP-VMI 50 keV	–36.19 ± 43.25	–12.14 ± 36.57	–44.93 ± 42.39	0.002*
AP-VMI 60 keV	–31.73 ± 35.71	–12.34 ± 30.88	–38.78 ± 34.92	0.001*
AP-VMI 70 keV	–29.72 ± 30.82	–12.33 ± 27.71	–36.04 ± 29.61	0.001*
AP-IC	0.26 ± 0.28	0.30 ± 0.22	0.25 ± 0.30	0.083
AP-NIC	0.03 ± 0.03	0.31 ± 0.20	0.02 ± 0.03	0.054
AP- Z_{eff}	6.98 ± 0.40	7.20 ± 0.31	6.90 ± 0.40	0.002*
VP-VMI 40 keV	–30.23 ± 54.07	–2.79 ± 49.04	–40.21 ± 52.66	0.004*
VP-VMI 50 keV	–28.24 ± 42.84	–4.10 ± 35.48	–37.01 ± 42.12	0.001*
VP-VMI 60 keV	–25.41 ± 34.34	–6.04 ± 29.38	–32.45 ± 33.46	0.001*
VP-VMI 70 keV	–23.15 ± 29.99	–5.68 ± 25.63	–29.50 ± 29.08	<0.001*
VP-IC	0.26 ± 0.22	0.33 ± 0.24	0.24 ± 0.21	0.083

Interobserver agreement

The interobserver agreement for CT images measured by two radiologists was excellent, with ICC values ranging from 0.657 (95% CI: 0.479–0.774) to 0.944 (95% CI: 0.914–0.963), except for the arterial enhancement fraction of peritumoral fat (Supplementary Table 1). This may be attributed to increased image noise at low keV or the challenge of consistently placing ROIs in heterogeneous peritumoral fat.

Table 2. Continued				
Parameters	Overall	Tumor budding		P value
	n = 90	Intermediate–high n = 24	Low n = 66	
VP-NIC	0.61 ± 0.50	0.08 ± 0.06	0.05 ± 0.05	0.024*
VP-Z _{eff}	7.08 ± 0.39	7.23 ± 0.37	7.03 ± 0.38	0.019*
AEF	1.98 ± 3.29	1.25 ± 1.14	2.26 ± 3.80	0.436

*Independently significant variables. Values are presented as mean ± standard deviation. DLSCT, dual-layer spectral detector computed tomography; VMI, virtual monoenergetic image; Z_{eff}, effective atomic number; IC, iodine concentration; NIC, normalized iodine concentration; AEF, arterial enhancement fraction; AP, arterial phase; VP, venous phase.

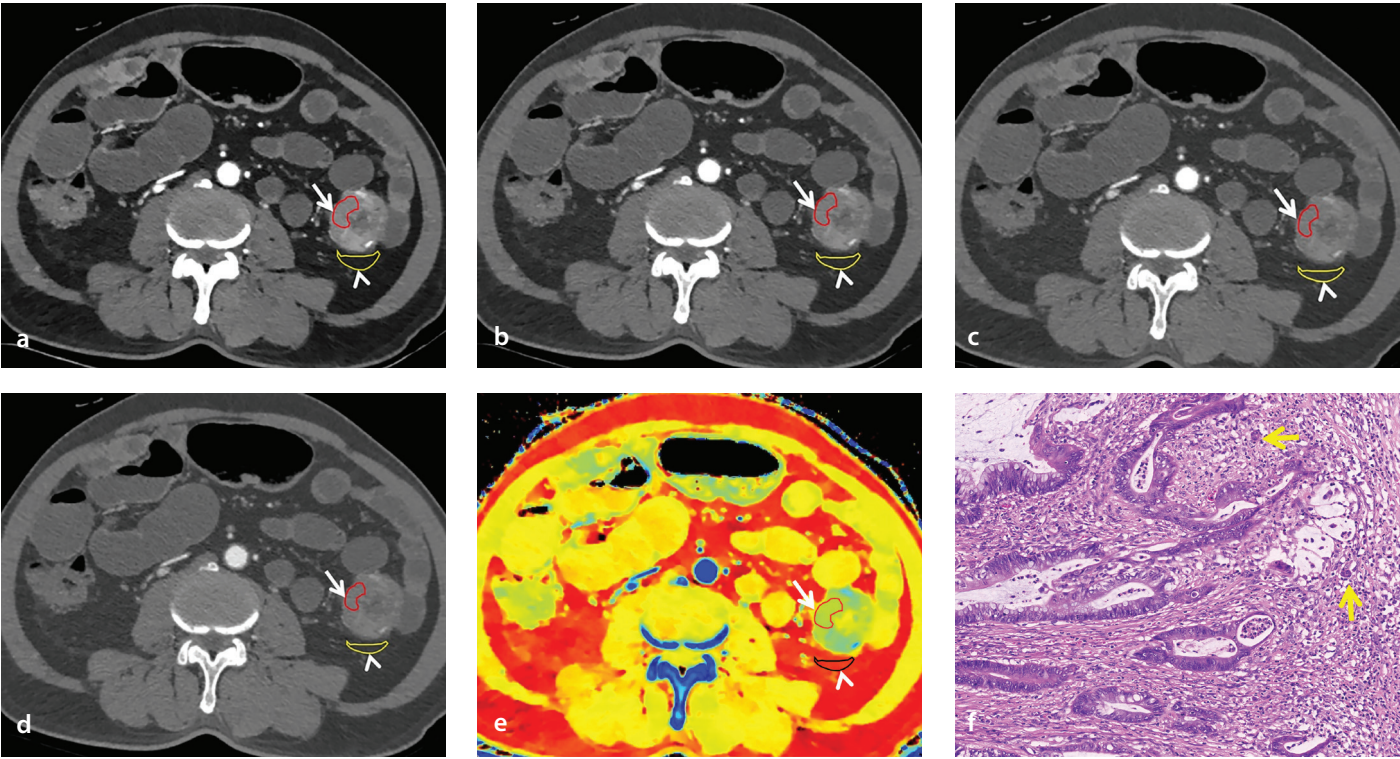


Figure 3. A 68-year-old man with low-grade TB descending colon adenocarcinoma. Regions of interest were established on the tumor (red) and peritumoral fat (yellow). (a) 40 keV-AP: tumor 133.05 HU (arrow), fat –147.15 HU (arrowhead); (b) 50 keV-AP: tumor 93.3 HU (arrow), fat –114.45 HU (arrowhead); (c) 60 keV-AP: tumor 75.9 HU (arrow), fat –95.3 HU (arrowhead); (d) 70 keV-AP: tumor 71.0 HU (arrow), fat –83.85 HU (arrowhead); (e) AP-Z_{eff}: tumor 7.91 (arrow), fat 6.21 (arrowhead); (f) histopathology shows < 5 buds (arrow) (hematoxylin–eosin, × 200). TB, tumor budding; HU, Hounsfield unit; AP, arterial phase; VP, venous phase.

Table 3. Univariate and multivariate analysis of predictors of tumor budding grade in patients with colorectal adenocarcinoma				
Parameters	Univariate parameter		Multivariate parameter	
	Z/χ ²	P value	OR	P value
Age	–1.196	0.232		
Gender	–1.889	0.062		
cEMVI	–1.618	0.109		
cT stage	0.295	0.768		
cN stage	–0.799	0.426		
CEA	0.958	0.338		
CA199	1.481	0.139		
Intratumor				
AP-VMI 40 keV	–1.305	0.192		
AP-VMI 50 keV	–1.049	0.294		
AP-VMI 60 keV	–0.871	0.384		
AP-VMI 70 keV	–0.808	0.419		

Parameters	Univariate parameter		Multivariate parameter	
	Z/ χ^2	P value	OR	P value
VP-Z _{eff}	−0.301	0.763		
AEF	1.615	0.106		
AP-IC	−1.501	0.133		
AP-NIC	−1.542	0.123		
AP-Z _{eff}	−1.442	0.149		
VP-VMI 40 keV	−0.287	0.774		
VP-VMI 50 keV	0.014	0.989		
VP-VMI 60 keV	0.073	0.942		
VP-VMI 70 keV	−0.009	0.993		
VP-IC	−0.255	0.798		
VP-NIC	0.547	0.584		
Peritumoral fat				
AP-VMI 40 keV	3.198	0.001*	1.026	0.001**
AP-VMI 50 keV	3.089	0.002*	0.990	0.002**
AP-VMI 60 keV	3.148	0.002*	0.849	0.002**
AP-VMI 70 keV	3.184	0.001*	1.157	0.002**
AP-IC	1.735	0.083		
AP-NIC	1.926	0.054		
AP-Z _{eff}	3.034	0.002*	6.811	0.002**
VP-VMI 40 keV	2.856	0.004*	0.913	0.004**
VP-VMI 50 keV	3.248	0.001*	1.091	0.001**
VP-VMI 60 keV	3.216	0.001*	1.062	0.001**
VP-VMI 70 keV	3.495	<0.001*	1.016	0.001**
VP-IC	1.738	0.082		
VP-NIC	2.249	0.024*	4.441	0.169
VP-Z _{eff}	2.345	0.019*	0.466	0.019**
AEF	−0.786	0.432		

*Variables included multivariable analysis based on $P < 0.1$; **Independently significant variables.
OR, odds ratio; cEMVI, extramural vascular invasion based on computed tomography; cT stage, clinical T stage based on computed tomography; cN stage, clinical N stage based on computed tomography; CEA, carcinoembryonic antigen; CA199, carbohydrate antigen199; PNI, perineural invasion; VMI, virtual monoenergetic image; Z_{eff}, effective atomic number; IC, iodine concentration; NIC, normalized iodine concentration; AEF, arterial enhancement fraction; AP, arterial phase; VP, venous phase.

Parameter	AUC (95% CI)	Cut-off	Sensitivity	Specificity
AP-VMI 40 keV	0.721 (0.610, 0.833)	−37.43	79.17	57.58
AP-VMI 50 keV	0.714 (0.601, 0.827)	−41.85	79.17	54.55
AP-VMI 60 keV	0.718 (0.606, 0.830)	−48.63	91.67	43.94
AP-VMI 70 keV	0.720 (0.609, 0.832)	−40.90	87.50	50.00
AP-Z _{eff}	0.710 (0.594, 0.826)	6.86	91.67	45.45
VP-VMI 40 keV	0.698 (0.575, 0.820)	−12.53	66.67	71.21
VP-VMI 50 keV	0.725 (0.608, 0.841)	−4.30	62.50	80.30
VP-VMI 60 keV	0.723 (0.604, 0.841)	−3.55	62.50	81.82
VP-VMI 70 keV	0.742 (0.623, 0.860)	−5.10	66.67	81.82
VP-Z _{eff}	0.662 (0.533, 0.792)	7.19	66.67	69.70
All parameters combined	0.809 (0.713, 0.906)	0.38	70.80	81.80

*Independently significant variables. AUC, area under curve; CI, confidence intervals; VMI, virtual monoenergetic image; Z_{eff}, effective atomic number; AP, arterial phase; VP, venous phase.

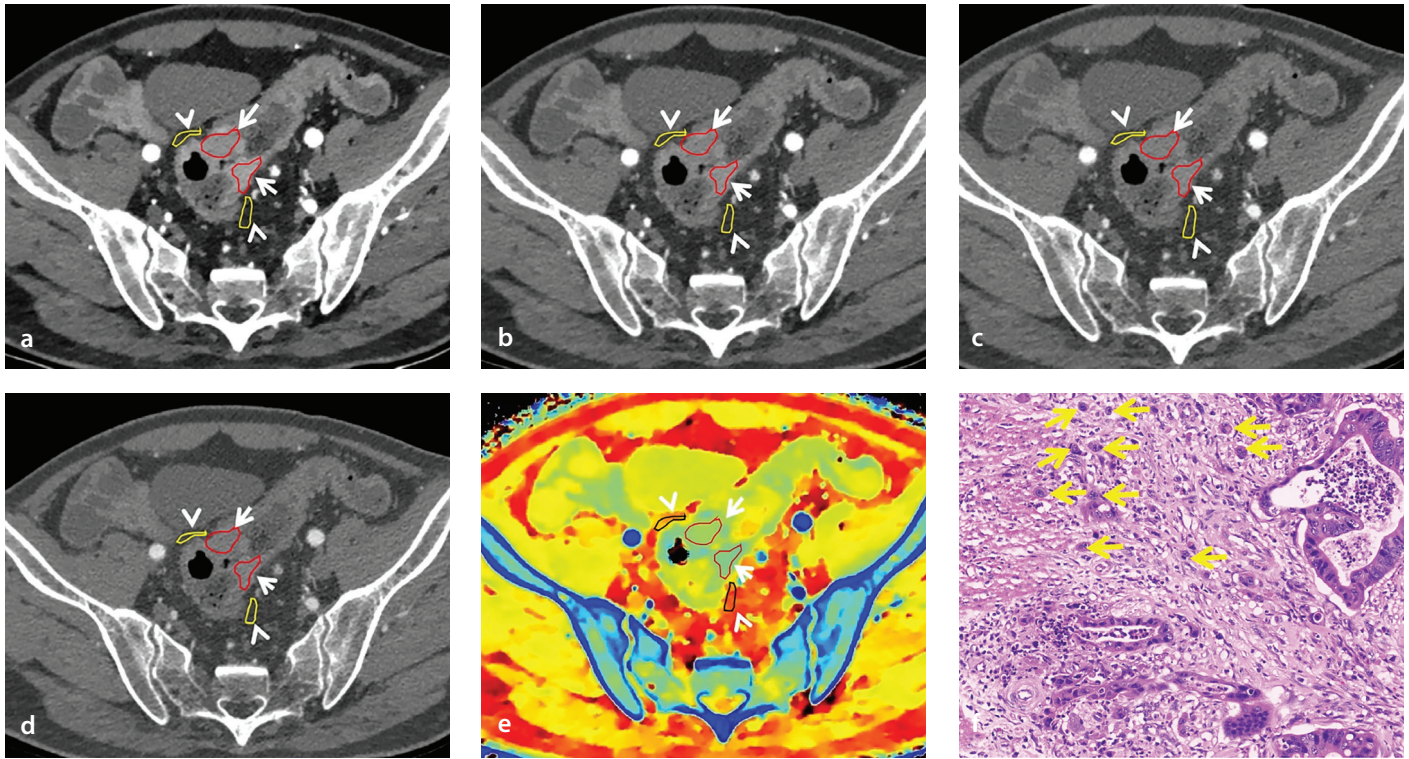


Figure 4. A 67-year-old man with high-grade-TB sigmoid colon adenocarcinoma. Regions of interest were established on the tumor (red) and peritumoral fat (yellow). (a) 40 keV-AP: tumor 162.85 HU (arrow), fat 22.45 HU (arrowhead); (b) 50 keV-AP: tumor 120.35 HU (arrow), fat 8.05 HU (arrowhead); (c) 60 keV-AP: tumor 95.4 HU (arrow), fat -0.4 HU (arrowhead); (d) 70 keV-AP: tumor 80.5 HU (arrow), fat -5.55 HU (arrowhead); (e) AP- Z_{eff} : tumor 8.09 (arrow), fat 7.44 (arrowhead); (f) histopathology shows > 10 buds (arrow) (HE, $\times 200$). TB, tumor budding; HU, Hounsfield unit; AP, arterial phase; VP, venous phase.

Discussion

This study identified DLSCT-derived quantitative parameters of peritumoral fat—specifically AP and VP VMI and Z_{eff} values—as independent predictors of TB grade in CRAC. This predictive utility is likely mediated by a mesenchymal transition of the peritumoral fat, characterized by marked metabolic re-wiring (e.g., upregulation of UCP-1, TMEM26, and PON3, and downregulation of Pref-1 and adiponectin), which generates dedifferentiated adipocytes that form a key stromal component within tumors.^{20,27}

Multiparametric DLSCT—including VMIs, IC, Z_{eff} and the HU curve slope (λHU)—provides valuable insights for tumor assessment.²⁸⁻³⁰ Iodine mapping quantitatively evaluates microcirculation via IC, offering clinical value in CRAC for staging, differentiation, PNI, and lymph node metastasis assessment.^{7,12,14,31} Low-energy 40-keV VMIs provide a superior signal-to-noise ratio and contrast-to-noise ratio, improving tumor detection clarity.^{32,33} This enhancement stems from the proximity of 40 keV to the iodine K-edge (33 keV), optimizing visualization of iodine-rich areas.^{33,34} Tumor vascularization, derived from intrinsic angiogenesis and co-option of host vessels,³⁵ correlates with growth and metastatic potential.^{35,36} The Z_{eff}

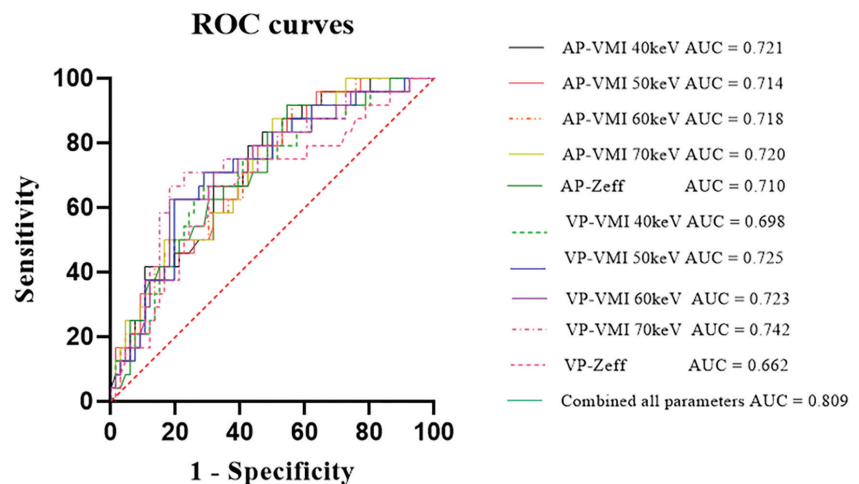


Figure 5. Receiver operating characteristic curves (ROCs) of peritumoral fat spectral parameters for tumor budding grade evaluation in colorectal adenocarcinoma. AUC, area under curve; AP, arterial phase; VP, venous phase; VMI, virtual monoenergetic image.

quantifies material composition by representing equivalent elemental atomic numbers in X-ray attenuation^{37,38} and can predict pathological subtypes at specific energies.³⁹

In our study, VMIs at 40, 50, 60, and 70 keV, Z_{eff} in both AP and VP, and NIC in the VP of peritumoral fat were significantly different between the low-grade TB and intermediate-to high-grade TB groups. This phenomenon likely results from improved visualization of contrast-enhanced lesions at lower VMI en-

ergy levels.⁴⁰⁻⁴² In our cohort, no significant correlations were found between tumor spectral parameters and TB grade, whereas peritumoral fat parameters showed strong associations. This spatial heterogeneity indicates distinct microenvironmental influences on spectral CT parameters. Our findings differ from those reported by Shao et al.,¹⁷ potentially due to divergent DLSCT protocols. Our AP scan was initiated 5 seconds after aortic enhancement reached 150 HU, potentially capturing a later, mixed AP/portal VP phase.

In contrast, Shao et al.¹⁷ triggered scanning immediately upon detection of 100 HU. This fundamental difference in temporal triggering could result in divergent enhancement patterns of both tumor and peritumoral fat, affecting absolute parameter values and their correlation with TB. This likely accounts for the discrepancies in VP characterization observed between the studies.

In the present study, we found that some new parameters are associated with the TB grade in CRAC. Peritumoral fat spectral parameters were not reported in previous studies to predict TB grade in CRAC. In the logistic regression analysis, VMIs at 40, 50, 60, and 70 keV, Z_{eff} in both AP and VP, and NIC in the VP of peritumoral fat were independent predictors of the TB grade in CRAC. We propose that the TB-spectral CT correlation stems from a structural–functional cascade in peritumoral fat. The increase in both VMI attenuation and Z_{eff} can be attributed to several concurrent pathological processes in the peritumoral fat: fibrosis (with higher cellularity and connective tissue deposition), microvascular

proliferation (increasing local blood volume), and inflammatory infiltration. These elements have intrinsically higher atomic numbers and density than pure adipose tissue. In parallel, a rise in NIC is a direct biomarker of increased iodine accumulation, signifying enhanced vascular perfusion, and potentially increased vascular permeability, likely mediated by tumor-derived angiogenic signaling. This architectural alteration directly alters X-ray absorption, producing measurable changes in spectral CT parameters, such as attenuation values and spectral λ HU.⁴³

Consequently, a clinically relevant workflow may be seamlessly integrated into preoperative DLST staging, including performing quantitative peritumoral fat analysis, appending the results (e.g., high-grade TB probability) to the radiology report, and thereby informing the multidisciplinary team. This could impact surgical planning (e.g., margin consideration), neoadjuvant therapy stratification in rectal cancer, and postoperative prognostic stratification—especially crucial in stage II colon cancer, where

TB is a high-risk feature—guiding follow-up intensity and adjuvant therapy.

This study has some limitations. First, its single-center, retrospective nature, coupled with its limited sample size, may have introduced selection bias. Moreover, the absence of internal validation (e.g., bootstrapping or cross-validation) for the logistic model necessitates future verification in larger, multicenter cohorts. Second, the study focused solely on adenocarcinomas, excluding other CRC histological variants, thereby limiting the applicability of our findings to all CRC subtypes. Third, subgroup distribution was uneven. Fourth, this study did not retrospectively collect body mass index or quantify visceral fat area. Finally, this study was restricted to a single CT vendor’s technology, and key parameters (VMI, IC, Z_{eff}) vary across different systems. Therefore, although the diagnostic concept is transferable, our specific thresholds are vendor sensitive, mandating multicenter studies across diverse platforms to derive generalizable standards. Future studies should also explore spectral CT’s role in evaluating tumor deposits, undertake multicenter investigations with larger sample sizes to confirm the robustness and reproducibility of these observations, and incorporate DLST spectral parameters into machine learning or radiomics models for automated, high-throughput analysis.

In conclusion, DLST-derived quantitative parameters of peritumoral fat, particularly VMI and Z_{eff} values from both AP and VP, serve as independent predictors of TB grade in CRAC.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

This study was supported by grants from the National Natural Science Foundation of China (No. 82471567, 81801757), Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515010279, 2023A1515010256).

Supplementary: <https://d2v96fxpocvxx.cloudfront.net/b65011c5-47cb-4d06-b7cc-d78a7e40a205/content-images/946a7c82-b1e4-4c07-9a2c-20174dc48e34.pdf>

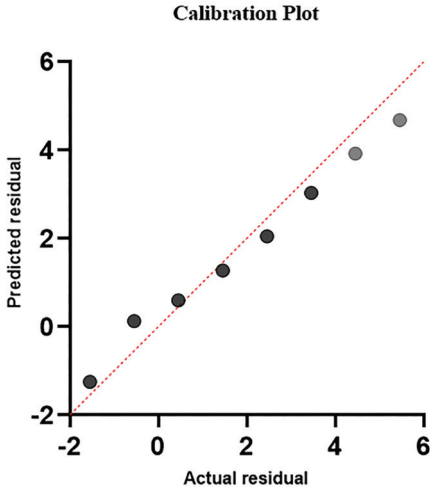


Figure 6. Calibration plot of the final combined logistic regression model.

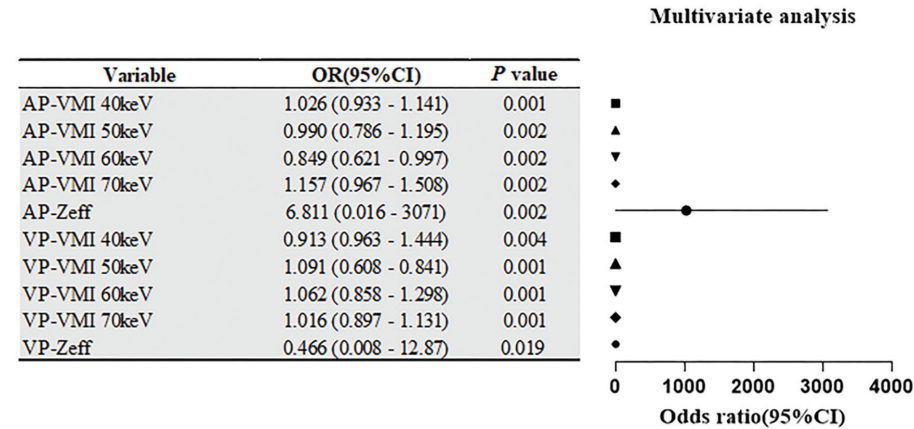


Figure 7. Forest plot of odds ratios from multivariable logistic regression of peritumoral fat. OR, odds ratio; CI, confidence interval; AP, arterial phase; VP, venous phase.

References

- Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin.* 2023;73(1):17-48. [\[Crossref\]](#)
- Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. *CA Cancer J Clin.* 2023;73:233-254. [\[Crossref\]](#)
- Benson AB, Venook AP, Adam M, et al. Colon Cancer, Version 3.2024, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw.* 2024;22:e240029. [\[Crossref\]](#)
- Kamall GH, Ulusoy C, Nikolovski A, Kamall S. Tumour budding - an additional prognostic factor in colorectal cancer survival. *Pol J Pathol.* 2023;74:36-41. [\[Crossref\]](#)
- Basile D, Broudin C, Emile JF, et al. Tumor budding is an independent prognostic factor in stage III colon cancer patients: a post-hoc analysis of the IDEA-France phase III trial (PRODIGE-GERCOR). *Ann Oncol.* 2022;33(6):628-637. [\[Crossref\]](#)
- Kijima S, Sasaki T, Nagata K, Utano K, Lefor AT, Sugimoto H. Preoperative evaluation of colorectal cancer using CT colonography, MRI, and PET/CT. *World J Gastroenterol.* 2014;20(45):16964-16975. [\[Crossref\]](#)
- Zhou XC, Chen QL, Huang CQ, Liao HL, Ren CY, He QS. The clinical application value of multi-slice spiral CT enhanced scans combined with multiplanar reformations images in preoperative T staging of rectal cancer. *Medicine (Baltimore).* 2019;98:e16374. [\[Crossref\]](#)
- Inamdar A, Shinde RK. The diagnostic impact of contrast-enhanced computed tomography (CECT) in evaluating lymph node involvement in colorectal cancer: a comprehensive review. *Cureus.* 2024;16:e61832. [\[Crossref\]](#)
- Li M, Qin H, Yu X, et al. Preoperative prediction of Lauren classification in gastric cancer: a radiomics model based on dual-energy CT iodine map. *Insights Imaging.* 2023;14(1):125. [\[Crossref\]](#)
- You Y, Wang Y, Yu X, et al. Prediction of lymph node metastasis in advanced gastric adenocarcinoma based on dual-energy CT radiomics: focus on the features of lymph nodes with a short axis diameter ≥ 6 mm. *Front Oncol.* 2024;14:1369051. [\[Crossref\]](#)
- Zhu Y, Wang P, Wang B, et al. Dual-layer spectral-detector CT for predicting microsatellite instability status and prognosis in locally advanced gastric cancer. *Insights Imaging.* 2023;14(1):151. [\[Crossref\]](#)
- Chen W, Ye Y, Zhang D, et al. Utility of dual-layer spectral-detector CT imaging for predicting pathological tumor stages and histologic grades of colorectal adenocarcinoma. *Front Oncology* 2022;12:1002592. [\[Crossref\]](#)
- Sun Q, Bian X, Sun D, et al. The value of preoperative diagnosis of colorectal adenocarcinoma pathological T staging based on dual-layer spectral-detector computed tomography extracellular volume fraction: a preliminary study. *Jpn J Radiol.* 2024;42:612-621. [\[Crossref\]](#)
- Lu W, Tan X, Zhong Y, et al. Spectral CT in the evaluation of perineural invasion status in rectal cancer. *Jpn J Radiol.* 2024;42:1012-1020. [\[Crossref\]](#)
- Yuan X, Quan X, Che X, et al. Preoperative prediction of the lymphovascular tumor thrombus of colorectal cancer with the iodine concentrations from dual-energy spectral CT. *BMC Med Imaging.* 2023;23(1):103. [\[Crossref\]](#)
- Chen J, Ni L, Gong J, et al. Quantitative parameters of dual-layer spectral detector computed tomography for evaluating differentiation grade and lymphovascular and perineural invasion in colorectal adenocarcinoma. *Eur J Radiol.* 2024;178:111594. [\[Crossref\]](#)
- Shao C, He C, Zheng P, Zhou P, Chen X. Preoperative prediction of tumor budding and lymphovascular invasion in colon cancer using dual-energy CT: a prospective study with internal model validation. *Abdom Radiol (NY).* 2025;50(8):3406-3414. [\[Crossref\]](#)
- Sun Q, Bian X, Feng F, Chen S, Chen G. A nomogram based on dual-layer spectral detector CT-derived 40KeV virtual monoenergetic images for preoperative prediction of simultaneous distant metastasis in colorectal cancer. *BMC Med Imaging.* 2025;25(1):455. [\[Crossref\]](#)
- Chen W, Jia Z, Wan L, et al. Interpretable radiomics model based on dual-layer spectral CT iodine maps for predicting microsatellite instability in colorectal cancer: a two-center study. *Eur J Radiol.* 2025;192:112357. [\[Crossref\]](#)
- Conti G, Calderan L, Quintero Sierra LA, et al. Tumor and peritumoral adipose tissue crosstalk: de-differentiated adipocytes influence spread of colon carcinoma cells. *Tissue Cell.* 2023;80:101990. [\[Crossref\]](#)
- Amor S, Iglesias-de la Cruz MC, Ferrero E, et al. Peritumoral adipose tissue as a source of inflammatory and angiogenic factors in colorectal cancer. *Int J Colorectal Dis.* 2016;31:365-375. [\[Crossref\]](#)
- Zoico E, Rizzatti V, Darra E, et al. Morphological and functional changes in the peritumoral adipose tissue of colorectal cancer patients. *Obesity (Silver Spring).* 2017; 25(Suppl 2):S87-S94. [\[Crossref\]](#)
- Liesenfeld DB, Grapov D, Fahrman JF, et al. Metabolomics and transcriptomics identify pathway differences between visceral and subcutaneous adipose tissue in colorectal cancer patients: the ColoCare study. *Am J Clin Nutr.* 2015;102:433-443. [\[Crossref\]](#)
- Conti G, Minicozzi A, Merigo F, et al. Morphogenetic events in the perinodal connective tissue in a metastatic cancer model. *Biomed Pharmacother.* 2013;67:1-6. [\[Crossref\]](#)
- Ahn H, Won Lee J, Jang SH, et al. Prognostic significance of imaging features of peritumoral adipose tissue in FDG PET/CT of patients with colorectal cancer. *Eur J Radiol.* 2021;145:110047. [\[Crossref\]](#)
- Lugli A, Kirsch R, Ajioka Y, et al. Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016. *Mod Pathol.* 2017;30(9):1299-1311. [\[Crossref\]](#)
- Pesapane A, Capasso L, Del Sorbo MR, et al. Adipose-tumor crosstalk in colorectal cancer: Identifying (Epi)genetic biomarkers for tumor progression and cachexia. *Cell Death Dis.* 2025;16(1):675. [\[Crossref\]](#)
- Yingying L, Zhe Z, Xiaochen W, Xiaomei L, Nan J, Shengjun S. Dual-layer detector spectral CT-a new supplementary method for preoperative evaluation of glioma. *Eur J Radiol.* 2021;138:109649. [\[Crossref\]](#)
- Li YX, Li WJ, Xu YS, et al. Clinical application of dual-layer spectral CT multi-parameter feature to predict microvascular invasion in hepatocellular carcinoma. *Clin Hemorheol Microcirc.* 2024;88:97-113. [\[Crossref\]](#)
- Zhang K, Zhang J, Li M, Liu X, Xu Y. Prediction of histologic grade of hepatocellular carcinoma using dual-layer spectral-detector computed tomography (CT): comparison of two region of interest plotting methods. *Quant Imaging Med Surg.* 2024;14(6):3887-3900. [\[Crossref\]](#)
- Luo M, Chen G, Xie H, et al. Correction: Preoperative diagnosis of metastatic lymph nodes by CT-histopathologic matching analysis in gastric adenocarcinoma using dual-layer spectral detector CT. *Eur Radiol.* 2023;33(12):9480. Erratum for: *Eur Radiol.* 2023;33(12):8948-8956. [\[Crossref\]](#)
- Xu YK, Chai TT, Wang JW, et al. Optimal virtual monochromatic images for assessing metastatic lateral cervical lymph nodes in patients with papillary thyroid carcinoma using dual-layer spectral detector computed tomography. *Eur J Radiol.* 2024;178:111623. [\[Crossref\]](#)
- Wu L, Cen C, Yue X, et al. A clinical-radiomics nomogram based on dual-layer spectral detector CT to predict cancer stage in pancreatic ductal adenocarcinoma. *Cancer Imaging.* 2024;24(1):55. [\[Crossref\]](#)
- Kim N, Bae K, Kim HC, Jeon KN. Added value of 40 keV virtual monoenergetic images for diagnosing malignant pleural effusion on chest CT. *Jpn J Radiol.* 2024;42(8):862-871. [\[Crossref\]](#)
- Zhao L, Zhou W, Fu Y, et al. Diagnostic value of one-stop CT energy spectrum and perfusion for angiogenesis in colon and rectum cancer. *BMC Med Imaging.* 2024;24:116. [\[Crossref\]](#)
- Sun H, Xu Y, Yang Q, Wang W. Assessment of tumor grade and angiogenesis in colorectal cancer: whole-volume perfusion CT. *Acad Radiol.* 2014;21:750-757. [\[Crossref\]](#)

37. Drljevic-Nielsen A, Donskov F, Mains JR, et al. Prognostic utility of parameters derived from pretreatment dual-layer spectral-detector CT in patients with metastatic renal cell carcinoma. *AJR Am J Roentgenol*. 2022;218(5):867-876. [\[Crossref\]](#)
38. Garcia LI, Azorin JF, Almansa JF. A new method to measure electron density and effective atomic number using dual-energy CT images. *Phys Med Biol*. 2016;61(1):265-279. [\[Crossref\]](#)
39. Zhu Y, Feng B, Cai W, et al. Prediction of microvascular invasion in solitary AFP-negative hepatocellular carcinoma ≤ 5 cm using a combination of imaging features and quantitative dual-layer spectral-detector CT parameters. *Acad Radiol*. 2023;30(Suppl 1):S104-S116. [\[Crossref\]](#)
40. Liang H, Zhou Y, Zheng Q, et al. Dual-energy CT with virtual monoenergetic images and iodine maps improves tumor conspicuity in patients with pancreatic ductal adenocarcinoma. *Insights Imaging*. 2022;13(1):153. [\[Crossref\]](#)
41. Noda Y, Takai Y, Suto T, et al. Effect of X-ray tube on image quality and pancreatic ductal adenocarcinoma conspicuity in pancreatic protocol dual-energy CT. *Clin Radiol*. 2024;79:e554-e559. [\[Crossref\]](#)
42. Woeltjen MM, Niehoff JH, Roggel R, et al. Pancreatic cancer in photon-counting CT: low keV virtual monoenergetic images improve tumor conspicuity. *Eur J Radiol*. 2024;173:111374. [\[Crossref\]](#)
43. Mehta A, Goswami M, Sinha R, Dogra A. Histopathological significance and prognostic impact of tumor budding in colorectal cancer. *Asian Pac J Cancer Prev*. 2018;19(9):2447-2453. [\[Crossref\]](#)



Letter to the Editor: comments on the diagnostic value of ADC texture analysis in PI-RADS 5 lesions

Esat Kaba

Recep Tayyip Erdoğan University Training and
Research Hospital, Department of Radiology, Rize,
Türkiye

Dear Editor,

I read with great interest the research article titled "Texture analysis enhances diagnostic accuracy of lesions scored as 5 in the Prostate Imaging Reporting and Data System in magnetic resonance imaging," published by Bai et al.¹ in *Diagnostic and Interventional Radiology*. In this study, the authors investigated the diagnostic value of apparent diffusion coefficient (ADC) mapping texture analysis in distinguishing prostate carcinoma from prostatitis in lesions scored as Prostate Imaging Reporting and Data System (PI-RADS) 5. This approach focuses on a problem that is extremely meaningful and important in clinical practice. I commend the authors for these clinically valuable and comprehensive study, and I would like to offer a few additional perspectives.

This study applied only first-order texture analysis to ADC maps and reported highly favorable diagnostic performance for the combined clinical and ADC texture model, achieving an accuracy of 96.8%, a sensitivity of 98.7%, and an area under the curve of 93.1%. The relatively low spatial resolution of ADC maps and the considerable variability in acquisition and post-processing parameters across scanners can limit detailed image analyses.² In this context, the first-order texture analysis chosen for the study appears to be an appropriate and rational approach for evaluating ADC maps, as it reflects the distribution of voxel values within the selected region of interest (ROI). In contrast, more advanced radiomics metrics (such as GLCM, GLRLM, and GLSZM) may exhibit lower reproducibility in ADC maps with limited spatial resolution. Therefore, it is noteworthy and encouraging that the authors achieved such remarkable results in diagnostic differentiation using first-order texture analysis, a simpler and more practical approach, without resorting to advanced radiomics analyses. Based on these results, I fully agree with the authors' view that first-order texture analysis can be a powerful diagnostic aid in differentiating PI-RADS 5 benign and malignant lesions.

Nevertheless, the study has several important limitations. First, the use of images acquired from different magnetic resonance imaging scanners introduces heterogeneity that may influence texture analysis and potentially bias the results. Second, the use of three different methods for histopathological confirmation (biopsy, TUR-P, and radical prostatectomy) represents a relevant limitation. In particular, in cases diagnosed as prostatitis based solely on biopsy, the possibility of sampling inadequacy should not be overlooked. Third, volumetric analyses of the ADC maps were performed by a single reader using slice-by-slice ROI delineation. Current studies recommend that at least two independent readers perform segmentation and that only features demonstrating high intraclass correlation coefficients be included in the final analysis.³ Lastly, robust machine learning models could be created using texture analysis data to develop clinical, texture analysis, and combined models.

In conclusion, the authors demonstrate that the model developed in this important study enables practical and highly accurate differentiation between benign and malignant PI-RADS 5 lesions, which is a critically important clinical problem in urogenital radiology practice. I believe that with further development and resolution of certain methodological limitations, such models have the potential for successful integration into clinical practice.

Handling editor: Sonay Aydın

Corresponding author: Esat Kaba

E-mail: esatkaba04@gmail.com

Received 16 December 2025; accepted 23 December 2025.



Epub: 04.02.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2025.253811

Conflict of interest disclosure

The author declared no conflicts of interest.

References

1. Bai Y, Xie XR, Hou Y, Zhang YD, Shi HB, Wu CJ. Texture analysis enhances diagnostic accuracy of lesions scored as 5 in the Prostate Imaging Reporting and Data System in magnetic resonance imaging. *Diagn Interv Radiol.* 2026;32(2):156-163. [\[Crossref\]](#)
2. Traverso A, Kazmierski M, Welch ML, et al. Sensitivity of radiomic features to inter-observer variability and image pre-processing in apparent diffusion coefficient (ADC) maps of cervix cancer patients. *Radiother Oncol.* 2020;143:88-94. [\[Crossref\]](#)
3. Koçak B, Yüzkan S, Mutlu S, et al. Influence of image preprocessing on the segmentation-based reproducibility of radiomic features: in vivo experiments on discretization and resampling parameters. *Diagn Interv Radiol.* 2024;30(3):152-162. [\[Crossref\]](#)



Reply: comments on the diagnostic value of ADC texture analysis in PI-RADS 5 lesions

Yan Bai^{1,2}

Chen Jiang Wu^{1,3}

¹The First Affiliated Hospital with Nanjing Medical University, Department of Radiology, Nanjing, China

²The First Affiliated Hospital of Baotou Medical College, Inner Mongolia University of Science and Technology, Department of Radiology, Baotou, China

³Nanjing Medical University School of Medical Imaging, Nanjing, China

Dear Editor,

We appreciate the opportunity to respond to the insightful comments regarding our recently published study. We are grateful for the reader's recognition of the clinical importance of distinguishing between prostate carcinoma and prostatitis in Prostate Imaging Reporting and Data System (PI-RADS) 5 lesions using texture analysis on apparent diffusion coefficient (ADC) maps.

As noted, although our model achieved an impressive accuracy of 96.8% and a sensitivity of 98.7%, we acknowledge that the specificity (60.0%) and negative predictive value (71.6%) indicate significant potential for further refinement. Improving the differentiation of these false-positive "mimickers" remains a priority for urogenital radiology.

Regarding the limitations raised, we would like to provide additional context. First, the challenge of image heterogeneity across different magnetic resonance imaging scanners is a well-known hurdle in radiomics. We consciously avoided traditional preprocessing techniques such as Z-score normalization or muscle-fat referencing, as these can introduce errors through tissue distribution compression or physiological variability.¹ Instead, we prioritized first-order (low-order) texture parameters, such as median and signal coefficient of variation. Research indicates that first-order features demonstrate significantly lower variability and higher robustness across different platforms compared with high-order metrics, especially in ADC images where spatial resolution is limited.² By utilizing consistent field strengths and parameters, we aimed to minimize noise while ensuring the methodology remains practical for clinical application.

Secondly, regarding the reliability of pathological confirmation, all PI-RADS 5 lesions included in this study had a diameter greater than 1.5 cm. Their larger size reduced the risk of missed diagnoses. To ensure reliability, we employed systematic biopsy supplemented by 1–2 targeted cores specifically at the suspicious sites. This combined approach, supported by clinical follow-up, substantially mitigates the risk of missing malignant foci in large lesions.

Third, our region of interest (ROI) delineation workflow involved a junior radiologist's initial segmentation followed by a rigorous audit by two experienced senior radiologists. We believe this consensus review model is often more representative of real-world clinical workflows than independent double-segmentation. It ensures a high level of expertise in the final ROI while maintaining feasibility in a high-volume clinical environment. Looking forward, the integration of artificial intelligence-assisted segmentation tools will likely standardize this process further.³

Finally, we agree that the development of more robust machine learning models and the execution of large-scale, multi-center trials are essential steps to validate these findings. We remain committed to refining these diagnostic tools to reduce unnecessary biopsies and improve patient outcomes. We thank the reader again for their constructive feedback and for contributing to the advancement of prostate imaging.

Handling editor: Sonay Aydin

Corresponding author: Chen Jiang Wu

E-mail: njmu_wcj@163.com

Received 16 January 2026; accepted 21 January 2026.



Epub: 04.02.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263881

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328-338. [\[Crossref\]](#)
2. Gourtsoyianni S, Doumou G, Prezzi D, et al. Primary rectal cancer: repeatability of global and local-regional MR imaging texture features. *Radiology*. 2017;284(2):552-561. [\[Crossref\]](#)
3. Boellaard TN, van Erck R, van der Graaf SH, et al. Comparing AI and manual segmentation of prostate MRI: towards AI-driven 3D-model-guided prostatectomy. *Diagnostics*. 2025;15(9):1141. [\[Crossref\]](#)



Human–AI interaction and collaboration in radiology: from conceptual frameworks to responsible implementation

Burak Kocak¹
 Renato Cuocolo²

¹Basakşehir Cam and Sakura City Hospital,
Department of Radiology, Istanbul, Türkiye

²Department of Medicine, Surgery and Dentistry,
University of Salerno, Baronissi, Italy

Artificial intelligence (AI) is entering routine radiology practice, but most studies evaluate algorithms in isolation rather than their interaction with radiologists in clinical workflows. This narrative review summarizes current knowledge on human–AI interaction in radiology and highlights practical risks and opportunities for clinical teams. First, simple conceptual models of human–AI collaboration are described, such as diagnostic complementarity, which explain when radiologists and AI can achieve synergistic performance exceeding that of either alone. Then, AI tool integration strategies along the imaging pathway are reviewed, from acquisition and triage to interpretation, reporting, and teaching, outlining common interaction models and physician-in-the-loop workflows. Cognitive and professional effects of AI integration are also discussed, including automation bias, algorithmic aversion, deskilling, workload management, and burnout, with specific vulnerabilities for trainees. Furthermore, key elements of responsible implementation are summarized, such as liability and oversight implications, continuous monitoring for performance drift, usable explanations, basic AI literacy, and co-design with radiology teams. Finally, emerging systems are introduced, including vision–language models and adaptive learning loops. This review aims to provide a clear and accessible overview to help the radiology community recognize where human–AI collaboration can add value, where it can cause harm, and which questions future studies must address.

KEYWORDS

Diagnostic imaging, artificial intelligence, human–AI interaction, clinical decision support, radiology workflow, responsible artificial intelligence, medical ethics

Artificial intelligence (AI), defined as the development of methods enabling machines to perform tasks that historically required human intelligence, is considered a revolutionary development in healthcare, particularly in diagnostic imaging, and has the potential to transform the medical imaging profession.^{1–3} Early perspectives on AI adoption in radiology were often characterized by automation anxiety, driven by impressive demonstrations of algorithmic performance that led some to speculate about the potential substitution of human practitioners.^{4,5} However, this narrative has matured, and a strong argument has emerged that AI's role is not to supplant human expertise but rather to function in a human–AI symbiosis as a cognitive partner.^{6,7}

The prevailing professional viewpoint is that AI should serve as a complementary assistive tool, augmenting human intelligence in the diagnostic process.^{5,7–9} The imaging workforce, including both radiologists and radiation technologists, demonstrates generally positive reception and optimism regarding the potential of AI.^{10–13} This favorable outlook is motivated by the expected advantages of AI, including improved efficiency, reduced workload, and optimized management of clinical practice.^{14–16} The primary goal is to establish a predominantly assistive and collaborative symbiotic relationship between humans and AI systems, yielding collective performance that exceeds what either could achieve alone.⁶ Realizing these clinical benefits requires the imaging workforce to adapt to and actively collaborate with such systems.

Handling editor: Tuğba Akıncı D'Antonoli

Corresponding author: Burak Kocak

E-mail: drburakkocak@gmail.com

Received 29 September 2025; revision requested 21
December 2025; accepted 11 January 2026.



Epub: 26.01.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263780

You may cite this article as: Kocak B, Cuocolo R. Human–AI interaction and collaboration in radiology: from conceptual frameworks to responsible implementation. *Diagn Interv Radiol*. 2026;32(5):535–547.

This narrative review first investigates the conceptual frameworks guiding effective human–AI collaboration, followed by an examination of the practical dynamics of integrating AI into clinical radiology workflows. Subsequently, it considers the resulting cognitive and professional impacts, details the required governance and ethical safeguards, and concludes by exploring emerging technological trajectories and proposing directions for future research.

This narrative review is based on targeted, topic-driven searches of the literature, complemented by expert knowledge and reference chaining (backward and forward), and is not designed to follow a formal systematic search strategy.

Conceptual frameworks for human–AI collaboration

The strategic adoption of AI systems in medical imaging relies on established conceptual models that clarify the optimal nature of human–AI cooperation. The key concepts used to describe human–AI interaction and collaboration in this review are summarized in Table 1.

The foundational principle guiding this interaction is diagnostic complementarity.¹⁷ This concept posits that combining two distinct interpretive agents, the human radiologist and the AI system, results in an overall diagnostic performance superior to that of either component acting alone. Importantly,

complementarity does not imply continuous or unstructured decision fusion; rather, it reflects the effective alignment of distinct strengths through clearly defined roles within the diagnostic workflow. This synergy, often described as human–AI symbiosis, stems from inherent differences in strengths and vulnerabilities.⁶ AI systems, particularly those using advanced algorithms such as neural networks, can identify subtle patterns and anomalies in medical images, efficiently perform repetitive, high-volume tasks at scale, and dramatically speed up the process of image interpretation.^{18,19} Conversely, human radiologists provide critical clinical context, common sense, intuition, and medical judgment, which are essential for synthesizing findings into meaningful patient care narratives.^{20–25} Because the errors made by human readers (often perceptual or related to fatigue or distraction) and those made by AI systems (often related to generalization or contextual limitations) only partially overlap,^{6,17} well-structured role separation may allow each to compensate for the other’s limitations, whereas poorly specified interaction can undermine this advantage.^{26,27}

AI is implemented across a spectrum of involvement, ranging from low-autonomy augmentative systems,²⁸ which often use a physician-in-the-loop model,²⁹ to highly autonomous systems that operate with limited direct human oversight.²⁸ In this paper, the term “physician-in-the-loop” is used to

emphasize the radiologist’s diagnostic liability; however, this role exists within a broader human-in-the-loop ecosystem in which technologists provide critical upstream functions, such as image acquisition and quality control.

Within this spectrum, AI can function as a decision support tool by marking suspicious areas or supplying confidence scores to assist the radiologist’s final determination.^{30–32} Alternatively, it can assume the role of an independent second opinion or safety net, flagging potentially overlooked regions to minimize false negatives and ensure quality control.^{33–35} Evidence strongly supports that this interaction model can yield measurable performance improvements, demonstrating the value of collaboration over fully autonomous operation. For instance, Lee et al.³⁶ explored this concept in a retrospective reader study involving 30 radiologists and residents who evaluated 120 chest radiographs (60 containing malignant nodules) in a controlled simulation. They found that a human–AI interaction model can improve performance, although the extent of improvement depends on the quality of the AI system. A conceptual overview of diagnostic complementarity and human–AI team performance is shown in Figure 1.

In mammography screening, different AI workflows are being evaluated to reduce radiologist workload and improve cancer detection.

Table 1. Core concepts in human–AI interaction and collaboration in radiology	
Term	Definition
Human–AI interaction	Observable exchange of information between humans and AI systems during clinical or research tasks.
Human–AI collaboration	Planned sharing of tasks and responsibilities between humans and AI systems within a defined workflow.
Human–AI team performance	Diagnostic performance of the combined human–AI system rather than either component alone.
Diagnostic complementarity	Situation in which human and AI errors only partially overlap, such that their combined decisions are more accurate than either alone.
Human–AI symbiosis	Stable pattern of collaboration in which humans and AI systems adapt to each other’s strengths and limitations over time.
Automation bias	Systematic tendency to accept AI suggestions even when they are incorrect.
Algorithmic aversion	Systematic tendency to ignore AI output because of mistrust, inconsistency, or prior negative experience.
Cognitive offloading	Transfer of tasks such as detection, measurement, or retrieval from humans to AI, which reduces effort but may lead to deskilling.
Trust calibration	Alignment between user trust in an AI system and its actual reliability for a given task and population.
Physician-in-the-loop design	Design principle in which imaging professionals retain oversight and final responsibility for decisions to which AI has contributed.
AI, artificial intelligence.	

Main points
- Human–artificial intelligence (AI) collaboration in radiology is best understood through diagnostic complementarity, where combined performance can exceed that of either the human or the algorithm alone. - Workflow-embedded, physician-in-the-loop designs can help determine AI’s real clinical value more than marginal gains in stand-alone accuracy. - Automation bias, algorithmic aversion, skill erosion, and context-dependent workload effects are central risks that require explicit mitigation, especially for trainees. - Responsible implementation depends on formal governance structures, continuous post-market surveillance, explainability standards, and systematic AI literacy for radiology teams. - Future research should prioritize prospective, multi-institutional evaluation of team performance, equity, and long-term training outcomes over isolated model-centric metrics.

One model is the “AI as supporting reader” workflow, proposed by Ng et al.³⁷ in a large-scale simulation study. The evaluation was based on a retrospective, multisite screening cohort of more than 280,000 mammography examinations from two countries, using multiple vendors, and compared simulated AI-supported workflows with standard human double reading. In this model, the AI acts as the second reader only when it agrees with the first human reader; discordant cases are referred to a second human reader. This simulation indicated that the “AI as supporting reader” workflow could maintain screening performance while substantially reducing the number of cases requiring a second human reading by up to 87%.

A different prospective study evaluated an “AI as additional reader” workflow in mammography screening.³⁸ This prospective implementation was conducted in routine clinical practice across multiple screening sites and evaluated tens of thousands of screening examinations using a commercially deployed AI system. In this implementation, AI was used as a safety net after standard human double reading was complete. If the two human readers agreed not to recall the patient but the AI flagged the case as suspicious, it was referred to an arbitrator for final review. This prospective study found that the additional reader workflow improved the cancer detection rate by 0.7–1.6 additional cancers per 1,000 women screened, with only a minimal increase in additional recalls. The additional cancers detected were primarily invasive (83.3%) and small in size (47.0% were 10 mm or less).

Radiologists, in a study by Zajac et al.,³⁹ envisioned triage workflows in which AI could pre-screen cases, for example, by filtering normal radiographs. This would allow senior radiologists to concentrate their efforts more efficiently on challenging cases with positive findings. In this cross-regional qualitative field study, which involved *in situ* observations and interviews with 18 radiologists across nine clinical sites in Denmark and Kenya, participants conceptualized AI-based case distribution as a potential tool to help clear reading backlogs during periods of high workload. The visions articulated by the participating radiologists focused on AI providing actionable support to help them work better or faster, rather than automating their tasks.

In a recent diagnostic study by the Prostate Imaging–Cancer AI Consortium,⁴⁰ AI assistance was associated with improved

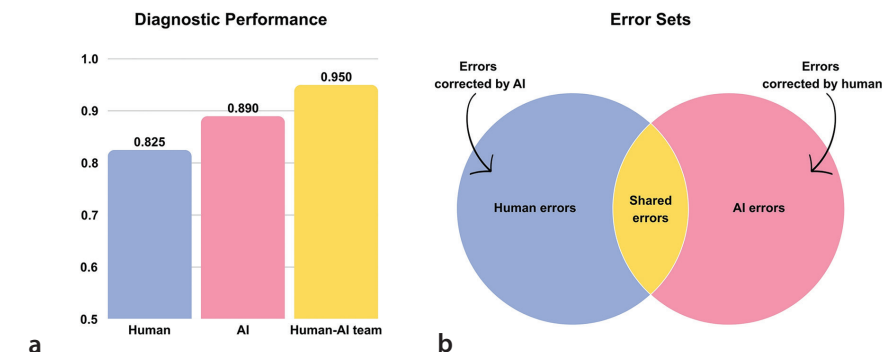


Figure 1. Conceptual model of diagnostic complementarity between radiologists and artificial intelligence (AI). Panel A illustrates that diagnostic performance [e.g., area under the receiver operating characteristic curve (AUC) or sensitivity] can be higher for the human–AI team than for either the radiologist or the AI system alone. Panel B represents the overlap between human and AI error sets; when errors only partially overlap, each can compensate for the other, leading to improved overall performance.

diagnostic accuracy for clinically significant prostate cancer on magnetic resonance imaging (MRI) compared with unassisted readings. The study, which included 61 readers from 53 centers across 17 countries who assessed 360 MRI examinations, found that AI assistance was associated with a statistically significant increase in the area under the receiver operating characteristic curve, from 0.882 to 0.916. At a Prostate Imaging Reporting and Data System threshold of 3 or more, sensitivity improved from 94.3% to 96.8%, and specificity increased from 46.7% to 50.1%. This resulted in three additional true-positive diagnoses and 10 fewer false-positive diagnoses with AI assistance.

Taken together, these studies illustrate both convergent and contrasting approaches to human–AI collaboration that can be organized into a limited number of recurring workflow patterns. Representative workflow architectures for these interaction models are shown in Figure 2, which arranges common diagnostic and screening workflows along a spectrum from low to high AI autonomy. To complement this visual overview, Table 2 presents a workflow-level taxonomy of human–AI collaboration across this autonomy spectrum.

For this partnership to succeed, mutual adaptation is required.^{29,41} The AI system must be iteratively optimized to incorporate human feedback, creating closed-loop learning that aligns algorithmic updates with expert judgment. Active learning frameworks operationalize this process by automatically identifying uncertain or informative cases for labeling by radiologists, enabling continuous model refinement while minimizing annotation burden.^{29,42,43} Conversely, radiologists must adapt their own practice by developing AI literacy to interpret model out-

puts, recognize failure modes, and calibrate trust according to task and context.^{6,44} This bidirectional adaptation fosters appropriate reliance, preventing both overtrust and algorithmic aversion and ensuring that human oversight remains central as systems evolve.

The considerations above can be operationalized as a set of design principles for human–AI collaboration in radiology, as outlined in Table 3.

Practical integration and workflow dynamics

Translating complementarity into clinical action requires strategic placement of AI within existing, often fast-paced operational workflows. AI systems interface with radiology at various stages, including image acquisition, triage, worklist prioritization, interpretation, and final reporting.^{45–48} The corresponding stages, along with human and AI tasks, are depicted in Figure 3.

AI can fundamentally affect the selection and prioritization of cases, for example, by providing prioritization cues or alerts through screening incoming studies for time-sensitive or high-suspicion findings, such as intracranial hemorrhage, pulmonary emboli, or pneumothorax.^{49–52} This AI-based triage has demonstrated substantial reductions in report turnaround times for critical cases. For example, one simulation showed that AI prioritization reduced the average reporting delay for critical chest radiograph findings from 11.2 to 2.7 days.⁵³ Similarly, another simulation study found that AI significantly reduced the average report turnaround time for critical chest X-ray findings (e.g., pneumothorax, from 80.1 to 35.6 min), although it also noted that the maximum report turnaround time could increase without specific safeguards.⁵¹ In contrast to

simulations, a real-world clinical deployment found that, for computed tomography pulmonary angiography examinations positive for pulmonary embolism, AI reprioritization significantly shortened the mean report turnaround time, from 59.9 to 47.6 min.⁵⁴ Furthermore, based on qualitative field studies, radiologists envision that future AI systems could be configured to route cases by user expertise, for example, by directing studies with positive findings to senior radiologists while filtering normal studies for junior radiologists to verify.³⁹

During image interpretation, AI can function as an interactive reporting assistant by automating highly manual tasks.⁴⁸ For instance, AI can perform time-consuming measurements or calculations.^{55,56} Additionally, AI can expedite the retrieval of historical examinations and automatically compare changes over time, highlighting relevant progression or new developments for the radiologist's attention.⁵⁷ Emerging vision-language models promise further assistance, enabling capabilities such as draft report generation and automated summarization of patient imaging history, highlighting key events from prior imaging records.^{58,59} Early expert-rated vision-language model systems (e.g., Flamingo-CXR) produced clinically acceptable draft reports under constrained conditions, although rigorous evaluation and guardrails remain essential.⁵⁹

Successful integration requires configurable AI tools to accommodate the varied needs of different clinical sites, local resources, and user expertise.^{39,60-62} Interface design must prioritize utility and efficiency, ensuring seamless workflow integration to minimize cognitive disruption.^{58,63} Clinicians generally prefer AI to be deployed as tool-based interactions for specific, functional tasks (e.g., quantification or data retrieval) rather than as open-ended, generalized conversational agents.⁵⁸

Maintaining clinical authority requires that final radiological supervision remain an indispensable component of all AI-supported activities.^{11,41,64} This human oversight is imperative for managing medicolegal liabilities and ensuring patient safety. Therefore, AI systems must be designed with explicit mechanisms for manual oversight and decision arbitration.⁴¹ For instance, in sophisticated prostate MRI protocols, AI may highlight suspicious lesions or provide risk scores for cancer detection; however, the human radiologist must retain the ultimate authority to validate or override these AI predictions

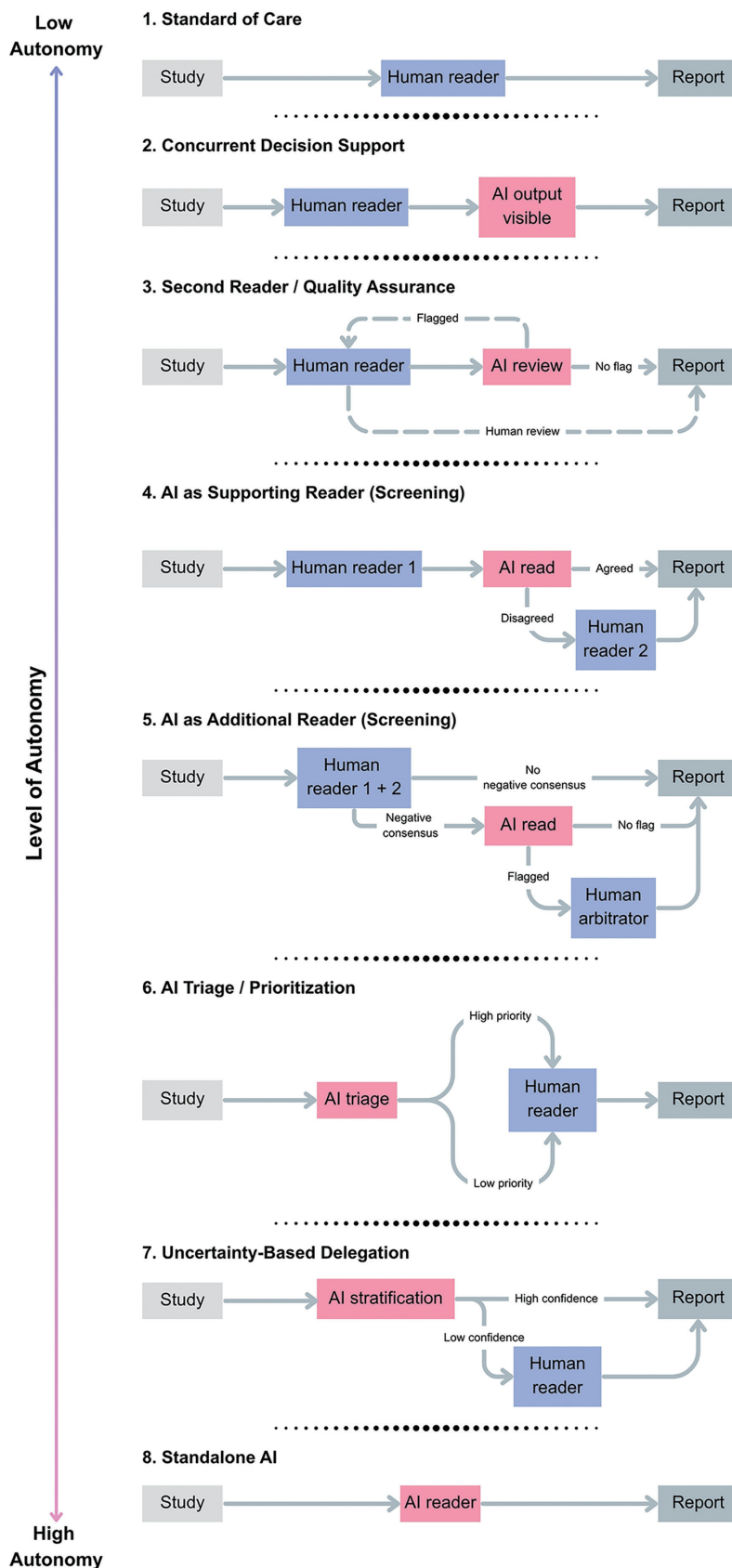


Figure 2. Oversimplified representative workflow architectures for human-artificial intelligence (AI) collaboration across diagnostic and screening tasks. Eight interaction models are shown, ranging from standard human reading to fully autonomous AI. The diagrams illustrate how AI output is incorporated as concurrent support, a second or additional reader, triage, or uncertainty-based delegation and how the need for human oversight changes as AI autonomy increases. These models primarily depict workflows for task-specific, narrow AI, such as detection and triage systems.

Table 2. Workflow-level taxonomy of human–AI collaboration in radiology, organized along the autonomy spectrum from human-only interpretation to conceptual autonomous systems

Workflow model	Human role	AI role	Output	Autonomy level
Standard of care	Radiologist performs the full primary and final interpretation	No AI involvement	Radiologist-authored report without AI involvement	Human only
Concurrent decision support	Radiologist performs the primary interpretation and integrates or ignores AI suggestions	Provides marks, scores, measurements, region proposals, or draft text during interpretation	Radiologist-authored report optionally informed by AI assistance	Low
Second reader and quality assurance	Radiologist performs the primary interpretation and reviews only AI-flagged discrepancies	Performs after-the-fact discrepancy checks, including image–report consistency	Radiologist-revised report incorporating AI-flagged discrepancies, when present	Low to moderate
AI as a supporting reader (screening)	Human reader 1 performs the primary interpretation; human reader 2 arbitrates discordant human–AI cases	Acts as a second reader only when concordant with human reader 1 and suppresses discordant outputs	Human-determined screening outcome after resolving human–AI agreement or disagreement	Moderate
AI as an additional reader (screening)	Two human readers perform primary and secondary interpretations; an arbitrator reviews AI-flagged cases	Acts as a safety-net reader for negative-consensus cases and flags suspicious findings	Screening outcome enhanced by AI safety-net detection and human arbitration, when flagged	Moderate
AI triage and worklist prioritization	Radiologist performs the primary interpretation in an AI-prioritized order	Screens studies and assigns priority, alerts, or routes by expertise or urgency	Radiologist report issued with improved timeliness for high-priority cases	Moderate to high
Uncertainty-aware delegation and high-confidence filtering	Radiologist interprets only escalated cases and audits a sample of auto-finalized normal cases	Stratifies cases by uncertainty, auto-finalizes normal cases, and escalates uncertain or positive cases	Combined output comprising AI-reported normal cases and radiologist-reported escalated cases	High
Standalone AI reporting	No radiologist involvement at the case level; oversight is limited to governance	Fully autonomous interpretation and reporting without case-level human input	AI-generated report without human interpretation (conceptual; not currently clinically viable)	Autonomous

AI: Artificial intelligence.

Table 3. Design principles for human–AI collaboration in radiology

Domain	Design principle	Practical implication for interaction
Task allocation	Define explicit roles for humans and AI for each task	Specify whether AI acts as a triage aid, second reader, quantification tool, or reporting assistant and which responsibilities are left exclusively to human
Workflow placement	Integrate AI at clearly defined points in the workflow	Decide when users first see AI output (e.g., after the initial read vs. upfront), and standardize this across readers
Mode of interaction	Choose simple, task-focused interaction modes	Prefer focused outputs (marks, scores, and structured suggestions) over open-ended conversational use for routine diagnosis
Trust calibration	Support appropriate, rather than maximal, reliance on AI	Use confidence estimates, performance summaries, and feedback mechanisms to align user trust with actual model reliability
Learning and feedback	Allow bidirectional learning between humans and AI	Enable users to flag problematic cases and use these for periodic model review and, where applicable, model updating
Training and supervision	Protect independent human reasoning, especially in trainees	Include AI-free reading periods and formal instruction on how to interpret, question, and, when necessary, override AI output

AI: Artificial intelligence.

based on contextual clinical information.⁶⁵ Similarly, in highly automated triage systems, although clear-cut cases may be filtered, all borderline or ambiguous examinations must be deferred to expert radiologists for final arbitration.^{6,44,65}

Furthermore, AI can be leveraged for internal quality assurance without imposing undue cognitive load. A practical example involves using natural language processing to compare the radiologist's transcribed

report with image detection findings, triggering an alert only when a discrepancy or missed finding is identified.³⁹ This approach provides seamless quality control by engaging the radiologist only when a potential error is detected, thereby avoiding workflow disruption in routine cases. However, the usefulness of such systems depends heavily on their performance, as excessive false positives or false negatives can undermine trust and disrupt workflow.

Table 4 summarizes the characteristic interaction patterns, including their main advantages and risks, across commonly used human–AI collaboration configurations.

Cognitive and professional impacts

Integrating AI into clinical radiology fundamentally alters the cognitive processes, trust, and diagnostic reasoning of human practitioners. These complex dynamics introduce specific risks related to human behavior, such as overreliance and interaction mismatches, for example, when workflow design or model context diverges from clinical reality.

A major behavioral concern is the management of automation bias, which refers to the uncritical acceptance of automated outputs, even when those outputs are incorrect (Figure 4).⁶⁶⁻⁶⁸ This bias represents a considerable risk, particularly for users with lower subject matter expertise, such as radiology trainees, who are more prone to accepting AI recommendations.^{66,67} For example, in a controlled mammography reader study, trainees’ diagnostic accuracy dropped from approximately 80% when the AI was correct to 20% when misled by incorrect AI output. The same study also found that when AI incorrectly suggested downgrading a correct finding (an error of omission), all experience levels, including experts, were equally susceptible to this bias.⁶⁷

Studies show that the accuracy of the AI tool is a key factor. High diagnostic accuracy in an AI model is associated with improved radiologist performance and increased susceptibility to accepting AI suggestions.³⁶ Conversely, flawed AI output can be detrimental. Randomized evidence shows that systematically biased AI predictions significantly reduced clinician diag-

nostic accuracy by approximately 9.1–11.3 percentage points compared with baseline performance.⁶⁹ This negative impact persisted even when explanations were provided; the study found that AI-generated explanations did not mitigate the adverse effects of biased predictions. This aligns with separate findings that interpretability methods can be flawed, as saliency maps in medical imaging may localize pathology poorly, with one study reporting localization utility, measured

as the area under the precision–recall curve, as low as 0.024–0.16 for some methods on chest radiographs.⁷⁰

The opposite challenge, algorithmic aversion (mistrust) (Figure 4), can also impede successful integration.⁷¹ A lack of trust may arise from perceived inconsistencies in AI tools’ technical performance or concerns about frequent false positives or false negatives, which impose an additional burden on radiologists who must double-check AI

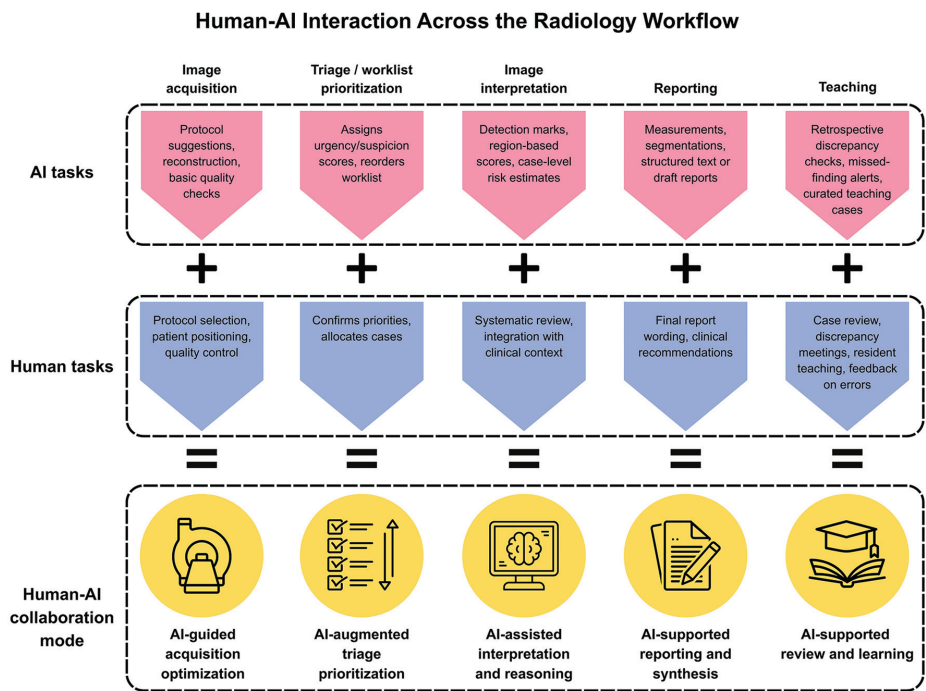


Figure 3. Human–artificial intelligence (AI) interaction across the radiology workflow. The figure illustrates how AI and radiologists can contribute complementary tasks from acquisition to teaching. At each workflow stage, AI provides technical or analytic functions (top row), radiologists perform contextual and clinical judgment tasks (middle row), and their combination yields specific collaboration modes.

Table 4. Key interaction characteristics of common human–AI collaboration configurations in radiology, summarizing core features, operational advantages, and associated risks			
Interaction model	Key interaction features	Main advantages	Main risks
AI-assisted triage and worklist prioritization	One-way interaction; AI output is typically seen before image review	Potentially shorter time to report for urgent findings; improved focus on high-risk cases	Risk of over-reliance on triage flags, particularly when AI confidence is not transparently communicated, and possible under-attention to low-priority studies
Concurrent decision support (marks and scores shown during reading)	Continuous visual interaction during reading; strong influence on search patterns	Potential increase in sensitivity; support for less experienced readers	Automation bias, anchoring on AI marks, and altered search behavior
Second reader and quality assurance mode	Sequential interaction; AI output is revealed after the initial judgment	Likely reduction in false negatives; explicit safety net with lower automation bias	Additional time for review and risk of alert fatigue if the false-positive rate is high
Quantification and structured reporting assistant	Task-focused interaction; users edit or overwrite AI outputs	Time savings for repetitive tasks; more consistent quantification and terminology	Cognitive offloading of basic skills, especially with prolonged, uncritical use, and potential propagation of systematic AI errors
Interactive teaching and feedback tools	Used in dedicated training sessions; not directly linked to the clinical workflow	Supports skill development and reflection on errors; may improve AI literacy	If uncalibrated, risk of conveying misleading confidence or incorrect “ground truth”
AI, artificial intelligence.			

interpretations.¹⁰ In a 2024 EuroAIM/EuSoMIL survey, 47.2% of respondents anticipated an increased total reporting workload; other major barriers cited were costs (49.5%), legal issues (43.7%), and lack of validation (35.5%).¹² This aversion may also be linked to ethical concerns regarding AI's vulnerability to bias and discrimination.^{10,71-73} Research has demonstrated that algorithmic bias can lead to unequal diagnostic performance across patient subgroups, thereby undermining clinician confidence. For example, a large study showed that deep learning chest radiograph models trained on heterogeneous hospital data systematically underdiagnosed disease in female and Black patients compared with other groups.⁷⁴ In a large international survey of 1,041 radiologists and residents, 37% of respondents cited a "lack of trust in AI by stakeholders" as a hurdle to implementation; this view was independently and significantly more often observed among those working outside Europe (adjusted odds ratio: 1.77; 95% confidence interval: 1.24–2.53; $P = 0.002$).⁷⁵

These behaviors can be conceptualized as a spectrum of reliance on AI, ranging from algorithmic aversion to automation bias, with optimal performance occurring at intermediate, calibrated reliance (Figure 5). It should also be noted that users may experience different biases over time when using AI tools in clinical practice. For example, an expert radiologist's algorithmic aversion may give way to automation bias as familiarity and confidence in the AI medical device increase or as the user learns to rely excessively on model output while losing confidence in his or her own reading skills.

The widespread deployment of AI detection tools may accelerate deskilling and hinder skill acquisition, particularly in training contexts.^{49,76} The introduction of tools designed to detect focal abnormalities, such as pulmonary nodules, intracranial hemorrhage, and pneumothorax, may unintentionally disrupt the training process required to acquire fundamental perceptual skills and efficient search patterns.⁴⁹ Furthermore, reader studies have shown that incorrect AI prompts can modify diagnostic judgment.⁶⁷ For instance, when AI automatically triages studies containing emergent findings to the top of a worklist, the immediate alert deprives the trainee of the opportunity to conduct an initial blinded evaluation of the imaging.⁴⁹ Beyond perceptual skills, AI systems that offer contextual diagnostic suggestions or provide automated scores for standardized reporting systems could compromise

mastery of foundational knowledge, such as learning complex differential diagnoses. If trainees become overly dependent on AI to identify findings, they may become more susceptible to automation bias, as studies show that inexperienced users are more vulnerable to following incorrect automated suggestions.^{67,77}

AI integration introduces an additional layer of cognitive workload and digital fatigue, contrary to expectations that it would uniformly reduce labor.⁶³ A nationwide survey of 6,726 radiologists from 1,143 hospitals in China demonstrated a dose–response relationship between the frequency of AI use for image interpretation and work-related emotional exhaustion and burnout.^{78,79} This

unexpected burden is partly attributable to the effort required to review and dismiss frequent false-positive detections, a problem already well recognized with mammography computer-aided detection systems and now echoed in newer AI tools.^{63,80} In a prospective study of 18,680 chest radiographs, AI use reduced overall reading time (13.3 s vs. 14.8 s) and clearly shortened reading times for studies without AI-detected abnormalities (10.8 s vs. 13.1 s). However, for cases with AI-detected abnormalities, reading times did not differ significantly (18.6 s vs. 18.4 s) and increased more steeply as abnormality scores rose.⁸¹ Taken together, these findings show that the workload impact of AI in radiology is highly context-dependent, being beneficial for some tasks, such as chest X-ray screening, but

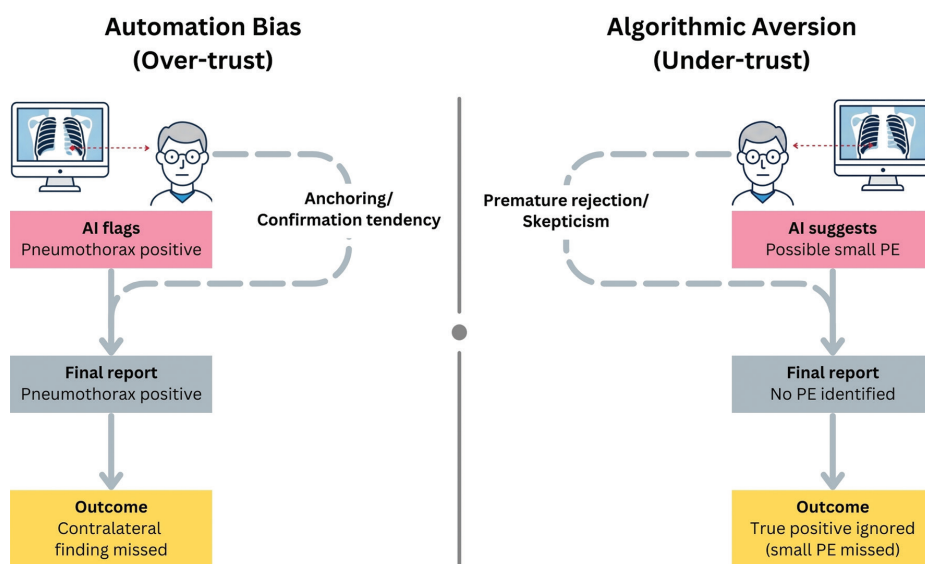


Figure 4. Illustration of automation bias and algorithmic aversion in radiology. The left panel shows automation bias, in which an artificial intelligence (AI)-generated pneumothorax flag anchors the radiologist's interpretation, leading to confirmation of the AI output and omission of a contralateral abnormality. The right panel shows algorithmic aversion, in which an AI suggestion of a small pulmonary embolism is prematurely dismissed because of prior negative experiences or skepticism, resulting in failure to identify a true-positive finding. Both extremes of trust distort independent image assessment and propagate avoidable diagnostic error. PE, pulmonary embolism.

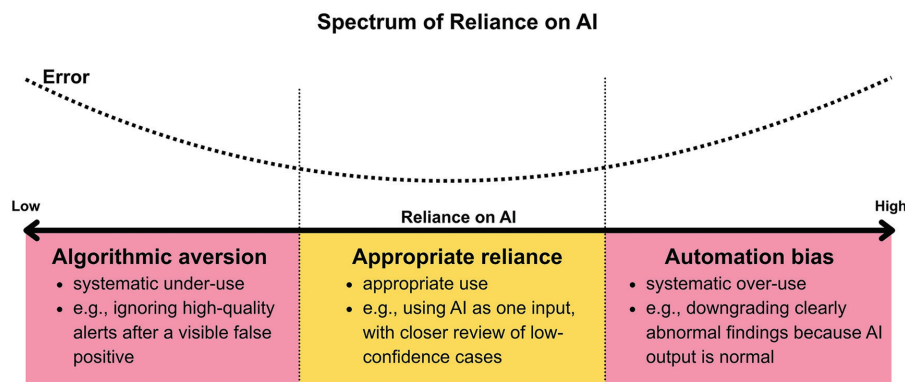


Figure 5. Spectrum of reliance on artificial intelligence (AI). The horizontal axis illustrates levels of reliance on AI, ranging from algorithmic aversion (systematic underuse) through appropriate reliance to automation bias (systematic overuse). The dashed curve indicates that error rates are expected to be higher at both extremes and lowest with appropriate reliance.

potentially burden-increasing in high-complexity settings or under conditions of high workload and low AI acceptance.

Survey data indicate that attitudes toward AI often differ by age and experience, although the patterns vary. One Italian survey found a U-shaped relationship, in which the youngest (< 30 years) and oldest (> 60 years) radiologists were the most optimistic, whereas a large international survey found that younger age was a positive predictor of a proactive attitude toward AI.^{11,13} Younger radiologists and residents frequently report feeling inadequately informed about AI. The survey of Italian radiologists found that 46% of younger members shared this sentiment.¹¹ A Singaporean survey found that a majority (64.8%) of residents and faculty described themselves as novices in their understanding of AI/ML and 59.2% of respondents felt that their residency programs had not adequately implemented AI or machine learning (ML) teaching, despite strong interest in the topic.^{82,83} This perceived gap in AI literacy is considered a factor inhibiting adoption; surveys suggest that limited AI knowledge is associated with fear of replacement, whereas intermediate to advanced knowledge correlates with a more positive attitude toward AI.^{11,13}

Professionals generally recognize that AI will necessitate an expansion of their roles, as evidenced by a 2024 EuroAIM/EuSoMIL survey of 572 European Society of Radiology (ESR) members, in which 98% agreed that radiology teams should participate in the development and validation of AI tools, and 45% stated that radiologists should retain full responsibility for AI outputs influencing clinical decisions.¹² However, skepticism remains regarding the delegation of high-risk functions, such as prognostication or complex treatment decisions, to AI. Surveys repeatedly show that radiologists favor AI as a second reader or workflow aid and insist that final image interpretation and clinical supervision remain their nondelegable responsibility.^{11,13,75,83} Accordingly, several education and human-computer interaction studies have warned that if core interpretive and reporting tasks become heavily automated, AI may contribute to progressive deskilling of radiologists unless training and system design explicitly safeguard independent perceptual and decision-making skills.⁸⁴⁻⁸⁶

Governance, ethics, and responsible implementation

The complex behavioral and cognitive challenges posed by AI necessitate strict

systemic responses, robust governance, and continuous oversight to ensure safe and responsible adoption.^{73,87,88}

A critical concern in scenarios involving AI assistance is medicolegal liability for errors arising from joint human-AI decisions.^{41,89-91} There is currently no transfer of liability to AI systems as long as the radiologist or clinician makes the final decision.⁹²⁻⁹⁵ Globally, experts affirm that final assessment and supervision of AI results by the radiologist are essential for managing legal risks and ensuring patient safety.^{11,88,96-98} Regulators, including the US Food and Drug Administration (FDA), treat AI-based diagnostic tools as medical devices whose potential harms include increased false-positive and false-negative rates and other incorrect outputs that can delay or misdirect care, leading to patient harm.^{41,99} Therefore, institutional governance bodies must establish safeguards to prevent patient harm, especially when deploying high-risk applications, such as screening tools for healthy populations or, in the future, models that might evolve toward treatment support roles.^{87,93} Institutional policies must also address ethical considerations, such as patient consent and the potential misuse of data for other purposes.^{73,87,91,97}

Effective AI implementation requires formal governance structures to guide the entire life cycle of clinical AI, encompassing evaluation, procurement, and ongoing support.²⁸ Daye et al.⁴¹ describe radiology-led, enterprise-level, and hybrid AI governance committees that oversee the selection, implementation, and continuous monitoring of imaging AI tools within large health systems. These governing bodies should be interdisciplinary, integrating clinical, technical, and governance expertise, including ethics and regulatory perspectives, as recommended by radiology AI governance frameworks that emphasize multidisciplinary teams and shared decision-making.^{41,100} Importantly, for such shared decision-making to be meaningful, governance models must also clearly delineate professional accountability for AI-informed clinical actions, as influence without responsibility risks undermining trust and patient safety.

Before any AI tool is deployed, it must undergo a rigorous assessment covering its clinical value, efficacy (benchmarked against average radiologist performance), technical readiness, and ethical implications.^{41,93,101-103} Multiple radiology-specific evaluation frameworks now formalize these dimensions, including the methodological guide by Park et

al.,¹⁰³ the ECLAIR guidelines for commercial tools,¹⁰⁴ and the RADAR deployment and assessment rubric.¹⁰⁵ Implementation should ideally follow a phased approach, beginning with shadow deployment, in which the AI runs in the background without influencing reports, followed by tightly scoped pilot deployment before full rollout.^{105,106}

Continuous post-market surveillance and monitoring are crucial for detecting performance degradation or drift after implementation, which can occur due to changes in patient populations, disease prevalence, acquisition protocols, or data pipelines.^{71,107-111} A robust monitoring plan must track established metrics and include mechanisms for early intervention if performance declines, as emphasized in radiology-specific monitoring frameworks and quality assurance proposals.^{105,112,113} Recently, the ESR published consensus recommendations clarifying that although legal responsibility for post-market surveillance lies with software providers, radiologists (acting as clinical deployers) are expected to actively contribute to the ongoing monitoring of AI safety and performance in routine practice, including output oversight, incident reporting, and structured clinical feedback.¹¹⁴ For algorithms designed for continuous learning, adherence to regulatory guidelines—including a Predetermined Change Control Plan for anticipated updates—is critical, as reflected in recent FDA guidance for AI/ML-enabled medical devices and in radiology AI governance statements.^{28,29,115,116} Collaboration among radiologists, AI scientists, and information technology staff is necessary for continuous quality control, as real-world implementation studies consistently show that sustained AI performance depends on this joint clinical-technical oversight.^{41,106,112}

To cultivate appropriate trust and counter bias, AI tools must provide transparency.¹¹⁷⁻¹¹⁹ Explainable AI systems aim to offer interpretive assistance; however, current user-level explanation tools, such as saliency maps (heat maps), have repeatedly been shown to be unstable and only weakly aligned with radiologists' localization needs, making them insufficient as the primary interface for human-AI interaction.^{70,120-122} Poorly articulated or non-sensical explanations can erode trust, whereas clear explanations aligned with established clinical reasoning may increase trust.¹²³⁻¹²⁷ Similarly, unreliable explanations may promote algorithmic aversion, whereas overly persuasive ones may increase automation bias, illustrating how explainability, trust, and user behavior are closely intercon-

nected. Beyond post hoc explanations, clinically useful AI systems should expose calibrated confidence or uncertainty estimates so that radiologists can preferentially scrutinize low-confidence cases and more readily detect potential AI errors.¹²⁸⁻¹³¹

AI literacy, defined as the competency to critically evaluate and collaborate with AI systems, remains a major barrier to safe integration and is also mandatory for deployers within the European Union under the AI Act.^{10,132,133} To address this, education must begin at the undergraduate level, as radiologists and radiation technologists—including residents and bachelor-level graduates—express a strong desire to enhance their AI and ML knowledge for practice improvement.^{10,83,134} Educational frameworks should be stratified by role (e.g., foundational, clinical user, and expert) to enable professionals to understand algorithmic principles and safety concerns appropriate to their scope of practice.¹³⁵ Several recent initiatives exemplify this structured approach. For instance, a multi-society collaboration (AAPM, ACR, RSNA, and SIIM) has released a comprehensive syllabus detailing competencies across different personas, from general users to purchasers and developers.¹³⁵ Similarly, practical implementation frameworks have been proposed, including a five-step model for integrating AI curricula into residency programs and condensed workshops focused on foundational literacy rather than technical proficiency, which have been shown to considerably improve resident confidence.^{136,137}

Finally, involving radiologists and radiation technologists in co-design efforts is vital to ensure that AI solutions address genuine clinical needs and integrate seamlessly with existing workflows.^{10,138} This collaboration aims to foster a symbiotic relationship with the technology, ensuring that standardized processes align machine-recommended procedures with professional judgment.

Emerging technologies and future directions

Recent advances in foundation models, particularly vision-language models, have extended the boundaries of human-AI interaction in radiology.¹³⁹⁻¹⁴¹ These systems combine image understanding and language generation to enable functions such as report drafting, segmentation, classification, image retrieval, and longitudinal case summarization.^{59,141-146} However, current vision-language models trained on general data remain limited in domain-specific reason-

ing and often underperform in specialized perception tasks.^{144,145,147,148} Therefore, their immediate utility is expected in constrained, task-specific roles—such as structured summarization, quantitative measurement, and retrieval—rather than open-ended conversational support.^{39,58} Qualitative studies suggest that radiologists and clinicians tend to prefer workflow tools (such as tool buttons or alerts) embedded within their reporting environment over general free-text conversational assistants, citing time constraints and a lack of trust in open-ended chat systems.⁵⁸

Physician-in-the-loop active learning, which facilitates interactive and continuous model improvement (Figure 6), aims to enhance physician-AI interaction and collaboration.²⁹ These frameworks allow radiologists to iteratively refine models through feedback collected during routine practice, with updates performed under predefined change control protocols and independent validation. Such designs support regulatory compliance and improve model adaptability and generalizability, although challenges such as annotation variability must be carefully managed to preserve data integrity.

To further strengthen human-AI collaboration and trust, models should incorporate both uncertainty quantification and transparent reasoning mechanisms.^{119,149-151}

Uncertainty-aware systems can guide role arbitration by allowing AI to handle clear, low-ambiguity cases while deferring complex or equivocal findings to expert review.¹⁴⁹⁻¹⁵¹ For explainability to be effective, it must provide human-centered, decision-relevant feedback; this includes not only visualizations linking predictions to evidential image regions but also calibrated measures of model confidence and uncertainty.^{119,149}

The near-term priority is not the development of larger models but the implementation of effective human-AI collaboration through trustworthy, auditable, and workflow-embedded systems that demonstrably enhance collective diagnostic accuracy, efficiency, and safety. Future research should focus on evaluating team performance metrics, workload implications, and long-term cognitive effects in prospective, multi-institutional settings.

Final remarks

Radiology is no longer debating whether AI will replace radiologists but rather how to structure accountable and effective human-AI partnerships. The evidence reviewed here demonstrates that performance gains are fragile when workflow integration, cognitive effects, and governance are neglected. Robust collaboration requires physician-in-the-loop design, calibrated trust, continuous

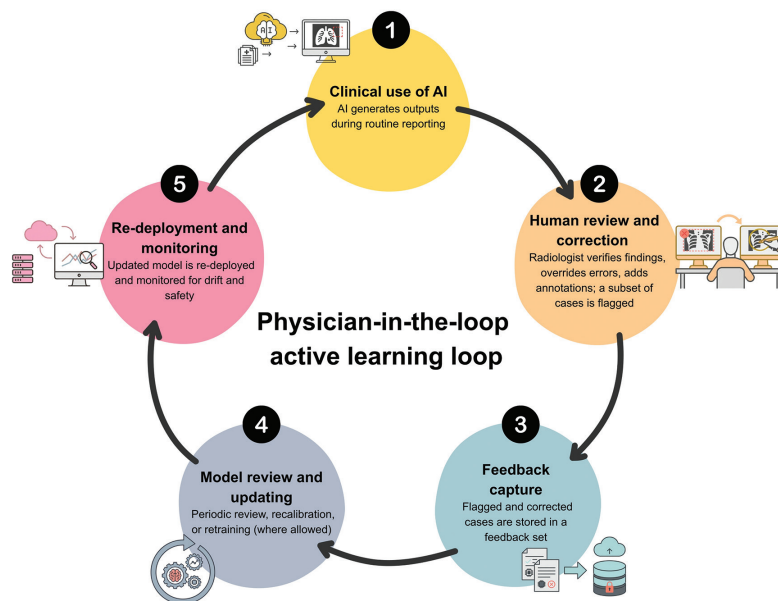


Figure 6. Physician-in-the-loop active learning design. The figure illustrates a cyclical process in which artificial intelligence (AI) is used during routine clinical reporting (step 1). Radiologists review and correct AI outputs as part of normal reporting, but only a subset of cases (e.g., uncertain or discordant results) is explicitly flagged for feedback (step 2). These flagged cases are stored in a feedback set (step 3), which informs periodic model review, recalibration, or retraining, where permitted (step 4), followed by re-deployment and monitoring for drift and safety (step 5). Unlike simple human-in-the-loop oversight at the case level, this loop emphasizes how targeted radiologist input continuously shapes the behavior of deployed AI systems over time while limiting additional workload.

monitoring, and explicit protection of training pathways and professional autonomy. Future work should prioritize prospective, multi-institutional studies of team performance, workload, equity, and long-term learning outcomes rather than isolated accuracy metrics. Under these conditions, AI can evolve from an opportunistic add-on into core clinical infrastructure that strengthens the safety and reliability of imaging care.

Acknowledgment

The language of this manuscript was refined with assistance from ChatGPT (versions 5.2) and Gemini 3 Pro. Only the icons in Figures 4 and 6 were created using Gemini Pro; all other graphics were manually created in Canva.com. All uses of these tools were performed under strict author supervision.

Footnotes

Conflict of interest disclosure

Burak Kocak, MD, previously served as a Section Editor for Diagnostic and Interventional Radiology. He had no involvement in the peer review of this article and had no access to any information related to its peer review.

References

- Moawad AW, Fuentes DT, ElBanan MG, et al. Artificial intelligence in diagnostic radiology: where do we stand, challenges, and opportunities. *J Comput Assist Tomogr*. 2022;46(1):78-90. [\[Crossref\]](#)
- Langlotz CP, Allen B, Erickson BJ, et al. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology*. 2019;291(3):781-791. [\[Crossref\]](#)
- Thrall JH, Li X, Li Q, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *J Am Coll Radiol*. 2018;15(3 Pt B):504-508. [\[Crossref\]](#)
- Gong B, Nugent JP, Guest W, et al. Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: a national survey study. *Acad Radiol*. 2019;26(4):566-577. [\[Crossref\]](#)
- Langlotz CP. Will artificial intelligence replace radiologists? *Radiol Artif Intell*. 2019;1(3):e190058. [\[Crossref\]](#)
- Geftter WB, Prokop M, Seo JB, Raoof S, Langlotz CP, Hatabu H. Human-AI symbiosis: a path forward to improve chest radiography and the role of radiologists in patient care. *Radiology*. 2024;310(1):e232778. [\[Crossref\]](#)
- Taylor-Phillips S, Freeman K. Artificial intelligence to complement rather than replace radiologists in breast screening. *The Lancet Digital Health*. 2022;4(7):e478-e479. [\[Crossref\]](#)
- Langlotz CP. The future of AI and informatics in radiology: 10 predictions. *Radiology*. 2023;309(1):e231114. [\[Crossref\]](#)
- Chen M, Wang Y, Wang Q, et al. Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation. *npj Digit Med*. 2024;7(1):349. [\[Crossref\]](#)
- Arkoh S, Akudjedu TN, Amedu C, Antwi WK, Elshami W, Ohene-Botwe B. Current radiology workforce perspective on the integration of artificial intelligence in clinical practice: a systematic review. *J Med Imaging Radiat Sci*. 2025;56(1):101769. [\[Crossref\]](#)
- Cè M, Ibba S, Cellina M, et al. Radiologists' perceptions on AI integration: an in-depth survey study. *Eur J Radiol*. 2024;177:111590. [\[Crossref\]](#)
- Zanardo M, Visser JJ, Colarieti A, et al. Impact of AI on radiology: a EuroAIM/EuSoMI 2024 survey among members of the European Society of Radiology. *Insights into Imaging*. 2024;15(1):240. [\[Crossref\]](#)
- Huisman M, Ranschaert E, Parker W, et al. An international survey on AI in radiology in 1,041 radiologists and radiology residents part 1: fear of replacement, knowledge, and attitude. *Eur Radiol*. 2021;31(9):7058-7066. [\[Crossref\]](#)
- Eng DK, Khandwala NB, Long J, et al. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: a prospective multicenter randomized controlled trial. *Radiology*. 2021;301(3):692-699. [\[Crossref\]](#)
- Wenderott K, Krups J, Zaruchas F, Weigl M. Effects of artificial intelligence implementation on efficiency in medical imaging—a systematic literature review and meta-analysis. *npj Digit Med*. 2024;7(1):265. [\[Crossref\]](#)
- Sridharan S, Hui ASX, Venkataraman N, et al. Real-world evaluation of an AI triaging system for chest X-rays: a prospective clinical study. *Eur J Radiol*. 2024;181:111783. [\[Crossref\]](#)
- Wagner I, Chakradeo K. Human-AI complementarity in diagnostic radiology: the case of double reading. *Philos Technol*. 2025;38(2):57. [\[Crossref\]](#)
- Khalifa M, Albadawy M. AI in diagnostic imaging: revolutionising accuracy and efficiency. *Computer Methods and Programs in Biomedicine Update*. 2024;5:100146. [\[Crossref\]](#)
- Hwang EJ, Goo JM, Park CM. AI Applications for thoracic imaging: considerations for best practice. *Radiology*. 2025;314(2):e240650. [\[Crossref\]](#)
- Briedis M. Sensing diagnostic images: skilful embodied cognition in oncoradiology. *Sens Imaging*. 2021;23(1):2. [\[Crossref\]](#)
- Friis JKBO. Gestalt descriptions embodiments and medical image interpretation. *AI & Soc*. 2017;32(2):209-218. [\[Crossref\]](#)
- Chen R, Friis JKBO. Vision, body and interpretation in medical imaging diagnostics. *Med Health Care and Philos*. 2024;27(2):253-266. [\[Crossref\]](#)
- Briedis M. Phenomenological ethnography of radiology: expert performance in enacting diagnostic cognition. *Phenom Cogn Sci*. 2020;19(2):373-404. [\[Crossref\]](#)
- Briedis M, Briediene R. Phenomenological analysis of diagnostic radiology: description and relevance to diagnostic errors. *Diagnosis*. 2020;7(3):215-225. [\[Crossref\]](#)
- Neri E, de Souza N, Brady A, et al. What the radiologist should know about artificial intelligence — an ESR white paper. *Insights into Imaging*. 2019;10(1):44. [\[Crossref\]](#)
- Agarwal N, Moehring A, Rajpurkar P, Salz T. Combining human expertise with artificial intelligence: experimental evidence from radiology. *National Bureau of Economic Research*. Preprint posted online July 2023. [\[Crossref\]](#)
- Rajpurkar P, Topol EJ. Beyond Assistance: The case for role separation in AI-human radiology workflows. *Radiology*. 2025;316(1):e250477. [\[Crossref\]](#)
- Brady AP, Allen B, Chong J, et al. Developing, purchasing, implementing and monitoring AI tools in radiology: practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR & RSNA. *Insights into Imaging*. 2024;15(1):16. [\[Crossref\]](#)
- Luo M, Yousefirizi F, Rouzrokh P, et al. Physician-in-the-loop active learning in radiology artificial intelligence workflows: opportunities, challenges, and future directions. *AJR Am J Roentgenol*. 2025;225(5):e2533364. [\[Crossref\]](#)
- Bayerl N, Adams LC, Cavallaro A, et al. Assessment of a fully-automated diagnostic AI software in prostate MRI: clinical evaluation and histopathological correlation. *Eur J Radiol*. 2024;181:111790. [\[Crossref\]](#)
- Caldas FAA, Caldas HC, Henrique T, et al. Evaluating the performance of artificial intelligence and radiologists accuracy in breast cancer detection in screening mammography across breast densities. *European Journal of Radiology Artificial Intelligence*. 2025;2(9310):100013. [\[Crossref\]](#)
- Saha A, Bosma JS, Twilt JJ, et al. Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. *Lancet Oncol*. 2024;25(7):879-887. [\[Crossref\]](#)
- Nakai E, Miyagi Y, Suzuki K, et al. Artificial intelligence as a second reader for screening mammography. *Radiology Advances*. 2024;1(2):umae011. [\[Crossref\]](#)
- van Winkel SL, Peters J, Janssen N, et al. AI as an independent second reader in detection of clinically relevant breast cancers within a population-based screening programme in the Netherlands: a retrospective cohort study. *Lancet Digit Health*. 2025;7(8):100882. [\[Crossref\]](#)

35. Topff L, Steltenpool S, Ranschaert ER, et al. Artificial intelligence-assisted double reading of chest radiographs to detect clinically relevant missed findings: a two-centre evaluation. *Eur Radiol.* 2024;34(9):5876-5885. [\[Crossref\]](#)
36. Lee JH, Hong H, Nam G, Hwang EJ, Park CM. Effect of human-AI interaction on detection of malignant lung nodules on chest radiographs. *Radiology.* 2023;307(5):e222976. [\[Crossref\]](#)
37. Ng AY, Glocker B, Oberije C, et al. Artificial intelligence as supporting reader in breast screening: a novel workflow to preserve quality and reduce workload. *J Breast Imaging.* 2023;5(3):267-276. [\[Crossref\]](#)
38. Ng AY, Oberije CJG, Ambrózay É, et al. Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. *Nat Med.* 2023;29(12):3044-3049. [\[Crossref\]](#)
39. Zając HD, Andersen TO, Kwasa E, et al. Towards clinically useful AI: from radiology practices in Global South and North to visions of AI support. *ACM Trans Comput-Hum Interact.* 2025;32(2):20:1-20:38. [\[Crossref\]](#)
40. Twilt JJ, Saha A, Bosma JS, et al. AI-Assisted vs unassisted identification of prostate cancer in magnetic resonance images. *JAMA Netw Open.* 2025;8(6):e2515672. [\[Crossref\]](#)
41. Daye D, Wiggins WF, Lungren MP, et al. Implementation of clinical artificial intelligence in radiology: who decides and how? *Radiology.* 2022;305(3):555-563. [\[Crossref\]](#)
42. Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med Image Anal.* 2021;71:102062. [\[Crossref\]](#)
43. Ren P, Xiao Y, Chang X, et al. A Survey of Deep Active Learning. *ACM Comput Surv.* 2021;54(9):180:1-180:40. [\[Crossref\]](#)
44. Leibig C, Brehmer M, Bunk S, Byng D, Pinker K, Umutlu L. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. *Lancet Digit Health.* 2022;4(7):e507-e519. [\[Crossref\]](#)
45. Korfiatis P, Kline TL, Meyer HM, et al. Implementing artificial intelligence algorithms in the radiology workflow: challenges and considerations. *Mayo Clinic Proceedings: Digital Health.* 2025;3(1):100188. [\[Crossref\]](#)
46. Tejani AS, Cook TS, Hussain M, Sippel Schmidt T, O'Donnell KP. Integrating and adopting AI in the radiology workflow: a primer for standards and integrating the healthcare enterprise (IHE) profiles. *Radiology.* 2024;311(3):e232653. [\[Crossref\]](#)
47. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics (Basel).* 2023;13(17):2760. [\[Crossref\]](#)
48. Pierre K, Haneberg AG, Kwak S, et al. Applications of artificial intelligence in the radiology roundtrip: process streamlining, workflow optimization, and beyond. *Seminars in Roentgenology.* 2023;58(2):158-169. [\[Crossref\]](#)
49. Li MD, Little BP. Appropriate reliance on artificial intelligence in radiology education. *J Am Coll Radiol.* 2023;20(11):1126-1130. [\[Crossref\]](#)
50. Topff L, Ranschaert ER, Bartels-Rutten A, et al. Artificial intelligence tool for detection and worklist prioritization reduces time to diagnosis of incidental pulmonary embolism at CT. *Radiology: Cardiothoracic Imaging.* 2023;5(2):e220163. [\[Crossref\]](#)
51. Baltruschat I, Steinmeister L, Nickisch H, et al. Smart chest X-ray worklist prioritization using artificial intelligence: a clinical workflow simulation. *Eur Radiol.* 2021;31(6):3837-3845. [\[Crossref\]](#)
52. Cotena M, Ayobi A, Zuchowski C, et al. Enhancing radiologist efficiency with AI: a multi-reader multi-case study on aortic dissection detection and prioritization. *Diagnostics (Basel).* 2024;14(23):2689. [\[Crossref\]](#)
53. Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G. Automated triaging of adult chest radiographs with deep artificial neural networks. *Radiology.* 2019;291(1):196-202. [\[Crossref\]](#)
54. Batra K, Xi Y, Bhagwat S, Espino A, Peshock RM. Radiologist worklist reprioritization using artificial intelligence: impact on report turnaround times for CTPA examinations positive for acute pulmonary embolism. *AJR Am J Roentgenol.* 2023;221(3):324-333. [\[Crossref\]](#)
55. Dundas J, Leipsic JA, Sellers S, et al. Artificial intelligence-based coronary stenosis quantification at coronary CT angiography versus quantitative coronary angiography. *Radiol Cardiothorac Imaging.* 2023;5(6):e230124. [\[Crossref\]](#)
56. Hamelink I, van Tuinen M, Kwee TC, van Ooijen PMA, Vliegthart R. Repeatability of AI-based, automatic measurement of vertebral and cardiovascular imaging biomarkers in low-dose chest CT: the ImaLife cohort. *Eur Radiol.* 2025;35(7):3833-3841. [\[Crossref\]](#)
57. Peters S, Schmill L, Gless CA, Stürner K, Jansen O, Seehafer S. AI-based assessment of longitudinal multiple sclerosis MRI: strengths and weaknesses in clinical practice. *European Journal of Radiology.* 2025;183:111941. [\[Crossref\]](#)
58. Yildirim N, Richardson H, Wetscherek MT, et al. Multimodal healthcare AI: identifying and designing clinically relevant vision-language applications for radiology. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. New York (NY): Association for Computing Machinery; 2024. p. 1-22. [\[Crossref\]](#)
59. Tanno R, Barrett DGT, Sellergren A, et al. Collaboration between clinicians and vision-language models in radiology report generation. *Nat Med.* 2025;31(2):599-608. [\[Crossref\]](#)
60. Alis D, Tanyel T, Meltem E, et al. Choosing the right artificial intelligence solutions for your radiology department: key factors to consider. *Diagn Interv Radiol.* 2024;30(6):357-365. [\[Crossref\]](#)
61. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ.* 2025;388:e081554. [\[Crossref\]](#)
62. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database (Oxford).* 2020;2020:baaa010. [\[Crossref\]](#)
63. Park SH, Langlotz CP. Crucial role of understanding in human-artificial intelligence interaction for successful clinical adoption. *Korean J Radiol.* 2025;26(4):287-290. [\[Crossref\]](#)
64. Chen C. Human-AI decision making with case studies in radiology [thesis]. Chicago (IL): The University of Chicago; 2025 [cited 2026 Jan 22]. Available from: <https://knowledge.uchicago.edu/record/15549>
65. Padhani AR, Papanikolaou N. AI and human interactions in prostate cancer diagnosis using MRI. *Eur Radiol.* 2025;35(9):5695-5700. [\[Crossref\]](#)
66. Kim SH, Schramm S, Riedel EO, et al. Automation bias in AI-assisted detection of cerebral aneurysms on time-of-flight MR angiography. *Radiol Med.* 2025;130(4):555-566. [\[Crossref\]](#)
67. Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology.* 2023;307(4):e222176. [\[Crossref\]](#)
68. Abdelwanis M, Alarafati HK, Tammam MMS, Simsekler MCE. Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis. *J Saf Sci Resil.* 2024;5(4):460-469. [\[Crossref\]](#)
69. Jabbour S, Fouhey D, Shepard S, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. *JAMA.* 2023;330(23):2275-2284. [\[Crossref\]](#)
70. Arun N, Gaw N, Singh P, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol Artif Intell.* 2021;3(6):e200267. [\[Crossref\]](#)
71. Koçak B, Ponsiglione A, Stanzione A, et al. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagn Interv Radiol.* 2025;31(2):75-88. [\[Crossref\]](#)
72. Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and

- recommendations. *Jpn J Radiol.* 2024;42(1):3-15. [\[Crossref\]](#)
73. Goisauf M, Cano Abadía M. Ethics of AI in radiology: a review of ethical and societal implications. *Front Big Data.* 2022;5:850383. [\[Crossref\]](#)
74. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med.* 2021;27(12):2176-2182. [\[Crossref\]](#)
75. Huisman M, Ranschaert E, Parker W, et al. An international survey on AI in radiology in 1041 radiologists and radiology residents part 2: expectations, hurdles to implementation, and education. *Eur Radiol.* 2021;31(11):8797-8806. [\[Crossref\]](#)
76. Macnamara BN, Berber I, Çavuşoğlu MC, et al. Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cogn Res Princ Implic.* 2024;9:46. [\[Crossref\]](#)
77. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc.* 2012;19(1):121-127. [\[Crossref\]](#)
78. Chisholm M, Magudia K. Beyond the AJR: Reevaluating the impact of artificial intelligence on radiologist burnout. *AJR Am J Roentgenol.* 2025;225(3):e2532713. [\[Crossref\]](#)
79. Liu H, Ding N, Li X, et al. Artificial intelligence and radiologist burnout. *JAMA Netw Open.* 2024;7(11):e2448714. [\[Crossref\]](#)
80. Shahrvini T, Wood EJ, Joines MM, et al. Artificial intelligence versus radiologist false positives on digital breast tomosynthesis examinations in a population-based screening program. *AJR Am J Roentgenol.* 2025. [\[Crossref\]](#)
81. Shin HJ, Han K, Ryu L, Kim EK. The impact of artificial intelligence on the reading times of radiologists for chest radiographs. *NPJ Digit Med.* 2023;6(1):82. [\[Crossref\]](#)
82. Salastekar NV, Maxfield C, Hanna TN, Krupinski EA, Heitkamp D, Grimm LJ. Artificial intelligence/machine learning education in radiology: multi-institutional survey of radiology residents in the United States. *Acad Radiol.* 2023;30(7):1481-1487. [\[Crossref\]](#)
83. Ooi SKG, Makmur A, Soon AYQ, et al. Attitudes toward artificial intelligence in radiology with learner needs assessment within radiology residency programmes: a national multi-programme survey. *Singapore Med J.* 2021;62(3):126-134. [\[Crossref\]](#)
84. Tzanis E, Adams LC, Akinci D'Antonoli T, et al. Agentic systems in radiology: principles, opportunities, privacy risks, regulation, and sustainability concerns. *Diagn Interv Imaging.* 2026;107(1):7-16. [\[Crossref\]](#)
85. Natali C, Marconi L, Dias Duran LD, Cabitza F. AI-induced deskilling in medicine: a mixed-method review and research agenda for healthcare and beyond. *Artif Intell Rev.* 2025;58(11):356. [\[Crossref\]](#)
86. Fischetti C, Bhattar P, Frisch E, et al. The evolving importance of artificial intelligence and radiology in medical trainee education. *Acad Radiol.* 2022;29 Suppl 5:S70-S75. [\[Crossref\]](#)
87. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. ISBN: 9789240029200. Available from: <https://www.who.int/publications/i/item/9789240029200>
88. Geis JR, Brady AP, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the joint european and north American multisociety statement. *Radiology.* 2019;293(2):436-440. [\[Crossref\]](#)
89. Mezrich JL. Is artificial intelligence (AI) a pipe dream? Why legal issues present significant hurdles to AI autonomy. *AJR Am J Roentgenol.* 2022;219(1):152-156. [\[Crossref\]](#)
90. Contaldo MT, Pasceri G, Vignati G, Bracchi L, Triggiani S, Carrafiello G. AI in radiology: navigating medical responsibility. *Diagnostics (Basel).* 2024;14(14):1506. [\[Crossref\]](#)
91. Jaremko JL, Azar M, Bromwich R, et al. Canadian Association of Radiologists white paper on ethical and legal issues related to artificial intelligence in radiology. *Can Assoc Radiol J.* 2019;70(2):107-118. [\[Crossref\]](#)
92. Naik N, Hameed BMZ, Shetty DK, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front Surg.* 2022;9:862322. [\[Crossref\]](#)
93. Aldhafeeri FM. Governing artificial intelligence in radiology: a systematic review of ethical, legal, and regulatory frameworks. *Diagnostics (Basel).* 2025;15(18):2300. [\[Crossref\]](#)
94. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. *Front Med (Lausanne).* 2023;10:1305756. [\[Crossref\]](#)
95. Jones C, Thornton J, Wyatt JC. Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. *Med Law Rev.* 2023;31(4):501-520. [\[Crossref\]](#)
96. Jungmann F, Jorg T, Hahn F, et al. Attitudes toward artificial intelligence among radiologists, IT specialists, and industry. *Academic Radiology.* 2021;28(6):834-840. [\[Crossref\]](#)
97. Tang A, Tam R, Cadrin-Chênevert A, et al. Canadian Association of Radiologists white paper on ethical and legal issues related to artificial intelligence in radiology. *Can Assoc Radiol J.* 2018;69(2):120-135. [\[Crossref\]](#)
98. Kohli M, Geis R. Ethics, artificial intelligence, and radiology. *J Am Coll Radiol.* 2018;15(9):1317-1319. [\[Crossref\]](#)
99. Hill DLG. AI in imaging: the regulatory landscape. *Br J Radiol.* 2024;97(1155):483-491. [\[Crossref\]](#)
100. Stogiannos N, Malik R, Kumar A, et al. Black box no more: a scoping review of AI governance frameworks to guide procurement and adoption of AI in medical imaging and radiotherapy in the UK. *Br J Radiol.* 2023;96(1152):20221157. [\[Crossref\]](#)
101. Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health Care Inform.* 2021;28(1):e100251. [\[Crossref\]](#)
102. Pesapane F, Hauglid MK, Fumagalli M, et al. The translation of in-house imaging AI research into a medical device ensuring ethical and regulatory integrity. *Eur J Radiol.* 2025;182. [\[Crossref\]](#)
103. Park SH, Han K, Jang HY, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology.* 2023;306(1):20-31. [\[Crossref\]](#)
104. Omoumi P, Ducarouge A, Tournier A, et al. To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *Eur Radiol.* 2021;31(6):3786-3796. [\[Crossref\]](#)
105. Boverhof BJ, Redekop WK, Bos D, et al. Radiology AI Deployment and Assessment Rubric (RADAR) to bring value-based AI into radiological practice. *Insights into Imaging.* 2024;15(1):34. [\[Crossref\]](#)
106. Bizzo BC, Dasegowda G, Bridge C, et al. Addressing the challenges of implementing artificial intelligence tools in clinical practice: principles from experience. *J Am Coll Radiol.* 2023;20(3):352-360. [\[Crossref\]](#)
107. Kore A, Abbasi Bavi E, Subasri V, et al. Empirical data drift detection experiments on real-world medical imaging data. *Nat Commun.* 2024;15(1):1887. [\[Crossref\]](#)
108. Roschewitz M, Khara G, Yearsley J, et al. Automatic correction of performance drift under acquisition shift in medical image classification. *Nat Commun.* 2023;14(1):6608. [\[Crossref\]](#)
109. Davis SE, Dorn C, Park DJ, Matheny ME. Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability. *J Am Med Inform Assoc.* 2025;32(5):845-854. [\[Crossref\]](#)
110. Feng J, Phillips RV, Malenica I, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *npj Digit Med.* 2022;5(1):66. [\[Crossref\]](#)
111. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. *Br J Radiol.* 2023;96(1150):20220878. [\[Crossref\]](#)
112. Dreyer KJ, Allen B, Wald C. Real-world surveillance of FDA-cleared artificial intelligence models: rationale and logistics. *J Am Coll Radiol.* 2022;19(2 Pt A):274-277. [\[Crossref\]](#)
113. Allen B, Dreyer K, Stibolt R, et al. Evaluation and real-world performance monitoring

- of artificial intelligence models in clinical practice: try it, buy it, check it. *J Am Coll Radiol*. 2021;18(11):1489-1496. [\[Crossref\]](#)
114. Cuocolo R, Bernardini D, Pinto dos Santos D, et al. AI medical device post-market surveillance regulations: consensus recommendations by the European Society of Radiology. *Insights Imaging*. 2025;16(1):275. [\[Crossref\]](#)
 115. Pantanowitz L, Hanna M, Pantanowitz J, et al. Regulatory aspects of artificial intelligence and machine learning. *Modern Pathology*. 2024;37(12):100609. [\[Crossref\]](#)
 116. U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: guidance for industry and Food and Drug Administration staff [Internet]. Silver Spring (MD): FDA; 2025 Aug. [\[Crossref\]](#)
 117. Cutillo CM, Sharma KR, Foschini L, et al. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. *npj Digit Med*. 2020;3:47. [\[Crossref\]](#)
 118. Tun HM, Rahman HA, Naing L, Malik OA. Trust in artificial intelligence-based clinical decision support systems among health care workers: systematic review. *J Med Internet Res*. 2025;27:e69678. [\[Crossref\]](#)
 119. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *npj Digit Med*. 2022;5(1):156. [\[Crossref\]](#)
 120. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745-e750. [\[Crossref\]](#)
 121. E Ihongbe I, Fouad S, F Mahmoud T, Rajasekaran A, Bhatia B. Evaluating Explainable Artificial Intelligence (XAI) techniques in chest radiology imaging through a human-centered Lens. *PLoS One*. 2024;19(10):e0308758. [\[Crossref\]](#)
 122. Zhang J, Chao H, Dasegowda G, Wang G, Kalra MK, Yan P. Revisiting the trustworthiness of saliency methods in radiology AI. *Radiol Artif Intell*. 2024;6(1):e220221. [\[Crossref\]](#)
 123. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655. [\[Crossref\]](#)
 124. Rosenbacke R, Melhus Å, McKee M, Stuckler D. How Explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. *JMIR AI*. 2024;3:e53207. [\[Crossref\]](#)
 125. Wysocki O, Davies JK, Vigo M, et al. Assessing the communication gap between AI models and healthcare professionals: explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*. 2023;316:103839. [\[Crossref\]](#)
 126. Alam L, Mueller S. Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Med Inform Decis Mak*. 2021;21(1):178. [\[Crossref\]](#)
 127. Hur S, Lee Y, Park J, et al. Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour. *npj Digit Med*. 2025;8(1):578. [\[Crossref\]](#)
 128. Singh Y, Hathaway QA, Keishing V, et al. Beyond post hoc explanations: a comprehensive framework for accountable AI in medical imaging through transparency, interpretability, and explainability. *Bioengineering (Basel)*. 2025;12(8):879. [\[Crossref\]](#)
 129. Salvi M, Seoni S, Campagner A, et al. Explainability and uncertainty: two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *Int J Med Inform*. 2025;197:105846. [\[Crossref\]](#)
 130. Faghani S, Moassefi M, Rouzrokh P, et al. Quantifying uncertainty in deep learning of radiologic images. *Radiology*. 2023;308(2):e222217. [\[Crossref\]](#)
 131. Huang L, Ruan S, Xing Y, Feng M. A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods. *Med Image Anal*. 2024;97:103223. [\[Crossref\]](#)
 132. Eltawil FA, Atalla M, Boulos E, Amirabadi A, Tyrrell PN. Analyzing barriers and enablers for the acceptance of artificial intelligence innovations into radiology practice: a scoping review. *Tomography*. 2023;9(4):1443-1455. [\[Crossref\]](#)
 133. Hua D, Petrina N, Young N, Cho JG, Poon SK. Understanding the factors influencing acceptability of AI in medical imaging domains among healthcare professionals: a scoping review. *Artif Intell Med*. 2024;147:102698. [\[Crossref\]](#)
 134. Waymel Q, Badr S, Demondion X, Cotten A, Jacques T. Impact of the rise of artificial intelligence in radiology: what do radiologists think? *Diagn Interv Imaging*. 2019;100(6):327-336. [\[Crossref\]](#)
 135. Kitamura F, Kline T, Warren D, et al. Teaching AI for radiology applications: a multisociety-recommended syllabus from the AAPM, ACR, RSNA, and SIIM. *J Imaging Inform Med*. 2025. [\[Crossref\]](#)
 136. Hu R, Rizwan A, Hu Z, Li T, Chung AD, Kwan BYM. An artificial intelligence training workshop for diagnostic radiology residents. *Radiol Artif Intell*. 2023;5(2):e220170. [\[Crossref\]](#)
 137. van Kooten MJ, Tan CO, Hofmeijer EIS, et al. A framework to integrate artificial intelligence training into radiology residency programs: preparing the future radiologist. *Insights Imaging*. 2024;15(1):15. [\[Crossref\]](#)
 138. Grootjans W, Krainska U, Rezazade Mehrizi MH. How do medical institutions co-create artificial intelligence solutions with commercial startups? *Eur Radiol*. 2025;35(12). [\[Crossref\]](#)
 139. Paschali M, Chen Z, Blankemeier L, et al. Foundation models in radiology: what, how, why, and why not. *Radiology*. 2025;314(2):e240597. [\[Crossref\]](#)
 140. Akinci D'Antonoli T, Bluethgen C, Cuocolo R, Klontzas ME, Ponsiglione A, Kocak B. Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects. *Diagn Interv Radiol*. 2025. [\[Crossref\]](#)
 141. Wu J, Wang Y, Zhong Z, et al. Vision-language foundation model for 3D medical imaging. *npj Artif Intell*. 2025;1(1):17. [\[Crossref\]](#)
 142. Ji J, Hou Y, Chen X, Pan Y, Xiang Y. Vision-language model for generating textual descriptions from clinical images: model development and validation study. *JMIR Form Res*. 2024;8:e32690. [\[Crossref\]](#)
 143. Zhong Z, Wang Y, Wu J, et al. Vision-language model for report generation and outcome prediction in CT pulmonary angiogram. *npj Digit Med*. 2025;8(1):432. [\[Crossref\]](#)
 144. Bian Y, Li J, Ye C, Jia X, Yang Q. Artificial intelligence in medical imaging: from task-specific models to large-scale foundation models. *Chin Med J (Engl)*. 2025;138(06):651-663. [\[Crossref\]](#)
 145. de Almeida JG, Alberich LC, Tsakou G, et al. Foundation models for radiology-the position of the AI for Health Imaging (AI4HI) network. *Insights Imaging*. 2025;16(1):168. [\[Crossref\]](#)
 146. Bradshaw TJ, Tie X, Warner J, Hu J, Li Q, Li X. Large language models and large multimodal models in medical imaging: a primer for physicians. *J Nucl Med*. 2025;66(2):173-182. [\[Crossref\]](#)
 147. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiology*. 2023;307(2):e230163. [\[Crossref\]](#)
 148. Keshavarz P, Bagherieh S, Nabipoorashrafi SA, et al. ChatGPT in radiology: a systematic review of performance, pitfalls, and future perspectives. *Diagn Interv Imaging*. 2024;105(7):251-265. [\[Crossref\]](#)
 149. Babyn PS. AI predictive uncertainty: a step forward. *Radiology*. 2023;308(3):e232144. [\[Crossref\]](#)
 150. Faghani S, Gamble C, Erickson BJ. Uncover this tech term: uncertainty quantification for deep learning. *Korean J Radiol*. 2024;25(4):395-398. [\[Crossref\]](#)
 151. Kurz A, Hauser K, Mehrtens HA, et al. Uncertainty estimation in medical image classification: systematic review. *JMIR Medical Informatics*. 2022;10(8):e36427. [\[Crossref\]](#)



Artificial intelligence in radiology examinations: a psychometric comparison of question generation methods

Emre Emekli^{1,2}

Betül Nalan Karahan¹

¹Eskişehir Osmangazi University Faculty of Medicine,
Department of Radiology, Eskişehir, Türkiye

²Eskişehir Osmangazi University, Translational
Medicine Application and Research Center, Eskişehir,
Türkiye

PURPOSE

This study aimed to evaluate the usability of artificial intelligence (AI)-based question generation methods—Chat Generative Pre-trained Transformer (ChatGPT)-4o (a non-template-based large language model) and a template-based automatic item generation (AIG) method—in the context of radiology education. The primary objective was to compare the psychometric properties, perceived quality, and educational applicability of generated multiple-choice questions (MCQs) with those written by a faculty member.

METHODS

Fifth-year medical students who participated in the radiology clerkship at Eskişehir Osmangazi University were invited to take a voluntary 15-question examination covering musculoskeletal and rheumatologic imaging. The examination included five MCQs from each of three sources: a radiologist educator, ChatGPT-4o, and the template-based AIG method. Student responses were evaluated in terms of difficulty and discrimination indices. Following the examination, students rated each question using a Likert scale based on clarity, difficulty, plausibility of distractors, and alignment with learning goals. Correlations between students' examination performance and their theoretical/practical radiology grades were analyzed using Pearson's correlation method.

RESULTS

A total of 115 students participated. Faculty-written questions had the highest mean correct response rate (2.91 ± 1.34), followed by template-based AIG (2.32 ± 1.66) and ChatGPT-4o (2.3 ± 1.14) questions ($P < 0.001$). The mean difficulty index was 0.58 for faculty, and 0.46 for both template-based AIG and ChatGPT-4o. Discrimination indices were acceptable (≥ 0.2) or very good (≥ 0.4) for template-based AIG questions. In contrast, four of the ChatGPT-generated questions were acceptable, and three were very good. Student evaluations of questions and the overall examination were favorable, particularly regarding question clarity and content alignment. Examination scores showed a weak correlation with practical examination performance ($P = 0.041$), but not with theoretical grades ($P = 0.652$).

CONCLUSION

Both the ChatGPT-4o and template-based AIG methods produced MCQs with acceptable psychometric properties. While faculty-written questions were most effective overall, AI-generated questions—especially those from the template-based AIG method—showed strong potential for use in radiology education. However, the small number of items per method and the single-institution context limit the robustness and generalizability of the findings. These results should be regarded as exploratory, and further validation in larger, multicenter studies is required.

CLINICAL SIGNIFICANCE

AI-based question generation may potentially support educators by enhancing efficiency and consistency in assessment item creation. These methods may complement traditional approaches to help scale up high-quality MCQ development in medical education, particularly in resource-limited settings; however, they should be applied with caution and expert oversight until further evidence is available, especially given the preliminary nature of the current findings.

KEYWORDS

Artificial intelligence, radiology education, multiple-choice questions, automatic item generation, clinical reasoning assessment

Handling editor: Tuğba Akıncı D'Antonoli

Corresponding author: Emre Emekli

E-mail: emreemekli90@gmail.com

Received 14 April 2025; revision requested 12 May 2025;
last revision received 10 June 2025; accepted 14 June
2025.



Epub: 21.07.2025

Publication date: 01.09.2026

DOI: 10.4274/dir.2025.253407

You may cite this article as: Emekli E, Karahan BN. Artificial intelligence in radiology examinations: a psychometric comparison of question generation methods. *Diagn Interv Radiol.* 2026;32(5):548-554.

In medical education, structured examination techniques are of great importance in objectively assessing students' knowledge, skills, and clinical decision-making competencies. In this regard, multiple-choice questions (MCQs) are widely used to measure theoretical knowledge. These questions also contribute to examination security and the standardization of the assessment process.^{1,2} However, generating high-quality MCQs requires attention to criteria such as appropriate difficulty level, discrimination index, distractor quality, and measurement validity, which demand significant time and expertise. Considering the intensity of educational programs, this creates a need for more efficient and sustainable methods in MCQ generation.³

In recent years, the rapid development of artificial intelligence (AI) has introduced potential alternatives to meet this need. With systems developed using large language models (LLMs), it is possible to generate a large number of questions in a short period. LLMs stand out as non-template-based methods for direct question generation and can create a high volume of case-based scenarios with minimal effort.⁴ In addition, template-based automatic item generation (AIG) techniques provide more structured and instructor-controllable outputs.⁵ Although both methods offer their advan-

tages, varying levels of expert oversight are required in terms of quality, accuracy, and contextual appropriateness.^{6,7} The literature suggests that questions created using these methods are psychometrically valid; however, direct comparative studies remain limited. This study distinguishes itself by directly comparing Chat Generative Pre-trained Transformer (ChatGPT)-4o and template-based AIG with faculty-written questions in the same real-life examination, while incorporating both objective psychometric analyses and subjective student evaluations in the context of radiology education.

In this study, the usability of questions generated by ChatGPT-4o (a non-template-based method) and questions developed through template-based AIG was examined within the context of assessment and evaluation. Specifically within radiology education, both AIG methods were compared with questions written by a faculty member. The evaluation was conducted based on students' examination performances, the psychometric properties of the questions, and perceived quality levels, aiming to provide empirical data on the applicability of AI-based tools in medical education.

Methods

Prior to the study, approval was obtained from the Non-Interventional Clinical Research Ethics Committee of Eskişehir Osmangazi University (decision number: 34, date: October 22, 2024). Participation in the study was voluntary, and consent was obtained from students who were to take the examination.

Participants

The study was conducted with 5th-year medical students enrolled in the Radiology internship at Eskişehir Osmangazi University Faculty of Medicine during the 2024–2025 academic year. A total of 163 students were invited to participate in the study.

Examination topics

Within the scope of the radiology internship, students were taught musculoskeletal radiology topics prepared in accordance with the European Society of Radiology⁸ and the National Core Education Program⁹, including "Radiology of the Skeletal System and Soft Tissue Tumors," "Radiology in Metabolic and Endocrine System Diseases," and

"Radiology in Rheumatologic Diseases." In addition to the routine radiology internship covering these subjects, an examination consisting of 15 MCQs was administered. Participation in this examination was voluntary and had no impact on students' internship grades. The examination was scheduled 7–10 days before the end of each internship rotation to ensure that students had received sufficient radiology training.

Question generation

Of the 15 questions, 5 were created by a faculty member, 5 by ChatGPT-4o, and 5 by the template-based AIG technique. The sources of the questions were not disclosed to the participants prior to the examination. The questions prepared by the faculty member were written by a physician with nine years of radiology experience and a PhD in medical education.

The ChatGPT-4o questions were prepared in August 2024. For each lecture topic, course materials (slide presentations used in lectures and lecture notes provided to students as resources) were shared with ChatGPT-4o. Subsequently, using prompts described in the literature, 5 questions were generated for each lecture topic.^{4,10} Each of the 5 questions used in the study was generated in a separate ChatGPT conversation to avoid memory retention effects and ensure independence between items. For each session, only one prompt was submitted, and the first complete set of 5 questions generated by ChatGPT-4o was directly included in the study without further selection or filtering. No modifications, edits, or refinements were made to the wording, content, or structure of these questions after generation.

In the template-based AIG method, three stages for question generation are defined in the literature. In the first stage, topic headings were identified.¹¹ For this study, it was planned to generate questions using the template-based AIG method within the lecture content "Radiology in Rheumatologic Diseases." Accordingly, it was decided to generate questions related to rheumatoid arthritis (RA), gout, psoriatic arthritis (PA), ankylosing spondylitis (AS), and osteoarthritis (OA). For these diseases, information variables were defined as patient age, sex, symptoms, medical history, and radiographic images. For each diagnosis, suitable age ranges (OA: 55–85 years; AS: 22–45; PA: 30–58; RA: 40–80; gout: 35–65),

Main points

- Artificial intelligence (AI)-generated questions using both Chat Generative Pre-trained Transformer (ChatGPT)-4o and template-based automatic item generation (AIG) methods had acceptable or very good discrimination indices, highlighting their consistency in performance difficulty and discrimination indices, making them suitable for use in radiology education.
- Faculty-authored questions outperformed generated ones in terms of student success, although template-based AIG questions showed the highest consistency in psychometric performance.
- Students rated the generated questions favorably, indicating that these methods can produce clear, plausible, and educationally aligned content.
- AI tools, such as ChatGPT-4o and template-based AIG systems, can alleviate faculty workload and support scalable question generation with proper validation and oversight.
- Combining AI-generated questions with expert oversight may improve the quality and efficiency of assessment development.

appropriate gender (OA, PA: female/male; AS, gout: male; RA: female), and potentially relevant radiographs were specified. Subsequently, appropriate symptoms and histories were written for each diagnosis. In the second stage, a question template was created. Two types of question templates were developed to ensure variety (Table 1). In this phase, the correct answer option was not directly tied to each diagnosis to assess clinical reasoning skills. Radiographic findings for each diagnosis were defined (e.g., OA: asymmetric joint space narrowing, subchondral sclerosis, osteophyte formation; AS: bamboo spine, shiny corner sign, vertebral squaring). One diagnosis was randomly assigned as the correct answer, and distractors were randomly selected from among the other diagnoses. Finally, questions were generated using Python-based software previously validated for use in item generation.¹² A total of 322 questions were generated using the template-based AIG method (5 for each diagnosis). One question was randomly selected from each diagnostic category to ensure diversity across clinical conditions, resulting in 5 questions used in the examination. These 5 questions were reviewed by two faculty members for accuracy, clarity, and appropriateness before inclusion.

The questions written by the faculty member were reviewed by a radiology research assistant for factual and grammatical errors before being included in the examination. Of the 5 questions generated by ChatGPT-4o and the template-based AIG method, 1 randomly selected question for each diagnosis (total of 5) was evaluated by two faculty members in terms of clarity, clinical appropriateness, presence of a single correct answer, factual accuracy, and distractor quality¹³; all were deemed appropriate and included in the study. The number of questions per method was limited to 5 to avoid placing an excessive testing burden on students during their radiology internship. Given that participation in the study was voluntary and that the examination was administered in addition to the standard curriculum assessments, it was deemed ethically and practically necessary to keep the examination length manageable. The primary aim was to explore the feasibility of applying AI-based question generation in a real-world educational setting. The study methodology is illustrated in the flowchart presented in Figure 1.

Question evaluation

After the examination, anonymous examination and question evaluation forms were distributed to students. Each question was presented individually, and students were asked to evaluate them using a previously developed form with a 5-point Likert scale ("the question text is clear"; "the question is of appropriate difficulty"; "the question has only one correct answer"; "the information provided is sufficient to find the correct answer"; "the distractors are logical"). Although students may not possess expert-level judgment on technical item quality, their feedback is valuable in assessing the clarity, plausibility, and perceived appropriateness of MCQs. Since the examination was designed for undergraduate education, student perceptions reflect real-world usability. Moreover, the Likert items used had been adapted from previous studies and paired with objective psychometric analyses to provide a more comprehensive evaluation.

Statistical analysis

The theoretical examination consisted of 32 MCQs, with 1 question corresponding to each hour of radiology instruction. The practical examination included 10 radiological cases, each accompanied by three components: identification of the imaging modality, description of radiological findings, and determination of the most likely diagnosis. Each component was scored separately to calculate the total practical score. Each student's scores from the radiology theoretical and practical examinations were recorded. The correlation between these scores and the number of correct answers in the study examination was assessed using Pearson's correlation method. Item discrimination was calculated using both the traditional 27% upper-lower group method and the Pearson item-rest correlation method. The latter was computed as the correlation between individual item scores and total test scores with the target item removed. This dual approach aligns with current psychometric recommendations and allows for more robust interpretation of item quality. Discrimination index thresholds were interpreted based on current guidelines in medical education literature, with values ≥ 0.30 considered very good, 0.20–0.29 acceptable, and < 0.20 poor¹⁴ students' perceptions of the examination were collected using a 5-point Likert scale. Due to the limited num-

ber of items ($n = 5$ per group), internal consistency measures, such as Cronbach's alpha, were not computed, as such estimates are considered statistically unreliable with small item sets.

Results

A total of 115 students (70.6%) agreed to participate in the study. The mean number of correct answers out of 15 questions in the examination was 7.53 ± 3.21 . When evaluated according to the source of the questions, the mean number of correct answers was 2.91 ± 1.34 for the faculty-written questions, 2.3 ± 1.14 for the ChatGPT questions, and 2.32 ± 1.66 for the template-based AIG questions, with a statistically significant difference ($P < 0.001$). The faculty-written questions were answered correctly more frequently than those generated by ChatGPT-4o and the template-based AIG method ($P = 0.010$; $P = 0.001$).

The average difficulty indices were calculated as 0.58 for the faculty-written questions, 0.46 for the template-based AIG questions, and 0.46 for the ChatGPT-4o questions. Regarding the discrimination index (27% upper-lower group), 5 of the faculty-written questions were classified as acceptable and 3 as very good; 4 of the ChatGPT questions were acceptable and 3 were very good; all of the template-based AIG questions were found to be appropriate (5 acceptable, 5 very good). In addition, item-rest correlation (Pearson) analyses were performed to further assess discrimination power. According to this method, 4 of the faculty-written questions, 3 of the ChatGPT-generated questions, and 5 of the template-based AIG questions showed acceptable or higher item-total correlations ($r \geq 0.20$), supporting the results obtained via the traditional method, providing preliminary support for the psychometric performance of the items. The difficulty and discrimination indices of the questions by source are presented in Table 2.

Students' evaluations of the questions and the examination are presented in Tables 3 and 4. Although the scores obtained from the prepared examination did not correlate with the theoretical examination scores ($P = 0.652$), a weak correlation was found with the practical examination scores ($P = 0.041$). Given the limited number of items per method, all psychometric results should be interpreted with caution and considered exploratory.

Table 1. Template-based question model				
CONTENT TO BE GENERATED: [Reason for Admission] [History] [Radiography] [Question Stem]				
	Template 1		Template 2	
Reason for admission	<SYMPTOM> complaint brought the <AGE>-year-old <GENDER> patient to the outpatient clinic		The <AGE>-year-old <GENDER> patient presents to the outpatient clinic with a <SYMPTOM> complaint	
History	In the patient's history, <HISTORY> is noted		The history reveals <HISTORY>	
Radiography	A <RADIOGRAPHY> is requested		A <RADIOGRAPHY> is performed for this patient	
Question stem	Which of the following findings is more likely to be seen in this patient compared with the others?		Which of the following radiographic findings would you most expect to see in this patient?	
Diagnosis	Symptoms	History	Radiography	Options
Osteoarthritis	• Pain in the right knee and restricted movement at both knees • Pain and inability to bend the left knee	• Pain worsens in the late afternoon • Pain worsens with physical activity	Knee /shoulder	Asymmetric joint space narrowing, subchondral sclerosis, osteophyte formation, Heberden's nodes, Bouchard's nodes
Ankylosing spondylitis	• Lower back pain • Morning stiffness	• Morning stiffness lasting for about an hour after waking up • Lower back stiffness decreases with physical activity	Vertebral/sacroiliac joints	Bamboo spine, vertebral squaring, shiny corner sign, syndesmophytes, ossification of the anterior longitudinal ligament
Psoriatic arthritis	• Red eyes and eye itching • Rash on extensor surfaces of extremities	• Morning stiffness lasting for about an hour after waking up • Pitting in the nails with joint pain	Hand	Arthritis mutilans, dactylitis, feather-like periostitis, pencil-in-cup deformity
Rheumatoid arthritis	• Swelling in fingers of both hands • Swelling and pain in both wrists	• Morning stiffness lasting for about an hour after waking up • Pain decreases with exercise	Hand/vertebral	Marginal erosions, rheumatoid nodules, juxta-articular osteoporosis, swan-neck deformity, Boutonnière deformity
Gout	• Severe swelling in the toe • Redness and increased warmth in the toe	• Symptoms started after alcohol consumption last night • Symptoms started after eating a rich meal	Elbow/foot	Rat-bite erosions, eccentric erosions, soft tissue tophi, Martel's sign

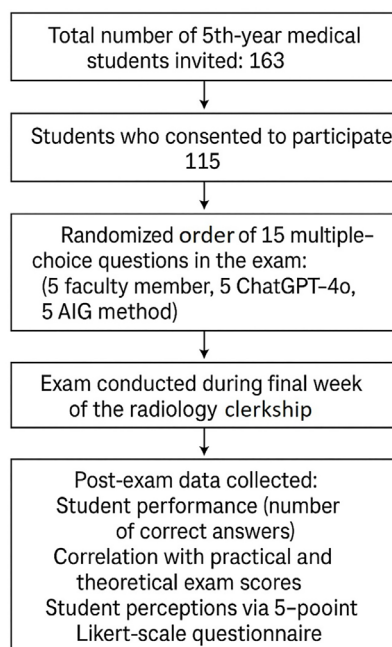


Figure 1. Flowchart of the study design. ChatGPT, Chat Generative Pre-trained Transformer; AIG, automatic item generation.

		Difficulty	Discrimination (27% upper-lower group method)	Item-rest correlation (Pearson)
Faculty member	1	0.60	0.53	0.29
	2	0.76	0.57	0.45
	3	0.61	0.36	0.27
	4	0.67	0.71	0.41
	5	0.28	0.28	0.12
Automatic item generation	1	0.57	0.76	0.48
	2	0.42	0.64	0.37
	3	0.59	0.74	0.54
	4	0.31	0.80	0.54
	5	0.43	0.74	0.46
ChatGPT-4o	1	0.28	0.65	0.41
	2	0.49	0.52	0.29
	3	0.14	0.00	−0.80
	4	0.77	0.43	0.33
	5	0.62	0.25	0.06

ChatGPT, Chat Generative Pre-trained Transformer.

	Faculty member					Automatic item generation					ChatGPT-4o				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
The question text is clear	4.73	4.76	4.88	4.68	4.54	4.65	4.49	4.73	4.58	4.63	4.53	4.63	4.28	4.73	4.69
The question is of appropriate difficulty	4.35	4.55	4.52	4.15	3.88	4.2	3.88	4.24	3.96	4.18	3.75	4.13	3.79	4.44	4.54
The question has only one correct answer	4.69	4.79	4.68	4.64	4.24	4.62	4.37	4.55	4.5	4.46	4.4	4.7	4.32	4.71	4.67
The information provided in the question is sufficient to identify the correct answer	4.52	4.65	4.81	4.48	4.3	4.51	4.24	4.5	4.2	4.41	4.28	4.54	4.19	4.73	4.57
The distractors are reasonable	4.5	4.52	4.69	4.3	4.23	4.42	4.29	4.49	4.39	4.43	4.27	4.46	4.23	4.64	4.59

ChatGPT, Chat Generative Pre-trained Transformer.

	Number of participants	1	2	3	4	5	Mean
The examination is clear	111	0	1	6	18	86	4.73
The examination is of appropriate difficulty	111	7	3	25	35	41	4.35
The examination is appropriate for the radiology internship	111	1	6	7	33	64	4.69
The examination aligns with the learning objectives	95	0	4	10	23	58	4.52

Discussion

There are publications in the literature indicating that high-quality questions can be generated using LLMs.^{15,16} In many of these studies, the questions were subject to expert review, and it was concluded that the use of these questions in examinations is appropriate.^{17,18} In studies involving student participation, the questions were found to be psychometrically acceptable when compared with human-written questions, and LLMs were shown to have the potential to serve as test developers for student assessment.¹⁹ Some studies have also calculated difficulty and discrimination indices for the questions.^{5,6,20,21} In a study by Emekli and Karahan²⁰ conducted with radiology technician students, the difficulty index for questions generated by ChatGPT was found to be 0.50, and 73.33% of the questions were deemed acceptable in terms of discrimination index. In another study in the field of anatomy, the discrimination index of the generated questions ranged between 0.29 and 0.54, and 11 out of 12 questions were reported to have a high discrimination index. In the same study, difficulty indices were calculated that ranged between 0.41 and 0.89.⁶ Similarly, in this study, 4 out of 5 ChatGPT-generated questions were found acceptable in terms of discrimination index, with an average difficulty index of 0.46. These findings appear consistent with the existing literature on questions generated with LLMs, though further validation is required.

The template-based AIG method is an automatic question generation technique that was studied and developed earlier than LLMs. The literature shows that questions of sufficient quality for assessment and evaluation can be generated using this method.²² Questions have been produced using this technique in different languages and medical specialties, and its effectiveness has been demonstrated.^{12,23,24} Recent studies have also implemented template-based AIG in multiple languages, including Polish.²⁵ In this study, the difficulty index was 0.46, and the discrimination index ranged between 0.64 and 0.8, which is classified as very good. Additionally, with the recent emergence of LLMs, hybrid methods that combine both question generation techniques have begun to be developed. This approach, combining the strengths of AI-based and template-based methods, presents a promising alternative in the field. In expert evaluations of questions generated using the hybrid method, it was reported that the correct an-

swer could be identified in 96.2% of cases, and consensus was reached among experts on all questions.²⁶

In addition to the findings in the literature, this study applied questions generated by both methods to the same student population and compared them with faculty-written questions. However, no prior study in radiology education has evaluated both question generation techniques in an examination format alongside faculty-authored questions. In one study, clinical reasoning skills in radiology were evaluated by experts, and questions generated by ChatGPT were rated 84.52% successful, faculty-written questions 82.14%, and template-based AIG-generated questions 78.57%. Based on these findings, both AI-based question generation methods were considered effective in assessing clinical reasoning skills.²⁷ Recent work by Mistry et al.²⁸ also supports the feasibility of LLMs, particularly GPT-4, in generating radiology board-style MCQs with high expert ratings in clarity, difficulty, and rationale quality. Although their study focused on expert review rather than student performance data, our findings complement theirs by providing psychometric and student-based validation of AI-generated items in undergraduate radiology education. Together, these studies contribute to a growing body of evidence supporting the role of LLMs in medical assessment design. In this study, consistent with the aforementioned research, both methods were found to produce questions with appropriate difficulty and discrimination indices. Additionally, the questions generated using the template-based AIG method demonstrated particularly high discrimination indices.

ChatGPT and template-based AIG offer significant advantages in accelerating and diversifying the process of preparing MCQs, while also reducing the workload on faculty members. The fact that the questions produced using these methods have acceptable levels of difficulty and discrimination suggests potential utility in assessment processes, although further research is needed to confirm this. However, due to the limited sample size of items in this study, these findings should be considered preliminary and exploratory, in line with similar small-scale investigations in the literature. Further large-scale studies are required to confirm these initial observations. These methods also have certain limitations. Questions generated by LLMs may sometimes contain content errors or semantic ambiguities, whereas the template-based AIG method requires ex-

tensive preparation and expertise, making it time-consuming. To minimize these limitations, hybrid methods that combine the speed and flexibility of AI with the controllability of template-based approaches may offer a more balanced and reliable question generation model.

This study has some limitations. First, it was conducted at a single medical school and only with 5th-year radiology internship students, which may limit the generalizability of the findings. Additionally, the examination administered during the study was voluntary and not included in official grading, which may have resulted in variable student motivation levels. The limited number of questions in the examination also restricted the evaluation of a broader range of item characteristics for each method. In particular, key psychometric metrics, such as reliability coefficients, could not be calculated due to the small item set, limiting the interpretability of overall examination consistency. Therefore, the findings should be viewed as preliminary and not generalizable without replication in larger datasets. The inclusion of only 5 questions per method constrains the reliability and generalizability of the psychometric analyses. Future research should consider increasing the number of items—ideally approaching the scale of studies such as Law et al.²⁹, which suggests 100 questions per method—to strengthen statistical power and validity. Although the prompt structures used in ChatGPT-4o were systematically controlled, the model's inherent variability may have led to some content differences. In the template-based AIG technique, questions were generated based on only one specific topic, limiting the evaluation of the method's effectiveness in other radiologic subfields. Some items in the examination showed discrimination index values >0.70, which may indicate high heterogeneity among participants or inconsistent motivation, as the test was voluntary and ungraded. Additionally, we used both the 27% upper-lower method and item-rest (Pearson) correlation for discrimination analysis. The latter provided more moderate values, suggesting that the items performed reasonably well across ability levels. Finally, psychometric analyses were conducted solely within the framework of classical test theory, limiting comparisons with alternative analytical methods. Therefore, future research is recommended to include larger participant groups, be conducted in different medical schools, and involve multicenter applications across various clinical domains. Another limitation is relat-

ed to the nature of the student evaluations of the MCQs. Although students provided valuable insights regarding question clarity, plausibility of distractors, and perceived alignment with learning goals, their ability to assess critically the psychometric and clinical quality of the questions may be limited. As undergraduate learners, they may not have the expertise required to fully evaluate item quality or the nuances of the clinical content presented.

In conclusion, this study demonstrates that AI-based question generation techniques in radiology education perform comparably with faculty-authored questions in a limited pilot setting. Both ChatGPT-4o and template-based AIG-generated questions possess acceptable levels of discrimination. However, due to the small number of items and single-institution design, these results should be interpreted with caution, given the small sample of items used and the single-institution design. The findings are preliminary and exploratory and should not be generalized without replication in larger item sets and multicenter studies. AI-based question generation methods may serve as supplementary tools in student assessment by reducing the time and expertise requirements traditionally needed in question development. Given the small item sample used in this study, the results should be interpreted as exploratory and hypothesis-generating rather than conclusive. Future research with larger item sets and multicenter designs is warranted to further validate the effectiveness of these approaches.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

- Pugh D, De Champlain A, Touchie C. Plus ça change, plus c'est pareil: making a continued case for the use of MCQs in medical education. *Med Teach*. 2019;41(5):569-577. [\[Crossref\]](#)
- Wrigley W, van der Vleuten CP, Freeman A, Muijtjens A. A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. *Med Teach*. 2012;34(9):683-697. [\[Crossref\]](#)
- Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? *Med Educ*. 2004;38(9):974-979. [\[Crossref\]](#)
- Kiyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. *Postgrad Med J*. 2024;100(1189):858-865. [\[Crossref\]](#)
- Abujabal A, Yahya M, Riedewald M, Weikum G. Automated template generation for question answering over knowledge graphs. In: Proceedings of the 26th International Conference on World Wide Web (WWW '17). Geneva, Switzerland; 2017:1191-1200. [\[Crossref\]](#)
- Kiyak YS, Soyulu A, Coşkun Ö, Budakoğlu İ, Peker T. Can ChatGPT generate acceptable case-based multiple-choice questions for medical school anatomy exams? A pilot study on item difficulty and discrimination. *Clin Anat*. 2025;38(4):505-510. [\[Crossref\]](#)
- Leslie T, Gierl MJ. Using automatic item generation to create multiple-choice questions for pharmacy assessment. *Am J Pharm Educ*. 2023;87(10):100081. [\[Crossref\]](#)
- European Society of Radiology. ESR curriculum for undergraduate radiological education [Internet]. Vienna: European Society of Radiology. [cited 2025 Apr 4]. Available from: <https://www.myesr.org/education/esr-curriculum>. [\[Crossref\]](#)
- Medical Faculty - National Core Curriculum 2020. *Tip Eğitimi Dünyası*. 2020;19(57-1):1-146. [\[Crossref\]](#)
- Kiyak YS. A ChatGPT prompt for writing case-based multiple-choice questions. *Span J Med Educ*. 2023;4(3). [\[Crossref\]](#)
- Gierl MJ, Lai H, Turner SR. Using automatic item generation to create multiple-choice test items. *Med Educ*. 2012;46(8):757-765. [\[Crossref\]](#)
- Emekli E, Emekli E, Kiyak YS, Hoşgören Alıcı Y, Coşkun Ö, Budakoğlu İ. Clinical reasoning in psychiatric education: development of multiple-choice questions with automatic item generation in Turkish. *Türk Psikiyatri Derg*. 2025;36:336-343. [\[Crossref\]](#)
- Pugh D, De Champlain A, Gierl M, Lai H, Touchie C. Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *Res Pract Technol Enhanc Learn*. 2020;15:12. [\[Crossref\]](#)
- Downing SM, Yudkowsky R. Assessment in health professions education (1st ed.). Routledge. 2009. [\[Crossref\]](#)
- Elkins S, Kochmar E, Serban I, Cheung JCK. How useful are educational questions generated by large language models? In: Wang N, Rebollo-Mendez G, Dimitrova V, et al., eds. Artificial intelligence in education. Communications in Computer and Information Science. Vol 1831. Springer; 2023:603-607. [\[Crossref\]](#)
- Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLoS One*. 2023;18(8):e0290691. [\[Crossref\]](#)
- Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, bard, and bing to generate reasoning-based multiple-choice questions in medical physiology. *Cureus*. 2023;15(6):e40977. [\[Crossref\]](#)
- Ayub I, Hamann D, Hamann CR, Davis MJ. Exploring the potential and limitations of chat generative pre-trained transformer (ChatGPT) in generating board-style dermatology questions: a qualitative analysis. *Cureus*. 2023;15(8):e43717. [\[Crossref\]](#)
- Lin Z, Chen H. Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System*. 2024;123:103344. [\[Crossref\]](#)
- Emekli E, Karahan BN. AI in radiography education: evaluating multiple-choice questions difficulty and discrimination. *J Med Imaging Radiat Sci*. 2025;56(4):101896. [\[Crossref\]](#)
- Zuckerman M, Flood R, Tan RJB, Kelp N, Ecker DJ, Menke J, Lockspeiser T. ChatGPT for assessment writing. *Med Teach*. 2023;45(11):1224-1227. [\[Crossref\]](#)
- Gierl MJ, Lai H. Evaluating the quality of medical multiple-choice items created with automated processes. *Med Educ*. 2013;47(7):726-733. [\[Crossref\]](#)
- Kurdi G, Leo J, Parsia B, Sattler U, Al-Emari S. A systematic review of automatic question generation for educational purposes. *Int J Artif Intell Educ*. 2020;30(1):121-204. [\[Crossref\]](#)
- Gierl MJ, Lai H, Tanygin V. Advanced methods in automatic item generation. 1st ed. New York, NY: Routledge; 2021. [\[Crossref\]](#)
- Kiyak YS, Kononowicz A, Górski S. Multilingual template-based automatic item generation for medical education supported by generative artificial intelligence models ChatGPT and Claude. *Bio-Algorithms and Med-Systems*. 2024;20(1):81-96. [\[Crossref\]](#)
- Kiyak YS, Emekli E, Coşkun Ö, Budakoğlu İ. Keeping humans in the loop efficiently by generating question templates instead of questions using AI: validity evidence on hybrid AIG. *Med Teach*. 2025;47(4):744-747. [\[Crossref\]](#)
- Emekli E, Karahan BN. Comparison of automatic item generation methods in the assessment of clinical reasoning skills. *Span J Med Educ*. 2024;6(1). [\[Crossref\]](#)
- Mistry NP, Saeed H, Rafique S, Le T, Obaid H, Adams SJ. Large language models as tools to generate radiology board-style multiple-choice questions. *Acad Radiol*. 2024;31(9):3872-3878. [\[Crossref\]](#)
- Law AK, So J, Lui CT, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Med Educ*. 2025;25(1):208. [\[Crossref\]](#)



Diagnostic accuracy and consistency of ChatGPT-4o in radiology: influence of image, clinical data, and answer options on performance

Kadir Atakır
 Kaan Işın
 Abdullah Taş
 Hakan Önder

University of Health Sciences Türkiye, Prof. Dr. Cemil Taşcıoğlu City Hospital, Clinic of Radiology, İstanbul, Türkiye

PURPOSE

This study aimed to evaluate the diagnostic accuracy of Chat Generative Pre-trained Transformer (ChatGPT) version 4 Omni (ChatGPT-4o) in radiology across seven information input combinations (image, clinical data, and multiple-choice options) to assess the consistency of its outputs across repeated trials and to compare its performance with that of human radiologists.

METHODS

We tested 129 distinct radiology cases under seven input conditions (varying presence of imaging, clinical context, and answer options). Each case was processed by ChatGPT-4o for seven different input combinations on three separate accounts. Diagnostic accuracy was determined by comparison with ground-truth diagnoses, and interobserver consistency was measured using Fleiss' kappa. Pairwise comparisons were performed with the Wilcoxon signed-rank test. Additionally, the same set of cases was evaluated by nine radiology residents to benchmark ChatGPT-4o's performance against human diagnostic accuracy.

RESULTS

ChatGPT-4o's diagnostic accuracy was lowest for "image only" (19.90%) and "options only" (20.67%) conditions. The highest accuracy was observed in "image + clinical information + options" (80.88%) and "clinical information + options" (75.45%) conditions. The highest interobserver agreement was observed in the "image + clinical information + options" condition ($\kappa = 0.733$) and the lowest was in the "options only" condition ($\kappa = 0.023$), suggesting that more information improves consistency. However, there was no effective benefit of adding imaging data over already provided clinical data and options, as seen in post-hoc analysis. In human comparison, ChatGPT-4o outperformed radiology residents in text-based configurations (75.45% vs. 42.89%), whereas residents showed slightly better performance in image-based tasks (64.13% vs. 61.24%). Notably, when residents were allowed to use ChatGPT-4o as a support tool, their image-based diagnostic accuracy increased from 63.04% to 74.16%.

CONCLUSION

ChatGPT-4o performs well when provided with rich textual input but remains limited in purely image-based diagnoses. Its accuracy and consistency increase with multimodal input, yet adding imaging does not significantly improve performance beyond clinical context and diagnostic options alone. The model's superior performance to residents in text-based tasks underscores its potential as a diagnostic aid in structured scenarios. Furthermore, its integration as a support tool may enhance human diagnostic accuracy, particularly in image-based interpretation.

CLINICAL SIGNIFICANCE

Although ChatGPT-4o is not yet capable of reliably interpreting radiologic images on its own, it demonstrates strong performance in text-based diagnostic reasoning. Its integration into clinical workflows—particularly for triage, structured decision support, or educational purposes—may augment radiologists' diagnostic capacity and consistency.

KEYWORDS

Artificial intelligence, large language models, ChatGPT, diagnostic accuracy, observer variation, clinical decision support systems, medical imaging, multimodal input

Handling editor: Burak Kocak

Corresponding author: Abdullah Taş

E-mail: atas997@gmail.com

Received 14 May 2025; revision requested 08 June 2025;
last revision received 28 July 2025; accepted 04 August 2025.



Epub: 22.09.2025

Publication date: 01.09.2026

DOI: 10.4274/dir.2025.253460

Artificial intelligence (AI) has emerged as a transformative tool in radiology. It offers the potential to organize diagnostic workflows, improve clinical decision-making, and reduce workload on radiologists. Among the different AI tools, large language models (LLMs), such as OpenAI's Chat Generative Pre-trained Transformer (ChatGPT), have shown marked capabilities in the medical field; for example, they can assist students in passing medical examinations and can support clinical decision-making.¹ In radiology, early applications of ChatGPT (ChatGPT-3.5 and 4) are variable and include activities such as data mining, structuring radiology reports, and answering text-based board-like examination questions.^{2,3} A recent study showed that ChatGPT-4 outperformed the previous version (ChatGPT-3.5) on written radiology questions.⁴ One of the latest updates of ChatGPT, ChatGPT-4 with vision (ChatGPT-4V), added the capability to evaluate images, which draws attention to applying this feature in radiological image interpretation. However, there have been few studies about ChatGPT-4V's accuracy in radiological image interpretation. Initial findings suggest that there is a considerable performance gap between the evaluation of images and the accomplishment of text-based tasks.^{2,5}

ChatGPT was originally developed as a text-based model. With the release of ChatGPT-4V, it gained the ability to interpret visual inputs. This became a turning point, as the model was able to process and analyze medical images along with clinical data. Multimodal functionality allows ChatGPT to open new opportunities for integration into radiology workflows and is prompting

researchers to evaluate its diagnostic performance.

Although numerous studies report high diagnostic accuracy (>80%) for LLMs in text-based tasks,^{1,6} only a limited number have demonstrated comparable performance in image-based scenarios; even then, this has been under highly controlled conditions, such as narrow diagnostic targets and optimized prompts.⁷ Conversely, many investigations have highlighted major limitations in "image only" scenarios, particularly in complex modalities, such as brain magnetic resonance imaging (MRI), where performance drops markedly in the absence of clinical context.^{4,8} Studies evaluating ChatGPT-4V against radiologists have revealed mixed outcomes, with some reporting comparable accuracy, whereas others found substantial underperformance by the model.^{9,10} This variability underscores the importance of input structure in multimodal tasks.

A growing body of literature emphasizes that prompt design, including differential diagnoses, clinical details, and structured questions, greatly influences LLM output.^{11,12} This variability raises concerns regarding reproducibility and reliability in diagnostic settings, where standardization is critical.^{3,13} Taken together, current evidence suggests that LLM performance in radiologic diagnosis is highly dependent on how and what information is provided. However, most prior studies have focused on isolated scenarios or specific modalities, without systematically comparing different input combinations under controlled conditions.^{14,15} Our study addresses this gap by evaluating the diagnostic accuracy and consistency of ChatGPT-4o across seven distinct input formats, ranging from image-only to full contextual prompts with multiple-choice options.

This study evaluates the diagnostic accuracy and consistency of ChatGPT-4o across varying structured input combinations and compares its performance with radiology residents under matched conditions. This benchmark is essential to contextualize the model's capabilities relative to human expertise and assess its potential clinical applicability.

Methods

Study designs and objective

This study aimed to evaluate ChatGPT-4o's diagnostic decision-making about radiological images and its potential role in real-life

diagnoses when provided with clinical information and differential diagnosis options. The objective was to investigate the radiological images' contribution in ChatGPT-4o's decision-making process and to understand the possible integration of ChatGPT for daily radiological tasks and workflows. An overview of the workflow is presented in Figure 1. Ethics committee approval was not required, as the study did not involve patient data or identifiable personal information. The participation of radiology residents was conducted in an educational setting and did not involve any intervention or collection of sensitive data. Informed consent was obtained from all participants prior to data collection. This study was conducted in compliance with the Minimum Reporting Items for Clear Evaluation of Accuracy Reports of LLMs in healthcare guidelines.¹⁶

Dataset

A total of 129 radiology cases were selected from two publicly available, peer-reviewed sources: the New England Journal of Medicine (NEJM) Image Challenge and the Radiological Society of North America (RSNA) Case Collection.^{17,18} The NEJM cases that matched the study format and included a radiological image were included in full. From the RSNA Case Collection, only cases published in 2023 that followed the same format were selected. The inclusion criteria required the presence of a diagnostic-quality radiological image (in single-slice static JPEG format) and a clearly defined radiologic diagnosis. Cases were excluded if they lacked radiological imaging (e.g., pathological or macroscopic photographs), if the diagnosis was uncertain, or if the case format did not conform to the five-option multiple-choice structure. Both sources were chosen for their educational value, structured content, and diagnostic clarity. The resulting dataset was designed to reflect a diverse range of real-world clinical scenarios rather than stylized board-examination questions. A detailed breakdown of cases by imaging modality and anatomical system is available in Supplementary Material 1.

Experimental setup and prompt design

To evaluate the impact of varying prompt components on the diagnostic performance of a large multimodal language model (ChatGPT-4o), a systematic prompt engineering framework was implemented inspired by recent studies on multimodal AI in radiology.^{2,11} A total of 129 radiological cases were compiled, each consisting of three main ele-

Main points

- Chat Generative Pre-trained Transformer version 4 Omni (ChatGPT-4o) achieved the highest diagnostic accuracy when clinical information and diagnostic options were provided and the lowest in the "image only" condition.
- Interobserver consistency improved with richer input combinations but was minimal in the "options only" setting, indicating a near-random output.
- Adding radiologic images to an already informative clinical prompt did not significantly enhance accuracy, highlighting current limitations in image interpretation.
- Compared with radiology residents, ChatGPT-4o outperformed humans in text-based tasks, whereas residents maintained an advantage in image-only diagnoses.

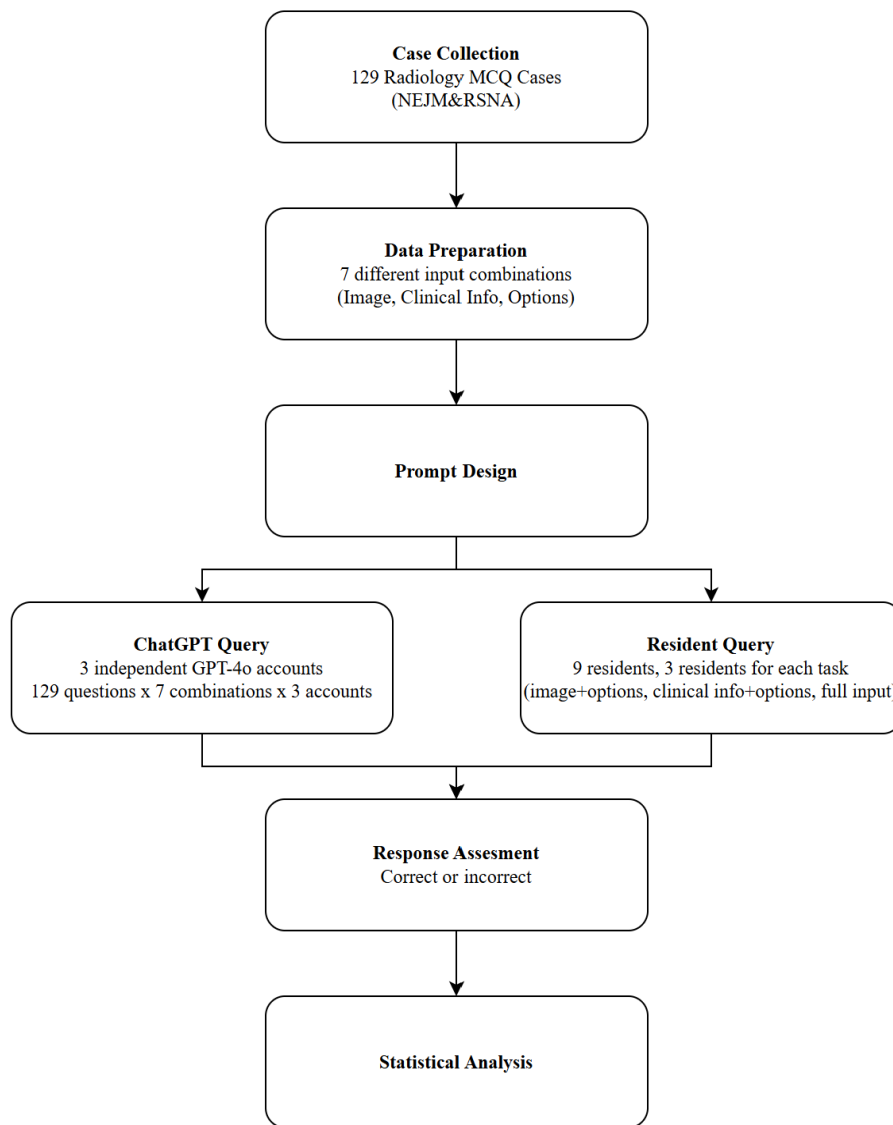


Figure 1. Flowchart of the study. MCQ, multiple choice question; NEJM, New England Journal of Medicine; RSNA, Radiological Society of North America; GPT-4o, Generative Pre-trained Transformer 4 Omni.

ments: a diagnostic-quality radiology image, structured clinical information, and five multiple-choice diagnostic options with a single correct answer. For each case, seven distinct prompt types were constructed by presenting the model with different combinations of these elements:

1. Radiological image only
2. Clinical information only
3. Differential diagnosis options only
4. Radiological image + clinical information
5. Radiological image + diagnostic options
6. Clinical information + diagnostic options

7. Radiological image + clinical information + diagnostic options

Each of these seven combinations was queried using ChatGPT-4o, accessed via the ChatGPT web interface (version released April 2024).

To explore how different types of input influence diagnostic performance, seven distinct prompt configurations were designed by selectively combining radiological images, clinical context, and diagnostic options. Each combination was chosen to isolate specific aspects of the model's reasoning process. For example, the "image only" prompt evaluates pure visual interpretation, whereas the "clinical only" prompt focuses on text-based decision-making. The "options only" condition was included to examine how the model performs without any

contextual information, relying solely on its internal knowledge and prior associations between disease names. This also helps reveal whether certain choices are favored due to frequency bias or random selection, offering a useful reference point for interpreting improvements seen in more informed conditions. Additionally, combinations of these inputs examine how multimodal integration affects performance. Together, these configurations allow for a structured comparison of how each input type contributes to both accuracy and consistency.

Prompts were automatically generated and submitted sequentially via a controlled Python-based interface. In all scenarios, the model was instructed to either choose the best-fitting diagnosis or provide a diagnostic decision based on the given input. The same base prompt format (e.g., "You will be provided with a radiologic image and associated clinical information. Based on both inputs, determine the most likely diagnosis...") was used to maintain consistency across conditions. No prompt engineering was performed beyond standardization for content structure and clarity. All interactions were conducted using the default ChatGPT interface settings, including default temperature and sampling parameters. All designed prompts are listed in Supplementary Material 2. The model was instructed to base its answers only on the information provided in each prompt. For each prompt, the model responded (Figures 2-5). The answers were then reviewed, and a record was made of whether the answer was correct. The model answered each question without knowledge of whether the previous answers were correct or not. All prior case context was then discarded before proceeding to the next prompt. Thus, the model was designed to approach each question independently. This design allowed for the isolated assessment of each input component's influence on diagnostic accuracy and ensured methodological reproducibility across all prompt combinations.

Repetition and observer simulation

First, to evaluate interobserver variability, all case combinations were submitted independently using three different ChatGPT-4o accounts, each accessed via a distinct computer and browser instance to eliminate any session overlap or caching artifacts. This resulted in a total of 2,709 model-generated responses (129 cases × 7 combinations × 3 accounts).

Answers were recorded as either correct (+) or incorrect (−), based on the alignment of

the model's final answer with the reference diagnosis provided in the source case material. In cases where multiple plausible answers were generated, responses were marked incorrect if the correct answer was not unambiguously prioritized.

Diagnostic comparison with radiology residents

To assess the performance of ChatGPT-4o, a supplementary analysis was conducted involving radiology residents to provide a benchmark for human diagnostic performance. Specifically, nine residents (post-graduate year 3 or 4 level) were recruited and randomly divided into three separate

groups, each assigned to one of the following input configurations: (1) image + options, (2) clinical information + options, and (3) image + clinical information + options. Among the seven configurations tested with ChatGPT-4o, only three were selected for human comparison, as these represented clinically meaningful scenarios; configurations lacking diagnostic options were excluded due to their anticipated low accuracy and limited comparative value. Each participant independently reviewed the same set of 129 cases in their assigned condition and submitted their diagnostic answers. In the second round (conducted 1 week later), they were allowed to use ChatGPT-4o as a supportive

tool and could revise their answers accordingly. This comparison is intended to explore the baseline performance of trained human readers and the potential additive value of LLM-generated input.

Data recording and analysis

All responses were manually recorded in structured Microsoft Excel spreadsheets (Microsoft Corporation, Redmond, WA, USA) by the research team. The accuracy rate (%) for each combination and account was calculated.

Statistical analysis

Descriptive statistics were used to calculate the mean diagnostic accuracy across all input combinations and ChatGPT-4o accounts. The distribution of variables was assessed with the Shapiro-Wilk test, revealing non-normal distributions for all combinations. For this reason, non-parametric tests were employed for further analysis. Fleiss' kappa was used to evaluate interobserver agreement across the three independent ChatGPT-4o accounts for each input combination. Pairwise comparisons between combinations were conducted using the Wilcoxon signed-rank test. All analyses were performed using the SPSS version 27.0 (IBM, USA) package, and statistical significance was set at $P < 0.05$.

Results

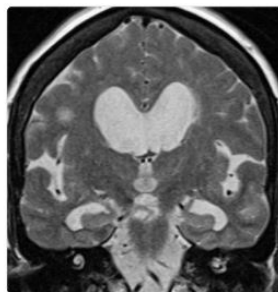
A descriptive analysis of diagnostic accuracy across the three independent ChatGPT-4o accounts revealed different patterns based on the type and combination of input information (Table 1). The lowest mean accuracies were observed when the model was presented with diagnostic options alone (20.67%) and with "image only" inputs (19.90%). In contrast, performance significantly improved with the inclusion of multiple input types. The "clinical information + options" configuration achieved a mean accuracy of 75.45%, whereas the "image + clinical information + options" combination yielded the highest overall performance at 80.88%. Accuracy scores were relatively consistent across accounts, indicating stable model behavior when sufficient contextual information was available. These results highlight the importance of multimodal input for maximizing diagnostic accuracy and reliability in LLMs within radiology.

To determine the distribution characteristics of the diagnostic accuracy data, the Shapiro-Wilk test was performed on case-level

ChatGPT 4o ▾

You will be presented with a radiologic image. Based solely on this image, determine the most likely diagnosis. Do not assume any clinical information or consider answer choices. Justify your decision based on visible imaging findings only.

Sure, please upload the radiologic image you'd like me to evaluate.



The coronal T2-weighted MRI image reveals **marked dilation of the lateral ventricles**, particularly involving the frontal horns, with **periventricular hyperintensities** suggestive of transependymal CSF flow. There is **no evidence of mass effect or obstructive lesion** at the foramen of Monro, third ventricle, or aqueduct visible on this slice.

Most Likely Diagnosis:

Normal Pressure Hydrocephalus (NPH)

Justification Based on Imaging:

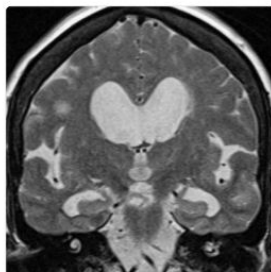
- **Ventriculomegaly** out of proportion to cortical atrophy (disproportionately enlarged ventricles).
- **Rounded configuration of the frontal horns** of the lateral ventricles.
- **Periventricular T2 hyperintensities** indicating interstitial edema due to CSF transudation.
- Absence of cortical sulcal effacement or mass effect, ruling out acute obstructive hydrocephalus or mass lesion.

This imaging pattern is classic for NPH, though definitive diagnosis requires clinical correlation and response to CSF diversion.

Figure 2. ChatGPT-4o query for the "image only" condition. No clinical information or answer options are provided. MRI, magnetic resonance imaging; CSF, cerebrospinal fluid.

You will be provided with a radiologic image and associated clinical information. Based on both inputs, determine the most likely diagnosis. Do not use any external answer choices. Explain your reasoning by integrating imaging and clinical features.

Of course—please upload the radiologic image and provide the associated clinical information so I can begin the analysis.



A 70-year-old woman presented to the neurology clinic with a 2-year history of gait disturbance, cognitive impairment, and urinary incontinence. What is the diagnosis?

The clinical triad of **gait disturbance**, **cognitive impairment**, and **urinary incontinence** in an elderly patient strongly suggests **normal pressure hydrocephalus (NPH)**.

Radiologic Findings:

The MRI (T2-weighted coronal image) demonstrates:

- **Ventriculomegaly:** Markedly enlarged lateral ventricles, especially the frontal horns, without corresponding cortical atrophy.
- **Disproportionately enlarged subarachnoid space hydrocephalus (DESH)** pattern:
 - Enlarged Sylvian fissures.
 - Narrowed high-convexity sulci, indicating compression.
- No mass lesion or obvious obstruction at the level of the aqueduct or foramen of Monro.

Figure 3. ChatGPT-4o query for radiological image + clinical information condition. No options are given. ChatGPT-4o, Chat Generative Pre-trained Transformer; MRI, magnetic resonance imaging.

mean accuracy scores for each of the seven input combinations. The results revealed that none of the distributions conformed to normality (all $P < 0.001$). This finding supports the use of non-parametric statistical methods, such as Fleiss' kappa and the Wilcoxon signed-rank test, for further analyses.

To assess interobserver consistency across different ChatGPT-4o accounts, Fleiss' kappa was calculated for each input combination (Table 2). The interobserver agreement varied considerably, with the highest consistency observed in the "image + clinical + options" condition ($\kappa = 0.733$), indicating substantial agreement. Conversely, the "options only" condition resulted in the lowest agreement ($\kappa = 0.023$), suggesting high variability in model outputs when contextual

information was not provided. Intermediate levels of agreement were seen in the "image + clinical" ($\kappa = 0.606$) and "image + options" ($\kappa = 0.652$) groups. These findings further emphasize the stabilizing effect of rich contextual input on the model's behavior.

Post-hoc comparisons using Wilcoxon signed-rank tests revealed statistically significant differences in diagnostic accuracy between most input combinations ($P < 0.001$) (Figure 6). Both "image only" and "options only" configurations performed significantly worse than any multimodal combination. No statistically significant difference was observed between "image + clinical" and "image + options" ($P = 0.051$) or between "clinical + options" and "image + clinical + options" ($P = 0.058$), indicating diminishing marginal

returns from image inclusion when comprehensive clinical context and differential options are already available. Together, these results emphasize the dominant contribution of clinical information and the synergistic value of multimodal input in enhancing diagnostic accuracy and reproducibility of LLM-driven image interpretation.

Although these results reflect intramodel variability, human-level diagnostic performance was further assessed for context. A benchmarking analysis was conducted involving nine radiology residents across the same 129-case dataset, grouped by input condition (Figure 7). Given the exploratory nature and limited size of the human comparator group, statistical inference was intentionally avoided to prevent overinterpretation. In the "image + options" scenario, the mean accuracy of residents (63.04%) was marginally higher than that of ChatGPT-4o (61.24%), suggesting that human readers retain a slight advantage in visual interpretation. However, when residents had access to ChatGPT-4o's image-based outputs, their accuracy improved markedly to 74.16%, indicating a clear benefit from AI-assisted interpretation. In the "clinical information + options" condition, ChatGPT-4o substantially outperformed the residents (75.45% vs. 42.89%). When residents were allowed to use ChatGPT-4o before answering, their performance rose significantly (74.93%), almost matching that of the model. In the full-input configuration (image + clinical + options), ChatGPT-4o again outperformed the resident group (80.88% vs. 71.57%), and the combination of both yielded the highest accuracy overall (83.47%). In all input combinations, ChatGPT-4o access provided significant benefit to residents, especially in structured clinical decision-making scenarios.

Discussion

The findings of this study provide several key insights into the diagnostic capabilities and limitations of LLMs, particularly ChatGPT-4o, within the context of radiological image interpretation. First, diagnostic performance was found to be remarkably limited when the model was presented with radiological images or diagnostic options in isolation. The "image only" and "options only" combinations resulted in the lowest mean accuracies, at 19.90% and 20.67%, respectively. In contrast, the inclusion of clinical information led to a marked improvement in performance, especially when paired with diagnostic options. This pattern may reflect the inherent design of instruction-tuned LLMs, which perform

You will receive clinical information and a list of possible diagnoses. Select the most likely diagnosis based only on the clinical details and the listed options. Do not incorporate any imaging data in your reasoning.

Understood. Please provide the clinical information and the list of diagnostic options.

A 70-year-old woman presented to the neurology clinic with a 2-year history of gait disturbance, cognitive impairment, and urinary incontinence. What is the diagnosis?

- a. Normal-pressure hydrocephalus
- b. Alzheimer's disease
- c. Parkinson's disease
- d. Benign intracranial hypertension
- e. Multiple system atrophy

The most likely diagnosis is:

a. Normal-pressure hydrocephalus (NPH)

Reasoning:

This patient presents with the classic triad of **gait disturbance, cognitive impairment, and urinary incontinence**, which is strongly suggestive of **normal-pressure hydrocephalus**. This triad is often summarized as:

- **"Wet"** – urinary incontinence
- **"Wobbly"** – gait disturbance (often magnetic gait)
- **"Wacky"** – cognitive decline

While other conditions like Alzheimer's disease (cognitive dominant), Parkinson's disease (bradykinesia, tremor), and multiple system atrophy (parkinsonism with autonomic features) may share overlapping features, none present as prominently with this triad.

Benign intracranial hypertension (idiopathic intracranial hypertension) is not typical in this demographic and usually presents with headache and visual symptoms rather than gait and cognitive issues.

demonstrated poor reproducibility across different accounts. Interestingly, although the addition of radiological images to the "clinical information + options" input did not improve accuracy, it did enhance interobserver agreement. This suggests that image input, although not increasing correctness, may contribute to more consistent model behavior across different interactions, thereby stabilizing diagnostic output when the clinical context is already rich.

As part of this study, we also conducted a supplementary analysis involving radiology residents to contextualize the performance of ChatGPT-4o. Doing so allowed for indirect insights into question difficulty, the impact of image and clinical information on decision-making, and the presence of guiding clinical hints embedded in the cases. Residents demonstrated relatively stronger performance in image-provided scenarios, reflecting their visual training and possible reliance on pattern recognition developed through clinical exposure; this contrasted with ChatGPT-4o, which performed particularly well on text-based tasks when clinical information was provided, likely due to its broad access to medical knowledge and its advanced capacity for textual reasoning. When residents were allowed to consult ChatGPT during interpretation, their accuracy improved further, especially when all inputs—images, clinical context, and diagnostic options—were available. Interestingly, in the "image + options" configuration, residents assisted by ChatGPT were able to outperform the model itself, whereas in other combinations, they barely matched its performance. This observation suggests that, at their current stage, LLMs may enhance human diagnostic accuracy in visually anchored tasks by offering auxiliary reasoning pathways or differential diagnosis support, rather than serving as standalone diagnosticians.

Although the presence of implicit diagnostic cues in clinical descriptions (commonly referred to as the "hint effect") may have contributed to the model's performance, this alone is unlikely to account for the substantial gap observed between the model and human readers, as radiology residents would also notice these cues. The markedly higher accuracy of ChatGPT-4o compared with residents in the "clinical information + options" scenario (75.45% vs. 42.79%) suggests that the model may be leveraging subtle linguistic patterns that elude human interpretation, as noted by earlier studies.¹⁹

Figure 4. ChatGPT-4o query for clinical information + diagnostic options condition. No image is given. ChatGPT-4o, Chat Generative Pre-trained Transformer.

best when decision space is constrained and explicitly framed. This relationship between input structuring and model performance is further illustrated when examining how different combinations interact. Although adding clinical information to the "image + options" condition significantly improved accuracy, the reverse scenario—adding images to the "clinical information + options" setting—did not yield a statistically significant gain. This asymmetry may indicate that ChatGPT-4o derives proportionally more diagnostic value from structured textual data than from single-slice imaging inputs, consistent with its language-dominant architecture. Furthermore, although a ceiling effect may have limited further gains in the "clinical

information + options" setting, the findings suggest that ChatGPT-4o's ability to interpret radiologic images independently remains limited. Accuracy was lowest with "image only" input, and adding images to text-based prompts did not yield significant improvements. These results indicate that the model relies more on structured text than on visual reasoning, which may be restricted by current architectural constraints.

Interobserver agreement analyses demonstrated that the highest Fleiss' kappa scores were observed in multimodal combinations, particularly when clinical information and diagnostic options were included. In contrast, scenarios lacking a clinical context

Consistent with prior research, our study found that the combination of clinical information and diagnostic options yielded relatively high diagnostic accuracy among

the tested input configurations, whereas the addition of radiological images to this setting did not lead to a significant improvement. This outcome mirrors the findings of Li et al.², who reported enhanced diagnostic performance when clinical context was added to imaging input. Similarly, Güneş et al.²⁰ and Horiuchi et al.²¹ observed substantially higher accuracy in text-based tasks than in image-based ones, highlighting the model's reliance on structured language. Elek et al.²² also demonstrated that although ChatGPT could recognize anatomical structures on computed tomography and MRI scans, its diagnostic accuracy remained low in the absence of guiding clinical information. These findings suggest that when structured textual context is available, visual input contributes only marginally to the model's decision-making process. Collectively, this evidence supports the presence of a text-biased processing pattern in current LLMs and underscores their suitability for structured clinical reasoning tasks, particularly those involving multiple-choice formats. Moreover, other studies have reported high accuracy rates in text-based radiology tasks that incorporate imaging findings as written input.^{21,23,24} Nevertheless, this strength does not generalize to all text-based formats. For instance, Sonoda et al.²⁵ compared ChatGPT-4o, Claude 3 Opus, and Gemini 1.5 Pro, and found that ChatGPT-4o performed poorly compared with Claude 3 Opus, even when provided with richly structured radiology findings. Similarly, studies by Fervers et al.²⁶ and Perchik et al.²⁷ have shown low model performance in narrative or open-ended scenarios, indicating that LLM effectiveness in radiology is strongly influenced by task design, question format, and the structure of the input prompt.

While the addition of radiological images did not significantly enhance diagnostic accuracy, our findings revealed a notable improvement in interobserver agreement with the inclusion of visual input. This suggests that image data, although insufficient

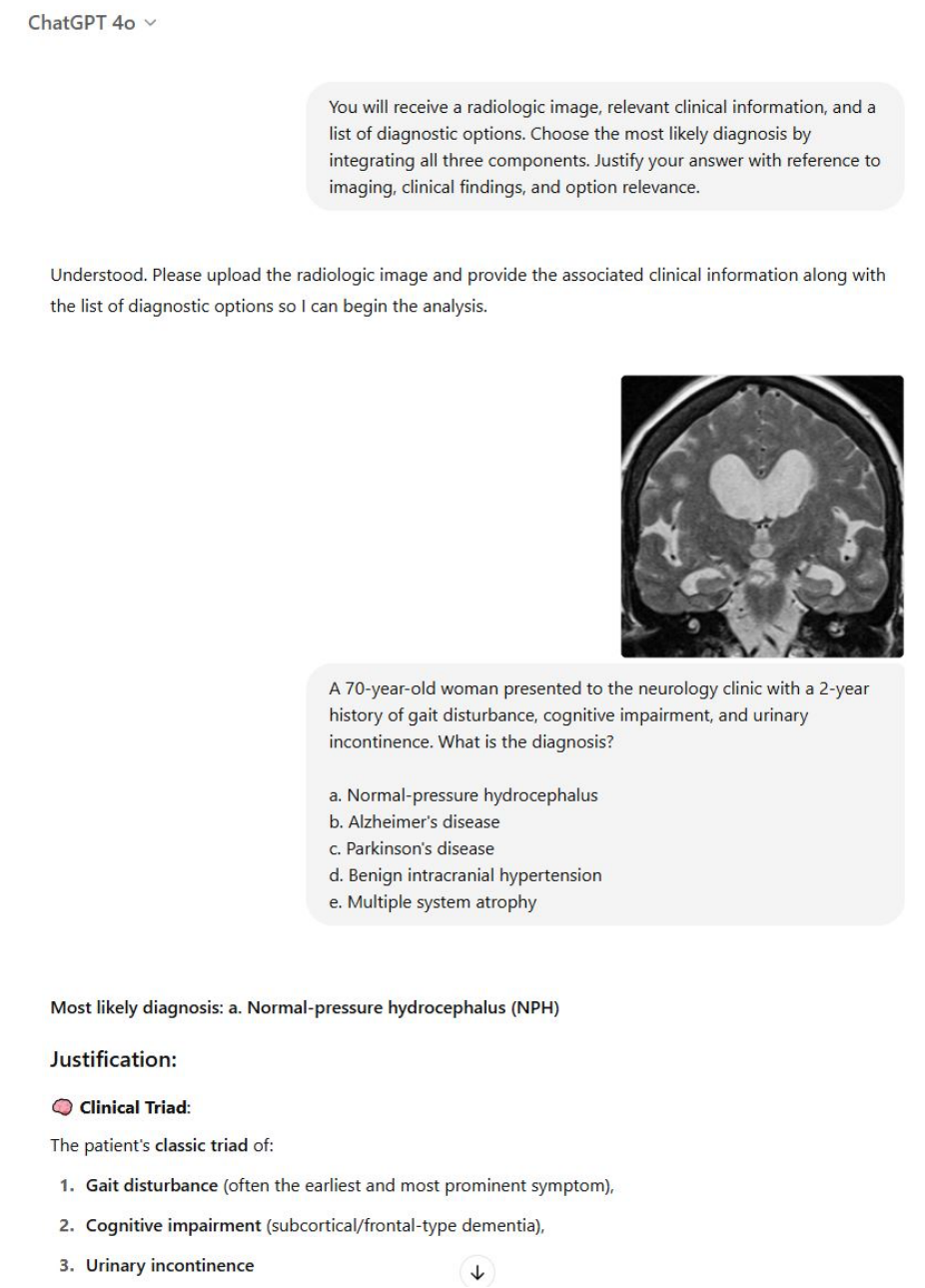


Figure 5. ChatGPT-4o query with radiological images, clinical information, and diagnostic options provided. ChatGPT-4o, Chat Generative Pre-trained Transformer.

Table 1. Accuracy rates of distinct accounts for each of the different input combinations				
Input combination	Account ¹	Account ²	Account ³	Average
Radiological image	19.38%	17.05%	23.26%	19.90%
Clinical information	41.86%	42.64%	49.61%	44.70%
Diagnostic options	10.85%	31.01%	20.16%	20.67%
Radiological image + clinical information	53.49%	50.39%	54.26%	52.71%
Radiological image + diagnostic options	59.69%	62.02%	62.02%	61.24%
Clinical information + diagnostic options	72.87%	76.74%	76.74%	75.45%
Radiological image + clinical information + diagnostic options	80.62%	79.07%	82.95%	80.88%

to improve accuracy, may help stabilize the model's internal decision processes across repeated runs. In our study, Fleiss' kappa scores were consistently higher in multimodal settings that combined images with clinical information and diagnostic options, indicating greater consistency across different

model instances. Similarly, Schramm et al.¹¹ reported that ChatGPT-4V produced more consistent responses in complex neuroimaging tasks when prompted with images supported by textual annotations. In contrast, Krishna et al.³ found that repeated prompts with identical text-only input often yielded

inconsistent outputs; however, they lacked investigations into image-included inputs. To our knowledge, this study is among the first to examine intramodel consistency systematically in a radiologic diagnostic setting.

Several studies have utilized open-access radiology datasets to evaluate LLM performance,^{4,28,29} raising the question of whether high accuracy in certain settings reflects genuine reasoning or simple memorization. In our study, the consistently low diagnostic accuracy in the "image only" condition—despite the use of publicly available cases—combined with the marked improvement following the addition of clinical context, strongly argues against memorization. These findings support the view that ChatGPT-4o relies on contextual understanding rather than the retrieval of previously encountered content.

Our comparative analysis involving radiology residents provided additional insight into the complementary strengths of human expertise and LLM-based diagnostic reasoning. Human readers outperformed the model in "image only" tasks, likely due to their training in visual pattern recognition—an area where current LLMs still fall short. Horiuchi et al.²¹ similarly found that ChatGPT did not reach the performance level of either residents or board-certified radiologists in challenging neuroradiology cases, consistent with our findings. Notably, our study also showed that ChatGPT-4o, when used as a supportive tool, improved the diagnostic accuracy of residents. In contrast, Mukherjee et al.⁴ reported no significant benefit when ChatGPT-4V was used to support human readers. This discrepancy may be explained by methodological differences; their study employed an earlier, less capable version of the model and included a smaller number of diagnostic cases, which may have limited statistical power to detect performance improvements.

Our study did not formally stratify question difficulty; however, the heterogeneous nature of our dataset inherently encompassed a wide range of case complexity. We did not label individual items as "easy" or "difficult"; however, the diagnostic accuracy achieved by radiology residents offers an indirect measure of relative difficulty. Recent studies have emphasized that LLMs, including ChatGPT-4 variants, tend to perform substantially worse on complex or ambiguous cases. For example, Horiuchi et al.²¹ reported significantly reduced LLM performance in challenging diagnostic tasks, highlighting

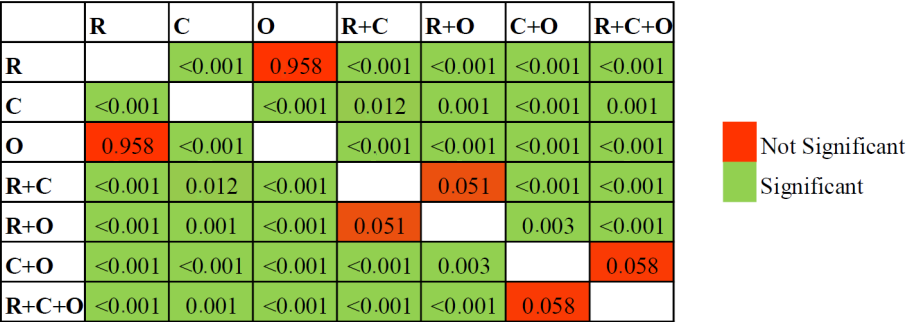


Figure 6. Wilcoxon signed rank test results for post-hoc pairwise comparisons. R, radiological image; C, clinical information; O, diagnostic options. Green: statistically significant ($P < 0.05$); Red: not significant.

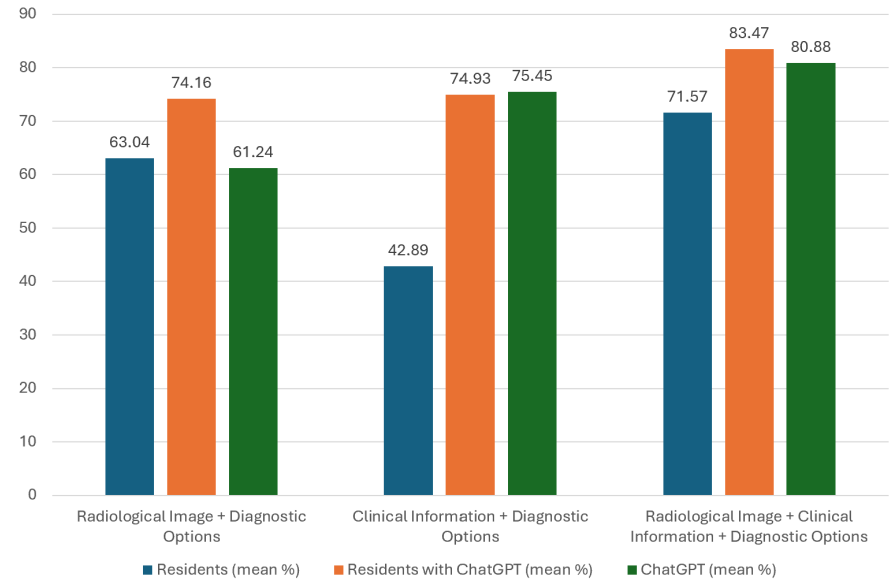


Figure 7. Mean diagnostic accuracy (%) of radiology residents, ChatGPT-4o, and residents with access to ChatGPT-4o responses ("wChatGPT") across three input configurations: "radiological image + diagnostic options," "clinical information + diagnostic options," and "radiological image + clinical information + diagnostic options." ChatGPT-4o outperformed residents in all settings except the "image only" evaluation, where residents had a slight advantage. In all conditions, residents showed a marked improvement when aided by ChatGPT-4o responses. ChatGPT-4o, Chat Generative Pre-trained Transformer.

Table 2. Interobserver agreement rates between accounts	
Input combination	Fleiss' kappa
Radiological image	0.497
Clinical information	0.613
Diagnostic options	0.023
Radiological image + clinical information	0.606
Radiological image + diagnostic options	0.652
Clinical information + diagnostic options	0.540
Radiological image + clinical information + diagnostic options	0.733

Notes: The Fleiss' kappa table shows interobserver variability for different input combinations. Substantial agreement is seen in the "radiological images + clinical information + options" combination.

persistent limitations under cognitive strain. Similarly, Bhayana et al.¹ identified question complexity as a key determinant of model accuracy. Some authors have further suggested that model performance can be optimized by adjusting temperature settings and prompt design strategies in accordance with question difficulty.³⁰ Collectively, these findings underscore the need for future benchmarking efforts to incorporate standardized complexity metrics to better evaluate LLM capabilities across varying levels of clinical challenge.

In comparison with prior literature, our study makes a unique contribution by systematically evaluating ChatGPT-4o's diagnostic accuracy across seven controlled input combinations and three independent repetitions. Unlike previous studies that focused on narrow clinical scenarios or language-specific settings, our methodology deliberately isolated each input type and assessed reproducibility through repeated testing.^{4,25} This design enabled a comprehensive evaluation of both diagnostic performance and output consistency, alongside a direct human comparison. Additionally, we included an input configuration containing only diagnostic options, without clinical context or image input. In contrast to our main input types, this allowed us to observe frequency bias and random selection tendencies in the model's decision-making indirectly—an aspect that, to our knowledge, has not been systematically explored in prior LLM-radiology studies. Moreover, none of the prior studies has made a post-hoc analysis regarding the inclusion of visual data input to already-provided clinical information and answer options. As a subtle but important difference, our study found that the addition of image data to clinical information and diagnostic options input did not improve accuracy; this contrasts with some prior assumptions suggesting that multimodal inputs would enhance performance, and, thus, offers a more nuanced understanding of ChatGPT's current limitations.

Overall, these findings highlight that although LLMs currently show strong potential in text-based reasoning tasks, their ability to interpret radiologic images independently remains limited. Although prior studies have reported mixed results regarding their role in radiologic decision-making, particularly in multimodal settings, it is evident that these models are still in the early stages of adapting to the complexities of visual diagnostic reasoning. As such, they may be more appropriately positioned as assistive tools, rather than autonomous diagnostic agents.

Several studies have suggested that LLMs can be beneficial when incorporated into structured decision-support systems, where clinical context is clearly defined and human oversight is preserved.^{2,14,28} However, their diagnostic precision declines markedly in the absence of textual input, reflecting a strong reliance on language-based information. This dependency raises important concerns about the safety, reproducibility, and clinical applicability of these models in radiological practice.

This study has several notable strengths. First, it employed a comprehensive and controlled experimental design that systematically evaluates ChatGPT-4o's diagnostic performance across seven different input combinations. This allowed us to assess not only diagnostic accuracy but also how the model responds to varying levels of contextual information. Second, by using three independent accounts for each scenario, we were able to measure interobserver variability, a rarely addressed aspect in previous LLM-focused radiology research. Third, the use of peer-reviewed multiple-choice diagnostic radiology questions enhances the clinical validity of our findings. Notably, to our knowledge, this is the first study to assess a prompt configuration that includes only diagnostic options, without accompanying clinical or imaging data, in the context of LLM performance in radiology. Finally, the study was designed and reported in accordance with the Minimum Information about Clinical Artificial Intelligence Modeling-LLM checklist,¹⁶ ensuring methodological rigor, transparency, and alignment with emerging standards in LLM evaluation.

Nevertheless, this study has several limitations. The radiological images provided to the model were static and isolated from the dynamic, interactive environment that is typical of clinical practice. Furthermore, even though the prompts were designed to isolate the effect of each input type, the interpretability of ChatGPT-4o's internal reasoning remains unclear. Finally, the forced-choice format used in the diagnostic tasks may not reflect the complexity of real-life radiological interpretation and reporting, which often involves open-ended reasoning and the integration of longitudinal patient data. Although case difficulty was not formally stratified to avoid subjective grading and maintain study focus, the comparison with radiology residents may serve as an indirect benchmark for overall difficulty, as previously discussed. In addition, although our dataset included a broad spectrum of

imaging modalities and anatomical regions, the case distribution was not balanced across these categories, potentially introducing subtle biases in modality or subspecialty representation. Furthermore, this study focused exclusively on ChatGPT-4o—currently among the most advanced and accessible multimodal LLMs—to ensure consistency in performance benchmarking under controlled conditions. Future studies should explore whether similar patterns hold across different models to assess the generalizability of our findings.

As LLMs continue to evolve, their integration into radiology is expected to become increasingly specialized and clinically aligned. A central priority for future development is improving the visual processing capabilities of these models through domain-specific multimodal training. Some studies underscore the potential of task-specified pre-trained models that are fine-tuned on radiology-specific image-text pairs, rather than relying solely on general-purpose architectures, such as ChatGPT.³¹ Another key area for advancement is the role of prompt engineering and fine-tuning strategies. Although our study used zero-shot prompting without any domain-specific customization, growing evidence suggests that carefully designed prompts and iterative fine-tuning can greatly influence diagnostic performance and consistency.^{9,12,32,33} As such, future research should systematically investigate how these factors can be optimized to balance model flexibility with clinical reliability, especially in high-stakes decision-making.

Beyond diagnostic reasoning, LLMs may also contribute to broader aspects of radiology practice, including structured reporting, generating patient-friendly summaries, prioritizing studies based on urgency, and integrating clinical context into triage or decision-support systems. Hybrid approaches that combine the interpretive strengths of LLMs with image-based AI tools or human oversight may provide a safer and more effective pathway toward real-world deployment. Finally, the use of LLMs in medical education deserves further exploration, particularly in simulation-based diagnostics and real-time AI-assisted feedback for trainees.

In conclusion, this study demonstrates that ChatGPT-4o's diagnostic accuracy in radiology is highly dependent on the presence of contextual information, particularly clinical data and structured diagnostic options. Its limited performance in "image only" scenarios and the lack of added value when ra-

diological images are paired with rich textual input emphasize the model's current limitations in autonomous radiological image interpretation. Even though ChatGPT-4o shows promise as a supportive tool in structured clinical environments—especially for triage or decision support—it cannot yet replace human expertise in image-driven diagnostic tasks. Future advancements must address the need for integrated, multimodal training and transparent reasoning mechanisms before these models can be safely adopted for real-world clinical radiology applications.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Materials 1 and 2:
<https://d2v96xfpocvxx.cloudfront.net/bda9171a-fae8-4995-8276-2138323f1e16/content-images/dfefbd7e-f7a2-43da-927b-e889250128eb.pdf>

References

- Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. 2023;307(5):230582. [Crossref]
- Li D, Gupta K, Bhaduri M, Sathiadoss P, Bhatnagar S, Chong J. Comparing GPT-3.5 and GPT-4 accuracy and drift in radiology diagnosis please cases. *Radiology*. 2024;310(1):232411. [Crossref]
- Krishna S, Bhambra N, Bleakney R, Bhayana R. Evaluation of reliability, repeatability, robustness, and confidence of GPT-3.5 and GPT-4 on a radiology board-style examination. *Radiology*. 2024;311(2):232715. [Crossref]
- Mukherjee P, Hou B, Suri A, et al. Evaluation of GPT large language model performance on RSNA 2023 case of the day questions. *Radiology*. 2024;313(1):240609. [Crossref]
- Hayden N, Gilbert S, Poisson LM, Griffith B, Klochko C. Performance of GPT-4 with vision on text- and image-based ACR diagnostic radiology in-training examination questions. *Radiology*. 2024;312(3):240153. [Crossref]
- Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology*. 2023;307(4):230424. [Crossref]
- Nguyen D, Rao A, Mazumder A, Succi MD. Exploring the accuracy of embedded ChatGPT-4 and ChatGPT-4o in generating BI-RADS scores: a pilot study in radiologic clinical support. *Clin Imaging*. 2025;117:110335. [Crossref]
- Doddi S, Hibshman T, Salichs O, et al. Assessing appropriate responses to ACR urologic imaging scenarios using ChatGPT and Bard. *Curr Probl Diagn Radiol*. 2024;53(2):226-229. [Crossref]
- Suh PS, Shim WH, Suh CH, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro vision using image inputs from diagnosis please cases. *Radiology*. 2024;312(1):240273. [Crossref]
- Brin D, Sorin V, Barash Y, et al. Assessing GPT-4 multimodal performance in radiological image analysis. *Eur Radiol*. 2024;35(4):1959-1965. [Crossref]
- Schramm S, Preis S, Metz MC, et al. Impact of multimodal prompt elements on diagnostic performance of GPT-4V in challenging brain MRI cases. *Radiology*. 2025;314(1):240689. [Crossref]
- Kim TT, Makutonin M, Sirous R, Javan R. Optimizing large language models in radiology and mitigating pitfalls: prompt engineering and fine-tuning. *RadioGraphics*. 2025;45(4):240073. [Crossref]
- Kao JP, Kao HT. Large language models in radiology: a technical and clinical perspective. *Eur J Radiol Artif Intell*. 2025;2:100021. [Crossref]
- Kuzan BN, Meşe İ, Yaşar S, Kuzan TY. A retrospective evaluation of the potential of ChatGPT in the accurate diagnosis of acute stroke. *Diagn Interv Radiol*. 2025;31(3):187-195. [Crossref]
- Kahalian S, Rajabzadeh M, Öçbe M, Medisoglu MS. ChatGPT-4.0 in oral and maxillofacial radiology: prediction of anatomical and pathological conditions from radiographic images. *Folia Med (Plovdiv)*. 2024;66(6):863-868. [Crossref]
- Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol*. 2024;25(10):865-868. [Crossref]
- The New England Journal of Medicine. June 26, 2025 [Internet]. Boston (MA): Massachusetts Medical Society; 2025 [cited 2025 May 10]. [Crossref]
- Home | RSNA Case Collection. Accessed May 10, 2025. [Crossref]
- Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*. 2024;310(1):232756. [Crossref]
- Güneş YC, Cesur T, Çamur E, Günbey Karabekmez L. Evaluating text and visual diagnostic capabilities of large language models on questions related to the Breast Imaging Reporting and Data System Atlas 5th edition. *Diagn Interv Radiol*. 2025;31(2):111-129. [Crossref]
- Horiuchi D, Tatekawa H, Oura T, et al. Comparing the diagnostic performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and radiologists in challenging neuroradiology cases. *Clin Neuroradiol*. 2024;34(4):779-787. [Crossref]
- Elek A, Ekizaloğlu DD, Güler E. Evaluating Microsoft Bing with ChatGPT-4 for the assessment of abdominal computed tomography and magnetic resonance images. *Diagn Interv Radiol*. 2025;31(3):196-205. [Crossref]
- Mitsuyama Y, Tatekawa H, Takita H, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. *Eur Radiol*. 2024;35(4):1938-1947. [Crossref]
- Kim SH, Schramm S, Adams LC, et al. Benchmarking the diagnostic performance of open source LLMs in 1933 Eurorad case reports. *NPJ Digit Med*. 2025;8(1):97. [Crossref]
- Sonoda Y, Kurokawa R, Nakamura Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. *Jpn J Radiol*. 2024;42(11):1231-1235. [Crossref]
- Fervers P, Hahnfeldt R, Kottlors J, et al. ChatGPT yields low accuracy in determining LI-RADS scores based on free-text and structured radiology reports in German language. *Front Radiol*. 2024;4:1390774. [Crossref]
- Perchik J, Godwin R, West J, Summerlin D, Zahid M, Smith A. Does ChatGPT pass the LIRADS test? Comparing quality of AI generated impressions to human reports. *Jpn J Gastroenterol*. 2023;10(5):1-5. [Crossref]
- Ueda D, Mitsuyama Y, Takita H, et al. Diagnostic Performance of ChatGPT from patient history and imaging findings on the diagnosis please quizzes. *Radiology*. 2023;308(1):231040. [Crossref]
- Suthar PP, Kounsai A, Chhetri L, Saini D, Dua SG. Artificial Intelligence (AI) in Radiology: a deep dive into ChatGPT 4.0's Accuracy with the American Journal of Neuroradiology's (AJNR) "case of the month." *Cureus*. 2023;15(8):43958. [Crossref]
- Milad D, Antaki F, Milad J, et al. Assessing the medical reasoning skills of GPT-4 in complex ophthalmology cases. *Br J Ophthalmol*. 2024;108(10):1398-1405. [Crossref]
- Lee RW, Lee KH, Yun JS, Kim MS, Choi HS. Comparative analysis of M4CXR, an LLM-based chest X-ray report generation model, and ChatGPT in radiological interpretation. *J Clin Med*. 2024;13(23):7057. [Crossref]
- Han T, Jeong WK, Shin J. Diagnostic performance of multimodal large language models in radiological quiz cases: the effects of prompt engineering and input conditions. *Ultrasonography*. 2025;44(3):220-231. [Crossref]
- Ray PP, Majumder P. ChatGPT in radiology: a deeper look into its limitations and potential pathways for improvement. *Can Assoc Radiol J*. 2024;75(1):202. [Crossref]



Potential radiation dose reduction in computed tomography with deep learning reconstruction: a retrospective monocentric study

Lucas Graber¹
 Melike Zeynep Akış²
 François Séverac³
 Luc Mertz⁴
 Serhat Akış²
 Catherine Roy¹
 Mickaël Ohana¹

¹Hôpital Civil, Department of Radiology, Strasbourg, France

²Dokuz Eylül University Hospital, Department of Radiology, İzmir, Türkiye

³Université de Strasbourg, Department of Biostatistics and Informatics, Strasbourg, France

⁴Université de Strasbourg, Department of Radiation Safety and Radiophysics, Strasbourg, France

PURPOSE

To evaluate whether deep learning reconstruction (DLR) can reduce the radiation dose in routine clinical computed tomography (CT) scans compared with iterative reconstruction (IR) while maintaining or improving image quality. The study assesses DLR's consistency and effectiveness across four distinct CT protocols—chest, head, chest–abdomen–pelvis (CAP) oncology, and lower limb CT angiography (CTA)—representing a wide range of clinical applications.

METHODS

Our study is retrospective and monocentric. It involves a total population of 13,060 patients who underwent a CT scan using either a DLR algorithm (CT-DLR) or an IR algorithm (CT-IR) in one of four different CT acquisition protocols. Image quality was evaluated qualitatively and quantitatively by measuring standardized signal-to-noise ratio and contrast-to-noise ratio values. Assessment was performed on a subsample of 200 patients (25 per protocol per group).

RESULTS

The overall reduction in radiation dose for the CT-DLR group compared with the CT-IR group was approximately 20%. By protocol, dose reductions were 22% for chest CT, 21% for CAP oncology, 20% for lower limb CTA, and 19% for head CT. The CT-DLR group exhibited superior subjective and objective image quality to the CT-IR group.

CONCLUSION

DLR algorithms allow for a significant reduction in radiation dose while achieving higher image quality compared with IR algorithms.

CLINICAL SIGNIFICANCE

This large-scale study confirms that DLR can significantly reduce the radiation dose in routine CT imaging while maintaining or enhancing diagnostic image quality. Its consistent performance across multiple protocols supports broader clinical adoption. Notably, the greatest dose reductions were observed in high-use protocols such as chest and CAP CT, underscoring DLR's potential to improve both individual patient care and long-term population-level radiation safety.

KEYWORDS

Deep learning reconstruction, radiation dose reduction, computed tomography scan, image quality

Handling editor: Burak Kocak

Corresponding author: Melike Zeynep Akış

E-mail: melikemza7@gmail.com

Received 16 June 2025; revision requested 27 July 2025;
last revision received 21 September 2025; accepted 24
October 2025.



Epub: 05.12.2025

Publication date: 01.09.2026

DOI: 10.4274/dir.2025.253513

The exponential growth in computational power over the last two decades has enabled substantial improvements in computed tomography (CT) image reconstruction techniques.¹ Statistical-based iterative reconstruction (IR) and model-based IR techniques were introduced between 2000 and 2010,²⁻⁴ resulting in a major reduction in image noise and the optimization of the signal-to-noise ratio (SNR) and the contrast-to-noise ratio (CNR) for lower irradiation doses,⁵⁻⁸ yet with mixed results regarding pure image quality.

Recently, new image reconstruction methods have been developed using artificial intelligence-based techniques, such as deep learning with convolutional neural networks.⁹ These neural networks are trained on tens of thousands of paired images to differentiate the noise

from the signal more efficiently, resulting in further improvements in image quality¹⁰⁻¹³ as well as a reduction in the radiation dose delivered.¹⁴⁻¹⁷

Although many studies have highlighted the superiority of deep learning reconstruction (DLR) algorithms in terms of image quality compared with previous algorithms, their potential for radiation dose reduction has also been confirmed. Kobayashi et al.¹⁸ demonstrated that the implementation of DLR reduced the radiation dose by 45% and lowered the estimated risk of radiation-induced cancer from 0.247% to 0.130%. Although their modeling approach illustrates the theoretical potential of DLR, our study focuses on real-world clinical protocols, revealing a more moderate reduction of 20%. This contrast underscores the complementary value of both perspectives and motivates the present study by addressing routine practice applicability.

The main objective of our study is to estimate the potential radiation dose reduction that can be achieved in a real-life setting using a DLR algorithm compared with an IR algorithm while maintaining equivalent image quality. Our secondary objective is to investigate the variation in radiation dose reduction potential as a function of body mass index (BMI) and patient sex.

Methods

Ethics

Our study was a retrospective analysis of prospectively collected clinical data. The study design was approved by the Ethics Committee of Strasbourg University Hospital (decision number: RNI-2021-149, date: 14.12.2021), and written consent from participating patients was not required.

Main points

- Deep learning reconstruction (DLR) significantly reduces the radiation dose in routine computed tomography (CT) scans compared with iterative reconstruction, with an average reduction of 20%.
- Despite lower radiation, DLR provides consistently better or equivalent image quality across all evaluated protocols: chest CT, head CT, chest–abdomen–pelvis oncology, and lower limb CT angiography.
- These findings support the clinical implementation of DLR as a safe and effective reconstruction method for dose optimization in a wide range of CT applications.

Population

This is a retrospective, single-center study analyzing prospectively collected dosimetry data from the Dose Archiving and Communication System (DACS) of the diagnostic radiology department of the Nouvel Hôpital Civil (Strasbourg University Hospital) between May 2018 and May 2021. All patients who underwent CT examinations during this period and had their radiation dose data stored on the DACS were considered for inclusion.

Specifically, we included patients who underwent CT examinations on either a third-generation wide-area detector CT scanner equipped with IR software [Aquilion One Genesis, Canon Medical Systems, Ōtawara, Tochigi, Japan, with three-dimensional adaptive iterative dose reduction (AIDR-3D)], which served as the control group (CT-IR), or a fourth-generation wide-area detector CT scanner equipped with DLR [Aquilion One Prism, Canon Medical Systems, with the *Advanced intelligent Clear-IQ Engine* (AiCE)], which served as the test group (CT-DLR). The Genesis system is equipped with PUREVISION detectors, the AIDR-3D hybrid IR algorithm, and SUREExposure 3D dose modulation, providing 16-cm z-axis coverage suitable for whole-organ imaging. The Prism system, although offering the same 16-cm z-axis coverage, incorporates improved PUREVISION Optics with lower electronic noise and higher light-collection efficiency,

enhanced dose-modulation features (including adaptive collimation and advanced exposure control), faster gantry rotation (up to 0.25 seconds), and optional single energy metal artifact reduction, in addition to the AiCE DLR algorithm. These hardware and software differences may have contributed to the observed reduction in radiation dose, in addition to the reconstruction algorithm itself.

Examinations had to meet predefined criteria, including being either a non-contrast chest CT, non-contrast head CT, chest–abdomen–pelvis (CAP) CT with injection, or CT angiography (CTA) of the lower limbs. Only examinations with a single series were included; thus, multiphasic CT was excluded.

A total of 13,060 patients were included in our study, with 4,361 patients in the CT-DLR group and 8,799 patients in the CT-IR group. Within the CT-DLR group, 1,567 patients were examined according to the chest CT protocol, 1,623 according to the head CT protocol, 1,109 according to the CAP oncology protocol, and 62 according to the lower limb CTA protocol. Within the CT-IR group, 3,060 patients were examined according to the chest CT protocol, 3,436 according to the head CT protocol, 2,058 according to the CAP oncology protocol, and 245 according to the lower limb CTA protocol (Figure 1).

To improve transparency and assess potential baseline imbalances between groups, we compiled a summary table of demo-

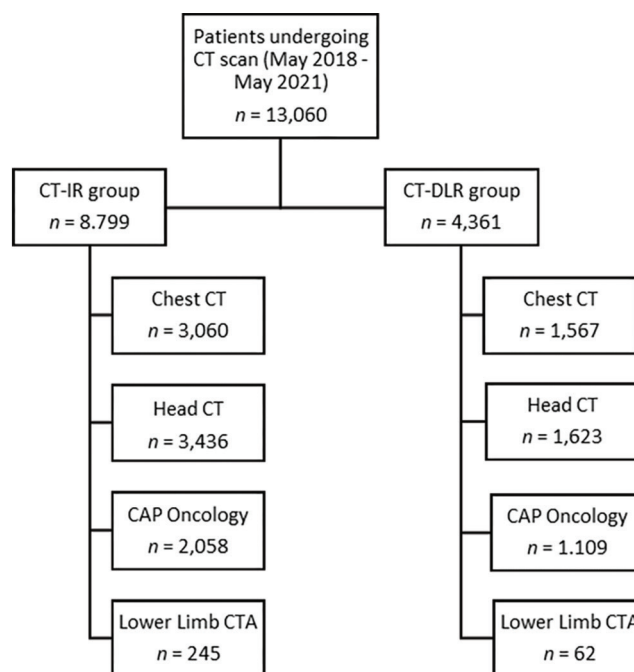


Figure 1. Flowchart of patient enrollment. CT-IR, computed tomography with iterative reconstruction; CT-DLR, computed tomography with deep learning reconstruction; CAP, chest–abdomen–pelvis; CTA, computed tomography angiography.

graphic characteristics, including age, sex distribution, BMI, and weight, for each protocol and reconstruction method (CT-IR and CT-DLR). These variables are known to affect radiation dose and were therefore reported to support comparability across groups (Table 1).

To homogenize data, examinations were selected based on anatomical area, contrast medium injection, and z-axis coverage (Table 2). The acquisition protocols were kept identical on both devices, except for the reconstruction algorithm used and the automatic z-axis dose modulation, which was adapted to these algorithms based on vendor recommendations.

Database

For each patient, we collected the date of the examination; the machine on which the examination was performed; the examination protocol; the patient's sex, weight, height, and BMI (if available); the dose-length product (DLP; mGy.cm); the CT dose index (mGy); the exposure time (ms); the acquisition field length (mm); the tube voltage

(kV); and the tension (mA). The radiation protection team of Strasbourg University Hospital conducted the data extraction.

Image quality assessment

For the specific qualitative and quantitative image analysis, we randomly selected 25 patients from the CT-IR group for each of the four examination protocols, using a stratified random sampling method with probabilities proportional to height. The stratification variable was BMI, which was categorized into four classes (BMI <20, 20 < BMI <25, 25 < BMI <30, BMI >30) to approximate clinical thresholds for underweight, normal weight, overweight, and obesity, in line with World Health Organization guidelines.¹⁹ These categories were chosen because patient body habitus substantially affects image noise and radiation dose modulation in CT examinations. This approach ensured that the sampled BMI distribution closely mirrored that of the overall study population.

Next, we selected 25 patients from the CT-DLR group using a 1:1 individual matching method that combined exact matching

for sex and the nearest neighbor method for BMI. A caliper of $\pm 0.5 \text{ kg/m}^2$ was applied during BMI matching to ensure comparable body habitus between CT-DLR and CT-IR groups.

A sample size of 25 patients per protocol and per group (CT-DLR and CT-IR) was deemed sufficient for the paired comparison of image quality metrics based on SNR, CNR, and Likert scores. This sample size represents a practical balance between statistical robustness and the feasibility of manual image evaluation.

Qualitative analysis

We conducted a subjective analysis of image quality using a Likert scale ranging from 1 to 5, following the European guidelines on image quality.²⁰ The analysis was performed by a single radiologist (LG, with 4 years of experience in CT). Evaluation criteria included visibility of relevant anatomical structures, sharpness of tissue boundaries, perceived image noise, soft tissue contrast resolution, and presence of reconstruction-related artifacts. The Likert scale categories were as follows:

Table 1. Baseline characteristics of patients in each protocol group, stratified by reconstruction method (IR vs. DLR)

	Protocol	Group	Age (mean)	Sex (male/female)	Body mass index (mean)	Weight (mean)
1.	Head CT	IR	76	17 M/8 F	26.73	81
2.	Head CT	DLR	72	17 M/8 F	26.64	81
3.	Chest CT	IR	51	17 M/8 F	25.97	77
4.	Chest CT	DLR	57	17 M/8 F	25.98	77
5.	CAP oncology	IR	61	13 M/12 F	27.89	83
6.	CAP oncology	DLR	62	13 M/12 F	27.17	81
7.	Lower limb CTA	IR	64	19 M/6 F	24.92	74
8.	Lower limb CTA	DLR	68	19 M/6 F	25.35	75

IR, iterative reconstruction; DLR, deep learning reconstruction; M, male; F, female; CT, computed tomography; CTA, CT angiography.

Table 2. Acquisition protocols according to the type of examination (identical on both devices)

Protocol	Chest CT	Head CT	CAP oncology	Lower limb CTA
Helical volume acquisition number	Single	Single	Single	Single
PoC injection prior to acquisition	No	No	70 seconds	30 seconds
Anatomical zone	Thorax	Skull	Chest–abdomen–pelvis	Abdomen and lower limbs
Minimum and maximum coverage (cm)	20–60	9–40	50–100	100–180
Collimation	0.5 × 80	0.5 × 40	0.5 × 80	0.5 × 40
Kilovoltage (kV)	Automatically defined according to scout	120	Automatically defined according to scout	100
Milliamperage (mA)	Modulation	Modulation	Modulation	Modulation
Rotation time (s)	0.275	0.75	0.5	0.4
Cutting thickness (mm)	1 every 0.8	1 every 0.8	1 every 0.8	1 every 0.8

PoC, point of contrast; CAP, chest–abdomen–pelvis; CTA, computed tomography angiography; CT, computed tomography.

1) non-diagnostic, 2) poor image quality, 3) limited image quality but diagnostic, 4) good image quality, and 5) optimal image quality.

Quantitative analysis

To ensure reproducibility and minimize inter-subject variability, all quantitative measurements (SNR and CNR) were standardized by using anatomically predefined, fixed-size regions of interest (ROIs). Background noise was consistently defined as the standard deviation of Hounsfield units within a homogeneous air region around the outside of the patient's body. This approach, adapted from prior validated methods, was based on Bhosale²¹ and Chatzaraki et al.²² for CAP oncology and lower limb CTA, Tang et al.²³ for chest CT, and Guziński et al.²⁴ for head CT.

Chest–abdomen–pelvis oncology and lower limb computed tomography angiography

Based on Bhosale,²¹ we placed an axial slice ROI of approximately 20 mm² in the right liver parenchyma behind the vena cava (Figure 2), with another ROI of approximately 20 mm² in the abdominal subcutaneous fat when possible (the size of the ROI was reduced if the thickness of the adipose tissue was insufficient) to measure the mean value. In cases where the adipose tissue thickness was insufficient, the size of the ROI was reduced accordingly. Although smaller ROIs are known to be more susceptible to local noise, potentially influencing the reproducibility of the SNR and CNR measurements—particularly in low-contrast regions—ROIs were carefully placed in homogeneous areas to mitigate this effect. This approach is consistent with the methodological considerations discussed in prior studies, including Chatzaraki et al.²² A similarly sized ROI was placed in the background around the outside of the patient to measure the distribution (standard deviation) of voxel values.

According to the literature review by Chatzaraki et al.,²² the most reliable and reproducible method for calculating the SNR and CNR in images is to use the distribution of voxel values in the background as the denominator.

The SNR is calculated as follows:

$$\frac{V_{\text{mean}}(\text{liver})}{\sigma(\text{background noise})}$$

The CNR is calculated as follows:

$$\frac{V_{\text{mean}}(\text{liver}) - V_{\text{mean}}(\text{fat})}{\sigma(\text{background noise})}$$

Chest computed tomography

We measured the mean voxel value by placing an ROI in the descending thoracic aorta, covering approximately half of the aortic lumen area in the axial section (Figure 2), similar to the study by Tang et al.²³ We also measured the mean voxel value in the adipose tissue of the left axillary region, using an ROI of at least 50 mm², and we placed an ROI of at least 50 mm² in the background around the outside of the patient to measure the voxel value distribution (standard deviation).

The SNR is calculated as follows:

$$\frac{V_{\text{mean}}(\text{Aorta})}{\sigma(\text{background noise})}$$

The CNR is calculated as follows:

$$\frac{V_{\text{mean}}(\text{Aorta}) - V_{\text{mean}}(\text{fat})}{\sigma(\text{background noise})}$$

Head computed tomography

We followed a method similar to that of Guziński et al.²⁴ We measured the mean voxel value in an ROI of at least 20 mm² located in the anterior portion of the right caudate nucleus in the axial slice (Figure 2). We also measured the mean voxel value in a similarly sized ROI of the fourth ventricle and an ROI with a minimum area of 10 mm² in the background around the outside of the patient to measure the distribution (standard deviation) of the voxel value.

The SNR is calculated as follows:

$$\frac{V_{\text{mean}}(\text{Basal ganglia})}{\sigma(\text{background noise})}$$

The CNR is calculated as follows:

$$\frac{V_{\text{mean}}(\text{Basal ganglia}) - V_{\text{mean}}(\text{CSF})}{\sigma(\text{background noise})}$$

The position of these ROIs was optimized to minimize the variation in voxel values within them.

Statistical analysis

General population

Continuous variables are reported as mean ± standard deviation, whereas categorical variables are presented as counts and proportions. To analyze the radiation exposure, linear regression models were used. Extreme values of radiation exposure (> 95th percentile) were excluded from the models,

as they may be caused by technical issues or multiple acquisitions during the same examination (such as patient movements and injection problems); their exclusion was intended to reduce the undue influence of outliers and improve the robustness of the analysis by focusing on typical cases.

To evaluate the independent effects of different factors, a multivariate model, including scanner type, protocol type, BMI, and sex, was constructed. Pairwise interactions between scanner type and other factors of interest were also included in the model to determine whether the effect of scanner type was dependent on other factors. The normality of the residuals was assessed using a quantile–quantile plot. The results are presented as differences with 95% confidence intervals (CIs).

Random samples

The analyses of image quality, SNR, and CNR were conducted on randomly selected samples with individual matching. Therefore, comparisons of the different variables of interest between the two types of scanners were performed using paired tests. Continuous variables were compared using the Wilcoxon signed-rank test, and categorical variables were compared using the Stuart–Maxwell test. A *P* value < 0.05 was considered statistically significant. The analyses were performed using R software version 4.1.1 (R Core Team, Vienna, Austria).

Results

Population

A total of 13,060 patients were included in the study—4,361 in the CT-DLR group and 8,799 in the CT-IR group—covering a range of clinical protocols: chest CT, head CT, CAP oncology, and lower limb CTA. Baseline demographic characteristics for each protocol and reconstruction group are summarized in Table 1.

Overall results

Compared with the CT-IR group, overall irradiation was significantly lower in the CT-DLR group, with a difference of 86.6 DLP units (95% CI: 78.5, 94.6; *P* < 0.001), corresponding to a radiation dose reduction of approximately 20%.

Further analysis by protocol showed an average reduction of 19.3% (95% CI: -21.6, -17) for head CT, 21.8% (95% CI: -24.1, -18.3) for chest CT, 20.9% (95% CI: -26.8, -20.2) for CAP

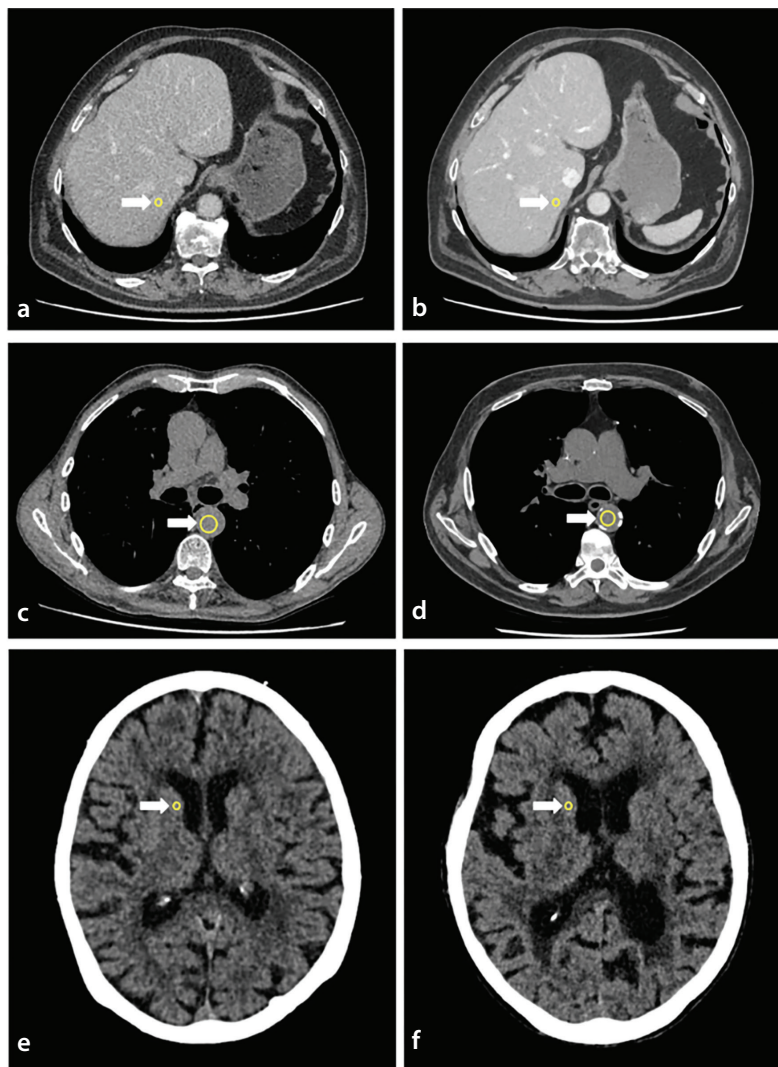


Figure 2. Representative axial computed tomography (CT) images from different anatomical regions and reconstruction algorithms: (a) and (b): Abdominal CT images from two different patients, reconstructed with iterative reconstruction (IR) and deep learning reconstruction (DLR), respectively. In both, a white arrow indicates a region of interest (ROI) placed in the right liver parenchyma posterior to the inferior vena cava: (c) and (d): Chest CT images reconstructed with IR and DLR, respectively. In both images, a white arrow indicates an ROI placed in the descending thoracic aorta in the axial section: (e) and (f): Head CT images from two different patients, reconstructed with IR and DLR, respectively. In both, a white arrow indicates an ROI placed in the anterior portion of the right caudate nucleus for mean voxel value measurement.

oncology, and 19.5% (95% CI: -30.4, -7.0) for lower limb CTA in the CT-DLR group (Figure 3).

Multivariate analysis adjusting for BMI showed that male patients had a mean excess radiation of 47.9 DLP units ($P < 0.001$), corresponding to a 12% difference in delivered radiation dose. However, there was no significant difference between the CT-DLR and CT-IR groups [47.8 (95% CI: 41.6, 54.1) vs. 47.9 (95% CI: 43.2, 52.7), $P = 0.965$], indicating that the dose reduction achieved with DLR was consistent across sexes.

In terms of BMI, there was no significant difference in irradiation between the two groups for the head CT and lower limb CTA protocols. However, for the chest CT protocol, patients with BMI >30 showed a significant increase in dose reduction with DLR

compared with those with BMI <20 [-62.1 (95% CI: -73.8, -49.1) vs. -30.7 (95% CI: -46.6, -14.9), $P = 0.003$]. Similarly, for the CAP oncology protocol, patients with BMI between 25 and 30 and those with BMI >30 showed a significant increase in dose reduction in the CT-DLR group compared with the CT-IR group [-115.8 (95% CI: -129.5, -102.1) and -131.0 (95% CI: -150.7, -111.4) vs. -57.6 (95% CI: -80.4, -34.9), $P < 0.001$ for both comparisons] (Figure 4).

Results samples

Table 3 displays the mean values for the SNR, CNR, and Likert scores used to assess image quality. Regardless of the acquisition protocol, the CT-DLR group exhibited significantly higher SNR, CNR, and Likert score values than the CT-IR group ($P < 0.001$).

Figure 5 displays the distribution of CNR values using a box-and-whisker plot.

Discussion

In a real-life retrospective analysis of over 10,000 CT examinations, we demonstrated that the use of DLR leads to a meaningful reduction in radiation dose compared with IR, with an average decrease of approximately 20%. This reduction was consistent and statistically significant across all four protocols, being most notable in the chest CT protocol and least notable in the head CT protocol. The observed differences remained significant even after adjusting for patient characteristics such as BMI and sex. Notably, dose reduction was greater in patients with higher BMI, especially in the CAP oncology

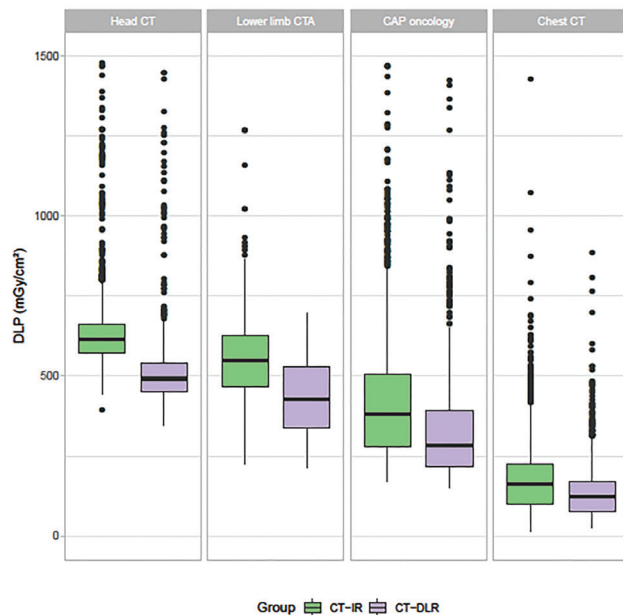


Figure 3. Box-and-whisker plot showing the distribution of patient dose by scanner type and acquisition protocol. DLP, dose-length product; CT-IR, computed tomography with iterative reconstruction; CT-DLR, computed tomography with deep learning reconstruction; CAP, chest–abdomen–pelvis; CTA, computed tomography angiography; CT, computed tomography.

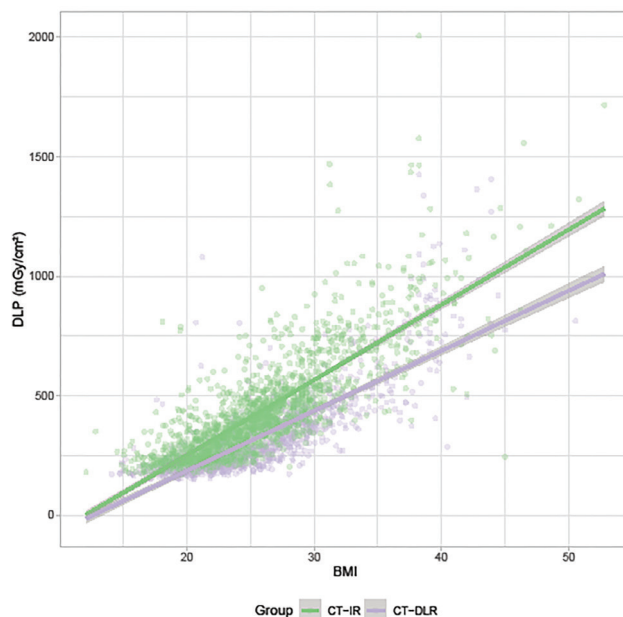


Figure 4. Correlation curves showing the radiation dose delivered to patients in relation to their body mass index. DLP, dose-length product; CT-IR, computed tomography with iterative reconstruction; CT-DLR, computed tomography with deep learning reconstruction; BMI, body mass index.

and chest CT protocols. Image quality was qualitatively and quantitatively evaluated as being at least equivalent to that of IR, confirming the results already published in the literature,^{16,17} while involving a larger number of patients.

The benefit of the DLR algorithm in terms of dose reduction compared with the IR algorithm appears to be slightly greater for the chest CT protocol than for the other protocols. This may be attributed to the lower attenuation of thoracic structures and the lung parenchyma having greater intrinsic contrast, making it more adapted to dose reduction.

Despite the dose reduction, we observed a superior objective (SNR, CNR) and subjective (Likert scale) image quality for each protocol, consistent with studies published in the literature.^{11–13} The DLR algorithm seems to have a greater advantage in terms of image quality for the chest CT and CAP oncology protocols than for the head CT and lower limb CTA protocols. This could be explained by the predominance of soft tissue on the thorax and abdomen, which the DLR algorithm is more effective at handling, compared with the lower limbs and skull, which have a larger proportion of bony structures. For lower limb CTA, however, the limited number of DLR cases resulted in wide CIs, and these findings should therefore be regarded as exploratory.

The clinical implications of this dose reduction are considerable. A recent large-scale risk-projection study by Smith-Bindman et al.²⁵ estimated that approximately 103,000 future cancers may result from the 93 million CT examinations performed in the United States in 2023 alone. Notably, the largest share of projected cancers was attributed to CAP and chest CT scans—protocols in which we observed the highest dose reductions using DLR.²⁵ Hence, a consistent 20% reduction in radiation dose across these high-utilization protocols could potentially contribute to reducing radiation-induced

Table 3. Average values of signal-to-noise ratio, contrast-to-noise ratio, and 5-point Likert score of image quality evaluated on the samples of the four protocols, according to the type of scanner

Protocol	Signal-to-noise ratio		Contrast-to-noise ratio		Image quality (Likert / 5)		P value*
	CT-DLR group	CT-IR group	CT-DLR group	CT-IR group	CT-DLR group	CT-IR group	
Chest CT	3.78	2.27	22.4	8.6	4.74	4	< 0.001
Lower limb CTA	61.49	46.24	78.74	58.24	4.8	4.32	< 0.001
CAP oncology	18.71	7.32	35.33	13.84	4.96	4.52	< 0.001
Head CT	13.7	13.3	13.6	12.6	4.36	3.88	< 0.001

*Wilcoxon signed-rank test (continuous variables) and Stuart–Maxwell test (categorical variables). CT-DLR, computed tomography with deep learning reconstruction; CT-IR, computed tomography with iterative reconstruction; CTA, computed tomography angiography; CAP, chest–abdomen–pelvis.

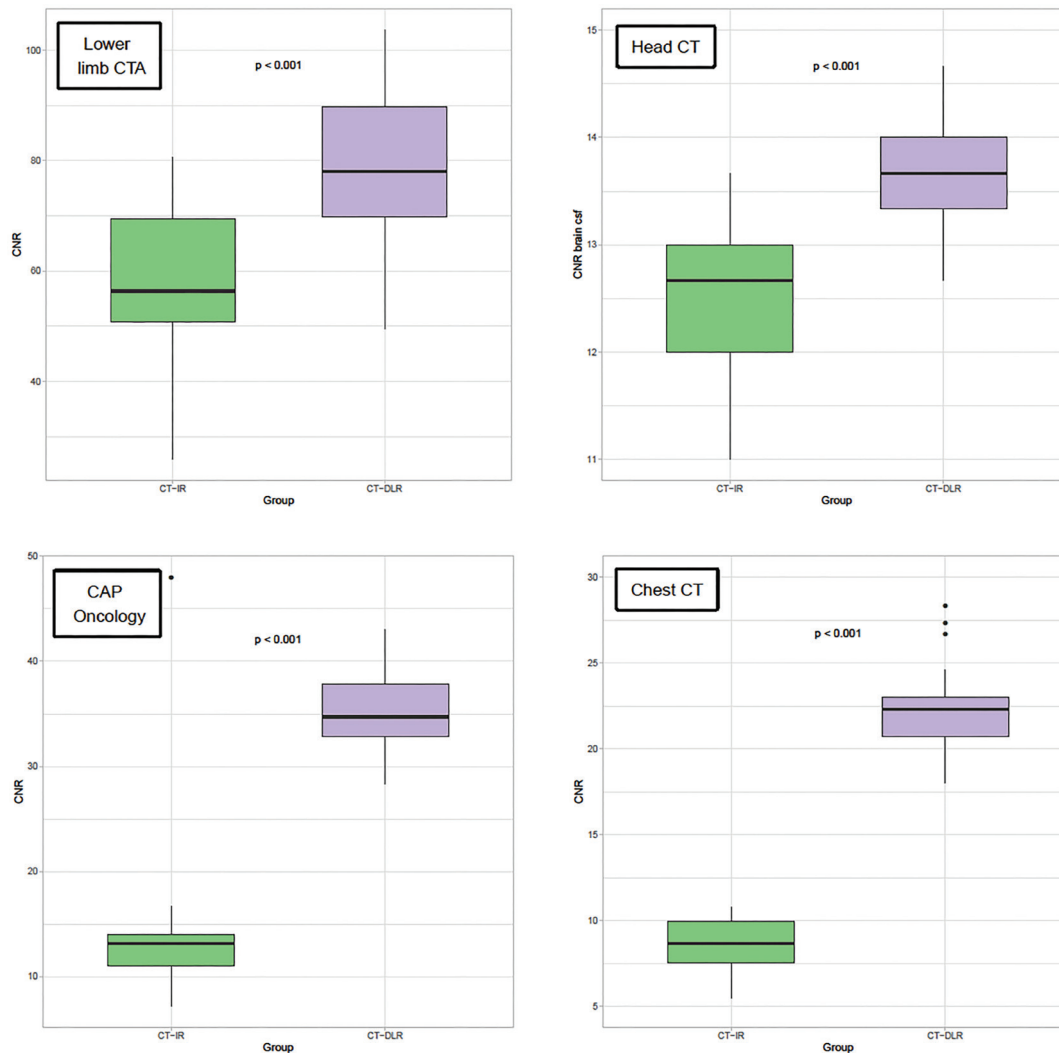


Figure 5. Box-and-whisker plot showing the distribution of the contrast-to-noise ratio for each scanner type and protocol. CTA, computed tomography angiography; CT, computed tomography; CAP, chest–abdomen–pelvis.

cancer risk at the population level. These findings highlight the dual benefit of DLR: achieving superior image quality while also supporting long-term cancer risk mitigation. A recent study by Kobayashi et al.¹⁸ demonstrated that the implementation of DLR could reduce the radiation dose by up to 45% and, importantly, lower the projected risk of radiation-induced cancer compared with IR. Taken together with our finding of a consistent 20% dose reduction in a large real-life cohort, these results underscore the broader clinical importance of adopting DLR, highlighting its potential not only for routine dose optimization but also for contributing to long-term patient safety.

Additionally, we found that male patients received, on average, 48 DLP units (approximately 12%) more radiation than female patients, independent of the type of examination, scanner, or BMI. This finding indicates a potential sex-related difference in radiation

dose. However, the current literature on this topic is limited, and further dedicated investigations are required to better understand the underlying factors and confirm this observation.

We were able to demonstrate that the difference in irradiation between the two scanners is significantly greater for higher BMIs in the CAP oncology and lower limb CTA protocols. This difference may be attributed to the DLR algorithm's superior ability to eliminate artifacts associated with anatomical factors related to body structure, thereby maintaining satisfactory image quality with more effective dose modulation.

Another important aspect to consider is that the CT-IR and CT-DLR groups were scanned on different generations of scanners (Aquilion One Genesis vs. Aquilion One Prism). Although acquisition protocols were identical, the fourth-generation Prism integrates incremental hardware advances such

as improved detector efficiency, lower electronic noise, and more sophisticated dose modulation. Consequently, the 20% reduction in radiation dose observed in our study should not be attributed solely to the DLR algorithm but rather to the combined contribution of both reconstruction and hardware improvements. This potential confounding effect is important when interpreting our results and may partly explain the variability in dose reduction observed across protocols and patient subgroups.

Our findings of an average 20% reduction in radiation dose with DLR differ from the 45% reduction reported by Kobayashi et al.,¹⁸ who additionally performed lifetime attributable risk modeling. Several factors may explain this discrepancy. First, scanner hardware generation differed: our analysis compared the Aquilion One Genesis (IR) with the newer Aquilion One Prism (DLR), whereas Kobayashi et al.¹⁸ employed the

Aquilion Precision. Such differences in detector design, dose modulation, and overall system architecture may partly confound the isolated contribution of the reconstruction algorithm. Second, protocol composition differed: our study included a broader range of routine protocols (head CT, chest CT, CAP oncology, and lower limb CTA), whereas Kobayashi et al.¹⁸ focused solely on body CT examinations (e.g., abdomen and pelvis), potentially delivering a different dose distribution. Finally, methodological differences must be acknowledged, particularly their use of risk-projection modeling vs. our emphasis on real-world clinical applicability. Taken together, these differences suggest that although absolute dose-reduction magnitudes may vary depending on study design, scanner generation, and protocol composition, the available evidence—including our large real-world cohort—supports the potential of DLR to meaningfully reduce patient radiation exposure.

Our study has several limitations that should be acknowledged. First, due to its retrospective design, it was not possible to strictly match patients between the CT-IR and CT-DLR groups. Additionally, the number of patients scanned with IR was higher, as the IR-equipped scanner was installed earlier and thus had a longer period of data collection. This temporal difference may have introduced slight variations in patient characteristics, possibly influenced by individual radiologists' preferences for one scanner over the other. Nonetheless, no statistically significant differences were observed between the two populations. As a further limitation, being a single-center non-randomized study, some degree of temporal or operator-related bias cannot be entirely excluded, particularly given the longer accrual period of the IR scanner.

Being a single-center study, the generalizability of our results to other institutions or populations may be limited despite the large sample size. Some inaccuracies in protocol labeling may also have occurred, such as injected head CTs being mislabeled or CAP oncology scans being performed without contrast. These errors did not affect the radiation dose measurements and only minimally influenced the SNR and CNR results in subgroup analyses.

A few extreme values were excluded during multivariate modeling to ensure robust analysis; however, the exclusion rate was balanced across both groups. In addition, qualitative and quantitative image assessments were performed by a single radiologist with 4 years of CT experience. Although

full blinding was not feasible due to the visually distinguishable differences between IR and DLR images, the consistency between subjective Likert scores and objective metrics (SNR, CNR), together with concordance with the existing literature, supports the validity of our findings. Furthermore, the small sample size in the lower limb CTA subgroup limits the precision of dose estimates, and these results should be interpreted with caution.

Finally, because IR and DLR data were acquired on scanners from different generations, hardware advances in the Prism system may have contributed to the observed dose reduction, representing a potential confounding factor.

In conclusion, in this retrospective study, we found that implementing DLR algorithms resulted in a mean 20% reduction in radiation dose compared with IR algorithms, across a wide range of patients and scanning protocols, with—at least—non-inferior qualitative and quantitative image quality. Future studies could explore the difference in radiation dose between men and women, independent of other factors, and determine the optimal balance between dose reduction and image quality enhancement offered by these DLR algorithms.

Acknowledgements

The authors would like to thank the Department of Radiology at both Hôpital Civil, Strasbourg and Dokuz Eylül University Hospital for their valuable support.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Willemink MJ, Noël PB. The evolution of image reconstruction for CT—from filtered back projection to artificial intelligence. *Eur Radiol*. 2019;29(5):2185-2195. [\[Crossref\]](#)
2. Patino M, Fuentes JM, Singh S, Hahn PF, Sahani DV. Iterative reconstruction techniques in abdominopelvic CT: technical concepts and clinical implementation. *AJR Am J Roentgenol*. 2015;205(1):W19-W31. [\[Crossref\]](#)
3. Southard RN, Bardo DME, Temkit MH, Thorkelson MA, Augustyn RA, Martinot CA. Comparison of iterative model reconstruction versus filtered back-projection in pediatric emergency head CT: dose, image quality, and image-reconstruction times. *AJNR Am J Neuroradiol*. 2019;40(5):866-871. [\[Crossref\]](#)

4. Shuman WP, Green DE, Busey JM, et al. Model-based iterative reconstruction versus adaptive statistical iterative reconstruction and filtered back projection in liver 64-MDCT: focal lesion detection, lesion conspicuity, and image noise. *AJR Am J Roentgenol*. 2013;200(5):1071-1076. [\[Crossref\]](#)
5. Frijia G. Recent updates in radioprotection. *Eur Radiol*. 2021;31(2):599-600. [\[Crossref\]](#)
6. Brady SL, Trout AT, Somasundaram E, Anton CG, Li Y, Dillman JR. Improving image quality and reducing radiation dose for pediatric CT by using deep learning reconstruction. *Radiology*. 2021;298(1):180-188. [\[Crossref\]](#)
7. Smith EA, Dillman JR, Goodsitt MM, Christodoulou EG, Keshavarzi N, Strouse PJ. Model-based iterative reconstruction: effect on patient radiation dose and image quality in pediatric body CT. *Radiology*. 2014;270(2):526-534. [\[Crossref\]](#)
8. Andersen HK, Völgyes D, Martinsen ACT. Image quality with iterative reconstruction techniques in CT of the lungs—a phantom study. *Eur J Radiol Open*. 2018;5:35-40. [\[Crossref\]](#)
9. Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine learning and deep learning in medical imaging: intelligent imaging. *J Med Imaging Radiat Sci*. 2019;50(4):477-487. [\[Crossref\]](#)
10. Akagi M, Nakamura Y, Higaki T, et al. Deep learning reconstruction improves image quality of abdominal ultra-high-resolution CT. *Eur Radiol*. 2019;29(11):6163-6171. Erratum in: *Eur Radiol*. 2019;29(8):4526-4527. [\[Crossref\]](#)
11. Shirasaka T, Kojima T, Funama Y, et al. Image quality improvement with deep learning-based reconstruction on abdominal ultrahigh-resolution CT: a phantom study. *J Appl Clin Med Phys*. 2021;22(7):286-296. [\[Crossref\]](#)
12. Singh R, Digumarthy SR, Muse VV, et al. Image quality and lesion detection on deep learning reconstruction and iterative reconstruction of submillisievert chest and abdominal CT. *AJR Am J Roentgenol*. 2020;214(3):566-573. [\[Crossref\]](#)
13. Liu P, Wang M, Wang Y, et al. Impact of deep learning-based optimization algorithm on image quality of low-dose coronary CT angiography with noise reduction: a prospective study. *Acad Radiol*. 2020;27(9):1241-1248. [\[Crossref\]](#)
14. Greffier J, Hamard A, Pereira F, et al. Image quality and dose reduction opportunity of deep learning image reconstruction algorithm for CT: a phantom study. *Eur Radiol*. 2020;30(7):3951-3959. [\[Crossref\]](#)
15. Nagayama Y, Goto M, Sakabe D, et al. Radiation dose optimization potential of deep learning-based reconstruction for multiphase hepatic CT: a clinical and phantom study. *Eur J Radiol*. 2022;151:110280. [\[Crossref\]](#)
16. Nam JG, Hong JH, Kim DS, Oh J, Goo JM. Deep learning reconstruction for contrast-enhanced

- CT of the upper abdomen: similar image quality with lower radiation dose in direct comparison with iterative reconstruction. *Eur Radiol.* 2021;31(8):5533-5543. [\[Crossref\]](#)
17. Benz DC, Ersözlü S, Mojon FLA, et al. Radiation dose reduction with deep-learning image reconstruction for coronary computed tomography angiography. *Eur Radiol.* 2022;32(4):2620-2628. [\[Crossref\]](#)
 18. Kobayashi N, Nakaura T, Yoshida N, et al. Impact of deep learning reconstruction on radiation dose reduction and cancer risk in CT examinations: a real-world clinical analysis. *Eur Radiol.* 2025;35(6):3499-3507. [\[Crossref\]](#)
 19. Obesity: preventing and managing the global epidemic. Report of a WHO consultation. *World Health Organ Tech Rep Ser.* 2000;894:i-xii, 1-253. [\[Crossref\]](#)
 20. European Commission. European guidelines on quality criteria for computed tomography [Internet]. Luxembourg: Publications Office of the European Union; 2000 [cited 2025 Nov 5]. [\[Crossref\]](#)
 21. Bhosale P. Comparing CNR, SNR, and image quality of CT images reconstructed with soft kernel, standard kernel, and standard kernel plus ASiR 30% techniques. *Int J Radiol.* 2015;2(2):60-65. [\[Crossref\]](#)
 22. Chatzaraki V, Kubik-Huch RA, Thali M, Niemann T. Quantifying image quality in chest computed tomography angiography: evaluation of different contrast-to-noise ratio measurement methods. *Acta Radiol.* 2022;63(10):1353-1362. [\[Crossref\]](#)
 23. Tang H, Yu N, Jia Y, et al. Assessment of noise reduction potential and image quality improvement of a new generation adaptive statistical iterative reconstruction (ASiR-V) in chest CT. *Br J Radiol.* 2018;91(1081):20170521. [\[Crossref\]](#)
 24. Guziński M, Waszczuk Ł, Sasiadek MJ. Head CT: image quality improvement of posterior fossa and radiation dose reduction with ASiR - comparative studies of CT head examinations. *Eur Radiol.* 2016;26(10):3691-3696. [\[Crossref\]](#)
 25. Smith-Bindman R, Chu PW, Azman Firdaus H, et al. Projected lifetime cancer risks from current computed tomography imaging. *JAMA Intern Med.* 2025;185(6):710-719. [\[Crossref\]](#)



Multi-parametric magnetic resonance imaging-based radiomics for differentiation of skull base osteomyelitis from locally advanced nasopharyngeal carcinoma: a multi-center external validation study

Firat Atak¹
Hanife Avcı²
Yeliz Pekçevik³
Ayça Karaosmanoğlu⁴

¹Mamak State Hospital, Clinic of Radiology, Ankara, Türkiye

²Hacettepe University Faculty of Medicine, Department of Biostatistics, Ankara, Türkiye

³University of Health Sciences Türkiye, Tepecik Training and Research Hospital, Clinic of Radiology, Division of Neuroradiology and Head and Neck Radiology, İzmir, Türkiye

⁴Hacettepe University Hospital, Department of Radiology, Division of Neuroradiology and Head and Neck Radiology, Ankara, Türkiye

The findings of this study were presented as a scientific poster at ECR 2025; however, the full work has not been published elsewhere.

Handling editor: Burak Kocak

Corresponding author: Ayça Karaosmanoğlu

E-mail: ayca.akgoz@gmail.com

Received 23 July 2025; revision requested 03 September 2025; accepted 24 October 2025.



Epub: 05.12.2025

Publication date: 01.09.2026

DOI: 10.4274/dir.2025.253574

PURPOSE

Skull base osteomyelitis (SBO) and nasopharyngeal carcinoma (NPCa) are challenging to differentiate due to overlapping clinical and radiological features. This study aimed to develop and validate a multi-parametric magnetic resonance imaging (MRI)-based radiomics model with high sensitivity, enabling reliable diagnosis of SBO in adult patients presenting with equivocal imaging findings.

METHODS

This was a retrospective, multicenter study using institutional data. The training cohort, comprising 63 adult patients from two classes (31 SBO, 32 NPCa) with MRI data, was used for model development and optimization. An external test set ($n = 30$; 12 SBO, 18 NPCa) obtained from two different clinical centers was used for model performance analysis and generalizability. Lesion segmentation was performed using a manual volumetric technique on three axial MRI sequences (pre-contrast T1-weighted, fat-suppressed T2-weighted, and post-contrast fat-suppressed T1-weighted). Hand-crafted radiomic features ($n = 2,553$) were extracted using the Pyradiomics library. A multi-step process was used to select the final features, including reproducibility analysis using an intra-class correlation coefficient threshold of 0.9, pairwise Spearman correlation analysis with a threshold of 0.8 to reduce redundancy, and least absolute shrinkage and selection operator regression. The final set of five features were used to train six machine learning models. The models were internally validated using 5-fold cross-validation, and performance was confirmed using the unseen external test set. Traditional statistical tests, including the Mann-Whitney U test and chi-squared test, were used to compare baseline characteristics, with a P value of <0.05 considered significant.

RESULTS

Among the evaluated classifiers, the random forest model demonstrated the best diagnostic performance, yielding the highest area under the curve (AUC) value in the 5-fold cross-validation analysis. In the external test set, the semantic model demonstrated the best diagnostic performance, achieving an AUC of 0.940 [95% confidence interval (CI): 0.857–1.00], followed by the radiomics model (AUC: 0.903, 95% CI: 0.784–1). The apparent diffusion coefficient (ADC)-based model demonstrated limited discriminative ability (AUC: 0.694, 95% CI: 0.497–0.892). The difference between the semantic and radiomics models did not reach statistical significance ($P = 0.644$), whereas both significantly outperformed the ADC model ($P < 0.05$).

CONCLUSION

Radiomics achieved high and consistent performance in distinguishing SBO from advanced NPCa. Although expert-based semantic assessment performed slightly better, radiomics provides an objective alternative. ADC-based methods showed limited generalizability due to inter-center variability.

CLINICAL SIGNIFICANCE

Our study confirms the importance of expert radiologist assessment while demonstrating that radiomics offers a comparably effective and objective decision-support tool. Its ability to provide a consistent, quantitative output is particularly valuable for standardizing the diagnostic approach and empowering less experienced radiologists to make more confident assessments.

KEYWORDS

Radiomics, skull base osteomyelitis, magnetic resonance imaging, nasopharyngeal carcinoma, hand-crafted

Skull base osteomyelitis (SBO) is defined as an infection of the skull base bones, which often develops as a complication of external otitis. Although it is frequently considered to be synonymous with necrotizing otitis externa, radiologists prefer the broader term SBO because it avoids specific etiological assumptions. SBO can be classified into two subtypes: typical and atypical (central).¹ Typical SBO characteristically arises secondary to external otitis.^{2,3} Atypical SBO occurs without typical clinical presentation or a preceding ear infection, posing a diagnostic challenge. It may arise from local infections other than the ear (paranasal sinus or soft tissue) or be idiopathic.¹

Despite advances in diagnostic techniques and antibiotic regimens, SBO remains a potentially fatal disease. A recent systematic review reported high rates of treatment failure (22%), relapse (7%), and disease-specific mortality (2%) for otogenic SBO.⁴ Neurological sequelae are common, with rates reported as high as 48%.^{3,5}

Tissue sampling is often required to rule out neoplasms but often yields non-specific results, necessitating repeated biopsies.^{1,3} Microbiologic culture of tissue samples plays an important role in establishing a definitive diagnosis of SBO, but an average of 15.8% of cases may yield false-negative culture results.² Although computed tomography remains the preferred imaging modality, it has limited sensitivity and may not detect subtle early-stage changes;⁶ Tc-99m methylene diphosphonate scintigraphy can overcome this limitation but lacks specificity.^{3,7}

Main points

- Magnetic resonance imaging-based radiomics model demonstrated excellent diagnostic accuracy for differentiating skull base osteomyelitis (SBO) from nasopharyngeal carcinoma (NPca), showing statistically superior performance compared with diffusion parameters.
- Semantic assessment demonstrated the best performance; however, given their qualitative and subjective nature, radiomics may provide an important complementary role by offering quantitative and objective information.
- This is the first radiomics study aiming to differentiate SBO from NPca.
- Apparent diffusion coefficient (ADC) threshold values often lack generalizability due to center-specific variability, limiting their use as quantitative biomarkers; however, ADC remains important for qualitative assessment.

Magnetic resonance imaging (MRI) provides superior visualization of soft tissue and intracranial involvement, but differentiation from malignant pathologies can be challenging without typical findings, especially for inexperienced readers.⁸⁻¹² Given the absence of a definitive imaging modality for SBO, an often time-consuming multimodal approach is typically necessary to establish an accurate diagnosis. Consequently, the average time from symptom onset to diagnosis is approximately 2 months.² Previous studies have demonstrated that early diagnosis and aggressive treatment can improve disease outcomes.⁵ Therefore, accurate diagnosis of SBO is essential for appropriate management, with recent studies showing promise in using qualitative and quantitative MRI parameters, including diffusion and dynamic contrast enhancement (DCE) perfusion metrics for distinction.⁸⁻¹⁰

Radiomics, the extraction of quantitative data from medical images, has recently gained significant attention in radiology and oncology research.^{13,14} The applications include, but are not limited to, differential diagnosis, prediction of clinical outcomes, and treatment response, as well as characterization of tumor gene expression patterns (radiogenomics).¹⁵

To date, no prior radiomics studies have investigated the differentiation between SBO and nasopharyngeal carcinoma (NPca). This study evaluates whether MRI radiomics features can distinguish these conditions.

Methods

Study design and ethical approval

This multicenter, retrospective case-control study was conducted and reported in adherence with the CheckList for Evaluation of AI in Radiology guidelines (Supplementary Text).¹⁶ The study received approval from the Hacettepe University Clinical Research Ethics Committee (decision number: GO 23/631, date: 03.10.2023), which waived the need for informed patient consent due to the retrospective nature of the analysis. All patient data were de-identified to ensure data protection.

Patient cohort

Patients who underwent head and neck MRI at one of three institutions between January 2005 and December 2023 were retrospectively screened. The study included adult patients with a final diagnosis of either SBO or NPca. Patients were required to have a

definitive diagnosis of SBO or NPca and have the necessary MRI sequences available [axial pre-contrast T1 (T1w), fat-suppressed T2 (FS-T2w), and post-contrast fat-suppressed T1 (T1c)]. Patients were excluded for the following reasons: 1) significant motion artifacts affecting lesion evaluation ($n = 3$); 2) SBO confined to the temporal bone without further skull base extension ($n = 8$); 3) prior radiotherapy for NPca ($n = 10$); or 4) NPca with a tumor stage $< T3$ ($n = 23$). Patients with NPca with T-stage $< T3$ were excluded because skull base involvement is a defining feature of stage T3 disease, and the diagnostic challenge under investigation arises only in this clinical context. The reference standard for NPca was a biopsy-confirmed diagnosis for all patients. The diagnosis of SBO was established based on a combination of the following: 1) compatible clinical presentation; 2) histopathology showing chronic inflammation; 3) microbiological culture isolating non-skin flora pathogens; and 4) clinical and radiological evidence of treatment response on follow-up.

The NPca cohort included 33 (66%) T3 and 17 (34%) T4 tumors; pathologically, 95.5% of analyzed cases (43/45) were of the non-keratinized subtype (the remaining five patients were diagnosed at another center). Within the SBO cohort, seven patients (16.3%) were classified as atypical, presenting with central clivus involvement without evidence of otogenic infection.

Data

The study included 93 patients (43 SBO, 50 NPca). The training set comprised 63 patients (31 SBO, 32 NPca) from a single institution (Center 1). All feature selection, hyperparameter tuning, and model training were performed exclusively on this set. The external test set comprised an independent cohort of 30 patients from two different institutions (Center 2: $n = 14$; Center 3: $n = 16$), consisting of 12 SBO and 18 NPca cases. This set was used only once for the final evaluation of the selected models (Figure 1).

Systematic differences between the training and test cohorts, such as variations in MRI scanners and acquisition protocols, were inherent to the multicenter design and served to evaluate the model's generalizability rigorously (Supplementary Text). The classes in the training set (31 SBO, 32 NPca) were well-balanced; therefore, no specific techniques for handling class imbalance were applied.

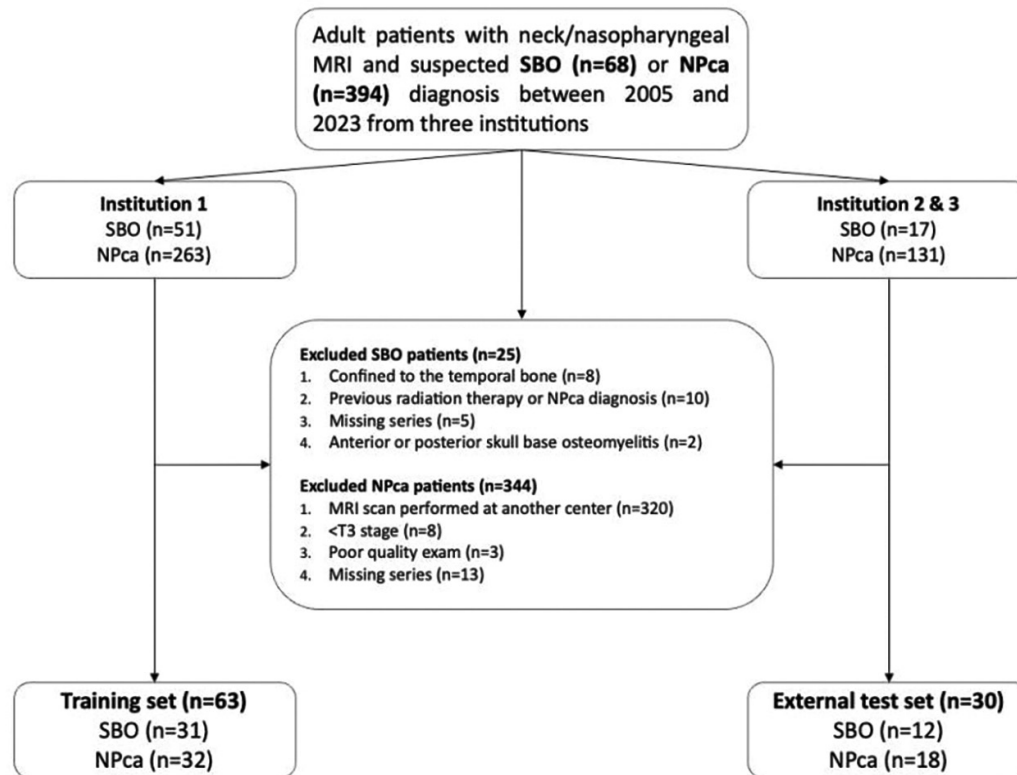


Figure 1. Flowchart of the study. SBO, skull base osteomyelitis; NPca, nasopharyngeal carcinoma; MRI, magnetic resonance imaging.

Non-radiomics data collection

Two radiologists (F.A., A.K.), blinded to diagnoses, evaluated nine predefined semantic MRI features, including the presence of lateral extension, abscess or fluid collection, architectural distortion, lymphadenopathy, increased T2 signal, marked contrast enhancement, dural venous thrombosis, dural thickening, and otologic inflammation (Supplementary Figure 1).

Lateral extension was considered positive if an abnormal signal was detected lateral to the temporomandibular joint.¹⁰ An abscess was considered positive only if diffusion restriction was present. Lesions exhibiting T2 hypointensity, ring-like contrast enhancement, facilitated diffusion, and T2 hyper- or hypointensity were classified as fluid collections. Lymphadenopathy was defined as retropharyngeal lymph nodes measuring >0.8 cm or cervical nodes measuring >1.1 cm in the shortest diameter.¹⁰ Increased T2 signal intensity and marked contrast enhancement were assessed by comparing the lesion signal characteristics with nasal mucosal tissue.¹⁰ Particular attention was given to the predominant signal features exhibited by the majority of the lesion volume rather than focal variations. Otologic inflammation was considered positive if contrast enhancement was observed in the ear cavities on post-con-

trast images and negative only in the presence of effusion.

A radiologist (F.A.), blinded to diagnoses, measured apparent diffusion coefficient (ADC) values by placing three 0.2 cm² regions of interest (ROIs) on ADC maps, excluding cystic or necrotic areas, and targeting the visually lowest signal regions of the lesion. The arithmetic mean of the ADC values (ADCmean) was recorded. Reference ADC (rADC) values were calculated by averaging measurements from ROIs placed over the spinal cord across two subsequent slices. Normalized ADC (nADC) values were derived by dividing ADCmean by rADC (Supplementary Figure 2). Diffusion-weighted imaging was unavailable in six patients (6.45%).

Radiomics workflow

The complete radiomics pipeline is illustrated in Figure 2.

Segmentation

A radiologist (F.A., with 7 years of experience) performed manual volumetric segmentation of the entire lesion on T1w, FS-T2w, and T1c sequences using Slicer 3D 5.2.2 (open source, <https://www.slicer.org>) (Figure 2).¹⁷ The segmentation volumes were comparable between the two groups, with a

mean volume of 29,057.97 mm³ for SBO and 17,282.68 mm³ for NPca.

Image preprocessing

The images were resampled to an isotropic voxel size of 2 × 2 × 2 mm³ using BSpline interpolation. Voxel signal intensities were normalized using z-score normalization (3σ) with a scaling factor of 200. To improve the reliability of texture feature calculations, voxels with intensities outside the range of $\mu \pm 3\sigma$ within the segmentation mask were excluded from the radiomics analysis.¹⁸ Gray-level discretization was performed using a fixed bin width (FBW) approach (Supplementary Figure 3). The bin width was set to 5, which in our data typically resulted in approximately 30–130 bins per volume of interest. This range has been suggested in the Pyradiomics documentation as a practical guideline to ensure that discretized features remain informative while retaining reproducibility.¹⁹ Although there are currently no universally accepted standards for MRI, FBW was considered the most appropriate choice following intensity normalization.^{19–21} Moreover, several reports have emphasized that selecting lower bin width values may enhance feature reproducibility.²¹

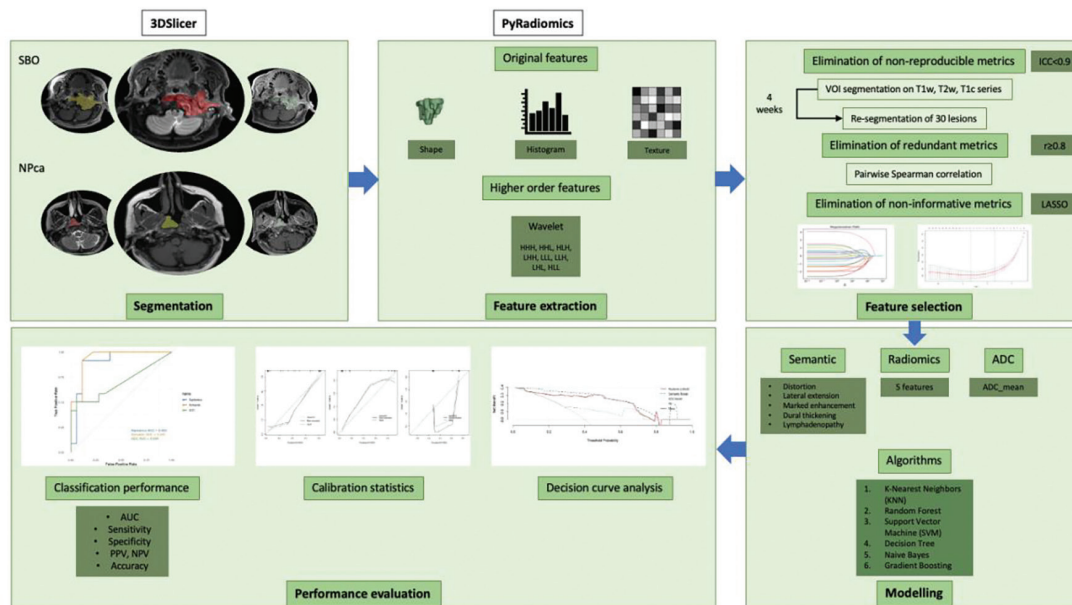


Figure 2. Radiomics workflow. SBO, skull base osteomyelitis; NPca, nasopharyngeal carcinoma; ICC, intraclass correlation coefficient; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve; ADC, apparent diffusion coefficient.

Feature extraction

From each of the three MRI sequences, 851 radiomic features were extracted, resulting in a total of 2,553 hand-crafted features per patient. These included shape descriptors, first-order statistics, texture features, and wavelet-transformed features. Feature extraction was performed using Pyradiomics (v3.0.1),¹⁹ an open-source Python package. All Pyradiomics settings not explicitly mentioned remained at their default configuration.

Feature selection and modeling

In order to reduce the risk of overfitting and select robust features, a multi-step feature selection was performed. To ensure reproducibility, two radiologists (F.A., A.K.) independently segmented a random subset of 30 cases. Radiomic features with poor inter-rater reliability [intraclass correlation coefficient (ICC) < 0.90] were excluded, yielding 1,064 reproducible features. Subsequently, pairwise Spearman correlation analysis was conducted to address multicollinearity; among highly correlated feature pairs ($|\rho| > 0.80$), the feature with the lower univariate discriminative power was discarded, resulting in 79 non-redundant features. The remaining feature set was then subjected to the binomial least absolute shrinkage and selection operator (LASSO) ($\alpha = 1$), with the regularization parameter (λ) optimized via leave-one-out cross-validation using the minimum cross-validated deviance criterion. To avoid overfitting, the 1-standard error rule

was applied, resulting in the selection of 5 features (Supplementary Table 1).

Six machine learning (ML) classifiers [random forest (RF), support vector machine (SVM), XGBoost, Decision Tree (DT), Naïve Bayes, k-nearest neighbors (kNN)] were trained. Model training and hyperparameter tuning were performed on the training set using 5-fold cross-validation. The classifier with the highest area under the curve (AUC) value during cross-validation was chosen as the final model for each feature set. During the entire model development process, the test dataset remained completely apart from the other data and was used only for final performance evaluation, thereby preventing the possibility of data leakage.

Univariate logistic regression analysis was performed to identify significant imaging predictors for the non-radiomic semantic and ADC features. Subsequently, semantic and ADC models were developed based on these selected predictors using the best ML algorithm.

Performance evaluation

The final models (semantic, ADC, and radiomics) were evaluated on the external test set using sensitivity, specificity, accuracy, balanced accuracy, positive predictive value, negative predictive value, and F1-score. Discriminative ability was assessed by generating receiver operating characteristic (ROC) curves and calculating the AUC. Pairwise comparisons of AUCs were conducted using

the DeLong test to determine statistical significance. Model calibration was evaluated with calibration curves, and clinical utility was assessed using decision curve analysis (DCA). Finally, radiomics model interpretability was enhanced by applying SHapley Additive exPlanations (SHAP), which allows quantification of the contribution of each radiomic feature to model predictions.

Statistical analysis

All statistical analyses were performed using R (v4.3.2; R Foundation for Statistical Computing, Vienna, Austria) and SPSS (v23.0; IBM Corp., Armonk, NY, USA) under the supervision of an experienced data analyst. Normality of continuous variables was assessed using the Shapiro–Wilk test, and homogeneity of variance was tested using Levene's test. Descriptive statistics were reported as median (25th percentile–75th percentile) for continuous variables and as frequencies (percentages) for categorical variables. Differences between groups were compared using the Student's t-test or Mann–Whitney U test for continuous variables, and the chi-squared or Fisher's exact test for categorical variables. Univariate logistic regression was used to identify significant predictors. A *P* value of <0.05 was considered statistically significant.

For radiomics feature selection, ICC was calculated using the “irr” and “lpSolve” packages, and redundant features were removed with pairwise Spearman correlation. Binomial LASSO regression was performed with the “glmnet” package. The RF classifier was im-

plemented using the “randomForest” package with 1,000 trees and mtry tuned across values of 2–4, with the best performance identified at mtry = 3. In addition, five other supervised ML algorithms were developed within the “caret” framework: SVM (“e1071”), Naïve Bayes (“e1071”), XGBoost (“xgboost”), kNN (“class”), and DT (“rpart”).

Model performance was assessed using 5-fold cross-validation in the training set and confirmed in the external test set. Discrimination was evaluated by ROC curves generated with the “pROC” package, and

AUCs were compared using DeLong’s test. Calibration was assessed using calibration plots (“rms” package), and clinical utility was evaluated by DCA (“rmda” package). Model interpretability was enhanced with SHAP implemented via the “fastshap,” “shapviz,” and “ggplot2” packages.

Results

Patient characteristics

The baseline radiological characteristics of the 93 patients are summarized in Table 1.

Patients with SBO had a mean age of 65 ± 11 years, whereas those with NPca had a mean age of 50 ± 14 years. Both groups were predominantly male (SBO ratio: 4:1; NPca ratio: 3.5:1).

A comparison of the training and test cohorts revealed no significant differences in the distribution of semantic MRI findings ($P > 0.05$). However, a statistically significant difference was observed in the ADC metrics (ADCmean and nADC) between the two cohorts ($P < 0.05$).

Table 1. Comparison of semantic MRI findings and ADC values between training and test sets

	Training set (n = 63)			Test set (n = 30)			
	SBO	NPca	P value	SBO	NPca	P value	P value
	(n = 31)	(n = 32)		(n = 12)	(n = 18)		
ADC variables							
ADCmean (10 ⁻³ mm ² /s)	1.078 (0.988–1.183)	0.699 (0.605–0.799)	< 0.001 ^b	0.967 ± 0.309	0.726 ± 0.118	0.023 ^a	0.037 ^b
nADC	1.33 (1.17–1.55)	0.92 (0.81–1.03)	< 0.001 ^b	1.14 ± 0.30	0.73 ± 0.15	< 0.001 ^a	< 0.001 ^b
Semantic MRI variables							
Architectural distortion							
No	27 (87.1%)	7 (21.9%)	< 0.001 ^c	10 (83.3%)	3 (16.7%)	< 0.001 ^c	0.461 ^c
Yes	4 (12.9%)	25 (78.1%)		2 (16.7%)	15 (83.3%)		
Involvement of lateral structures							
No	13 (41.9%)	30 (93.8%)	< 0.001 ^c	4 (33.3%)	18 (100%)	< 0.001 ^c	0.797 ^c
Yes	18 (58.1%)	2 (6.3%)		8 (66.7%)	0		
T2-signal							
Low	23 (74.2%)	32 (100%)	0.002 ^c	9 (75%)	18 (100%)	0.054 ^c	> 0.05 ^c
High	8 (25.8%)	0		3 (25%)	0		
Enhancement							
Low	11 (35.5%)	28 (87.5%)	< 0.001 ^c	2 (16.7%)	17 (94.4%)	< 0.001 ^c	> 0.05 ^c
High	20 (64.5%)	4 (12.5%)		10 (83.3%)	1 (5.6%)		
Dural thickening							
No	13 (41.9%)	19 (59.4%)	0.258 ^c	4 (33.3%)	16 (88.9%)	0.004 ^c	0.223 ^c
Yes	18 (58.1%)	13 (40.6%)		8 (66.7%)	2 (20%)		
Venous thrombosis							
No	23 (74.2%)	32 (100%)	0.002 ^c	9 (75%)	18 (100%)	0.054 ^c	0.973 ^c
Yes	8 (25.8%)	0		3 (25%)	0		
Abscess/fluid collection							
No	17 (54.8%)	31 (96.9%)	< 0.001 ^c	7 (58.3%)	18 (100%)	0.006 ^c	0.607 ^c
Yes	14 (45.2%)	1 (3.1%)		5 (41.7%)	0		
Lymphadenopathy							
No	27 (87.1%)	4 (12.5%)	< 0.001 ^c	8 (66.7%)	4 (22.2%)	0.024 ^c	0.542 ^c
Yes	4 (12.9%)	28 (87.5%)		4 (33.3%)	14 (77.8%)		
Otologic inflammation							
No	5 (16.1%)	31 (96.9%)	< 0.001 ^c	4 (33.3%)	18 (100%)	< 0.001 ^c	0.201 ^c
Yes	26 (83.9%)	1 (3.1%)		8 (66.7%)	0		

Data are presented mean ± SD or median (25th percentile–75th percentile), as appropriate. Categorical variables reported as frequency (percent). Differences between groups were analyzed using: ^a, Independent t-test; ^b, Mann–Whitney U test; ^c, Yates continuity correction chi-squared test; ^d, Fisher’s exact test. MRI, magnetic resonance imaging; ADC, apparent diffusion coefficient; SBO, skull base osteomyelitis; NPca, nasopharyngeal carcinoma; ADCmean, mean of the ADC values; nADC, normalized apparent diffusion coefficient; SD, standard deviation.

Non-radiomic data analysis

All nine semantic MRI findings showed statistically significant differences between the groups ($P < 0.05$, Table 1). Lateral extension, high T2 signal, marked enhancement, dural venous thrombosis, dural thickening, fluid loculation/abscess, and ear inflammation were more common in SBO, whereas architectural distortion and lymphadenopathy were more frequent in NPca. The most common MRI findings in SBO included absence of architectural distortion ($n = 27$, 87.1%), presence of ear inflammation ($n = 26$, 83.9%), and presence of marked enhancement ($n = 20$, 64.5%). Based on univariate logistic regression, five significant predictors of SBO were identified: presence of lateral extension [odds ratio (OR): 36.71], marked enhancement (OR: 20.77), dural thickening (OR: 3.57), absence of architectural distortion (OR: 0.04), and absence of cervical lymphadenopathy (OR: 0.044) (Supplementary Table 2).

In the training set, both the ADCmean and nADC values were significantly higher in the SBO group than in the NPca group ($P < 0.001$). The ADCmean (OR: 1.012) value was identified as a significant predictor of SBO based on univariate logistic regression ($P < 0.001$) (Table 2).

Cut-off values for ADCmean and nADC, determined using the Youden index, were $0.954 \times 10^{-3} \text{ mm}^2/\text{s}$ and $1.133 \times 10^{-3} \text{ mm}^2/\text{s}$, respectively. These predefined thresholds demonstrated good diagnostic performance in the training set [ADCmean: AUC = 0.849, 95% confidence interval (CI): 0.65–1.00; nADC:

AUC: 0.897, 95% CI: 0.70–1.00], but performance declined in the external test set (ADCmean: AUC: 0.688, 95% CI: 0.49–0.88; nADC: AUC: 0.705, 95% CI: 0.51–0.90) (Table 2).

Radiomics model performance

The performance of six candidate classifiers was assessed using 5-fold cross-validation. Among these, the RF achieved the highest mean AUC (0.985, 95% CI: 0.956–1.00) and was selected as a final model (Table 3).

Comparative analysis of final models

The top-performing radiomics RF model was compared with two other models developed using the same RF classifier: a semantic model (using the five selected qualitative features) and an ADC model (using ADCmean). The performance of all three models on the external test set is detailed in Table 4. The semantic model achieved the highest diagnostic performance (AUC: 0.940, 95% CI: 0.857–1.00), followed by the radiomics model (AUC: 0.903, 95% CI: 0.784–1.00) and ADC model (AUC: 0.694, 95% CI: 0.497–0.892). The difference between the semantic and radiomics models did not reach statistical significance ($P = 0.644$), whereas both significantly outperformed the ADC model ($P < 0.05$) (Figure 3). The confusion matrices corresponding to the three evaluated models are presented in Supplementary Figure 4.

Clinical utility and model interpretability

The calibration curves indicated that the radiomics model was well-calibrated, with

its bias-corrected curve closely tracking the ideal line for predicted probabilities between approximately 0.20 and 0.85 (Figure 4). The semantic model also demonstrated good overall calibration, though with a slight tendency to overestimate risk at very high probabilities (> 0.8) (Figure 5). In contrast, the ADC model was poorly calibrated, showing considerable underestimation in the mid-range and erratic overestimation at the upper end of the probability scale.

DCA demonstrated that both the radiomics and semantic models provided substantial clinical utility over default strategies within a 10%–70% threshold range. The semantic model was slightly superior at lower thresholds, whereas the radiomics model showed comparable or greater benefit in the mid-range. Conversely, the ADC model offered minimal net benefit, indicating a general lack of clinical utility. The clinical relevance of all models decreased markedly beyond a 0.70 threshold.

The SHAP analysis revealed that $t2_wavelet_LLL_firstorder_Minimum$ was the most influential feature in model predictions, exerting a strong negative contribution to the probability of SBO (i.e., lower values were associated with a higher likelihood of SBO). In contrast, $t1_original_firstorder_Skewness$ demonstrated a positive contribution, indicating that greater skewness increased the predicted probability of SBO (Supplementary Figure 5). The models' outputs and reference standard were presented through the use of sample cases to enhance the transparency of the models (Figures 6, 7).

Table 2. Performance analysis of ADC parameters in training and external test cohorts

Parameter	Cut-off	Dataset	Accuracy	Sensitivity	Specificity	AUC (95% CI)
ADCmean	$0.954 \times 10^{-3} \text{ mm}^2/\text{s}$	Training	0.91 (0.84–0.98)	0.84 (0.69–0.98)	1.00 (0.89–1.00)	0.85 (0.65–1.00)
		Test	0.77 (0.61–0.92)	0.50 (0.22–0.78)	0.94 (0.84–1.00)	0.69 (0.49–0.88)
nADC	1.133	Training	0.90 (0.82–0.97)	0.87 (0.74–1.00)	0.92 (0.84–1.00)	0.89 (0.70–1.00)
		Test	0.80 (0.66–0.94)	0.50 (0.22–0.78)	1.00 (0.81–1.00)	0.70 (0.51–0.90)

AUC, area under the curve; CI, confidence interval; ADCmean, mean apparent diffusion coefficient; nADC, normalized apparent diffusion coefficient.

Table 3. Five-fold cross-validation classification performance of six different modeling strategies

Model	Performance measures							
	Accuracy (95% CI)	Sensitivity	Specificity	PPV	NPV	AUC (95% CI)	BA	F1
Support vector machine	0.921 (0.824–0.974)	0.936	0.906	0.906	0.936	0.968 (0.920–1.00)	0.921	0.921
Naive bayes	0.936 (0.845–0.982)	0.968	0.906	0.909	0.967	0.962 (0.917–1.00)	0.937	0.937
Random forest	0.921 (0.824–0.974)	0.968	0.875	0.882	0.966	0.985 (0.956–1.00)	0.921	0.923
K-nearest neighbors	0.952 (0.867–0.990)	0.967	0.936	0.938	0.968	0.979 (0.952–1.00)	0.953	0.952
XGBoost	0.905 (0.804–0.964)	0.936	0.875	0.879	0.933	0.968 (0.928–1.00)	0.905	0.906
Decision tree	0.889 (0.784–0.954)	0.903	0.875	0.875	0.903	0.914 (0.840–0.989)	0.889	0.889

CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve; BA, balanced accuracy.

Table 4. Diagnostic performances of semantic, ADC, and radiomics (random forest) models in the cross-validation and external test set for differentiating skull base osteomyelitis from nasopharyngeal carcinoma

Model	Dataset	Accuracy (95% CI)	Sens.	Spec.	PPV	NPV	AUC (95% CI)	BA	F1
Semantic	Training (CV)	0.888 (0.784–0.954)	0.871	0.906	0.900	0.878	0.953 (0.900–1.00)	0.888	0.885
	Test	0.900 (0.735–0.978)	0.916	0.888	0.846	0.941	0.940 (0.857–1.00)	0.903	0.880
Radiomics	Training (CV)	0.921 (0.824–0.974)	0.968	0.875	0.882	0.966	0.985 (0.956–1.00)	0.921	0.923
	Test	0.800 (0.614–0.923)	0.917	0.722	0.688	0.929	0.903 (0.784–1.00)	0.819	0.786
ADC-only	Training (CV)	0.929 (0.830–0.980)	0.903	0.961	0.965	0.892	0.981 (0.955–1.00)	0.932	0.933
	Test	0.700 (0.506–0.852)	0.500	0.833	0.667	0.714	0.694 (0.497–0.892)	0.667	0.571

ADC, apparent diffusion coefficient; CI, confidence interval; Sens., sensitivity; Spec., specificity; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve; BA, balanced accuracy; CV, cross-validation.

Discussion

In this study, we developed and externally validated an MRI-based radiomics model to differentiate SBO from locally advanced NPCa. The model demonstrated robust and

generalizable performance on the external test set (AUC: 0.903), comparable with prior literature using quantitative DCE and ADC parameters.^{8,9} Importantly, only modest performance degradation was observed in the

independent cohort, highlighting the generalizability and robustness of our radiomics model. Notably, although a semantic model based on expert interpretation achieved a slightly higher performance (AUC: 0.940), the principal value of our radiomics model lies in its objectivity. By providing a reproducible, quantitative analysis, it can complement the radiologist's assessment, reduce inter-reader variability, and serve as a valuable decision-support tool, particularly for less experienced readers.

The diagnostic challenge of distinguishing SBO from NPCa is well-documented, as both can present with similar features, such as asymmetric soft tissue thickening and cranial nerve palsy.^{8,10–12} Our findings on semantic MRI features align with and expand upon previous work. We reproduced many of the key indicators of SBO identified by Goh et al.¹⁰ However, we noted some discrepancies. For instance, lateral extension was seen in 60.5% of our SBO cases, lower than the 92% reported by Goh et al.,¹⁰ which may be due to our inclusion of more non-otogenic

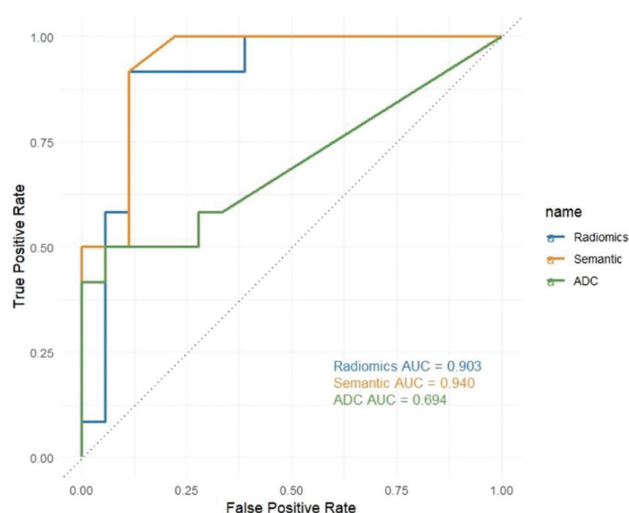


Figure 3. Receiver operating characteristic curves (ROC) and area under the curve (AUC) values of apparent diffusion coefficient (ADC), semantic, and radiomics models in the external test set.

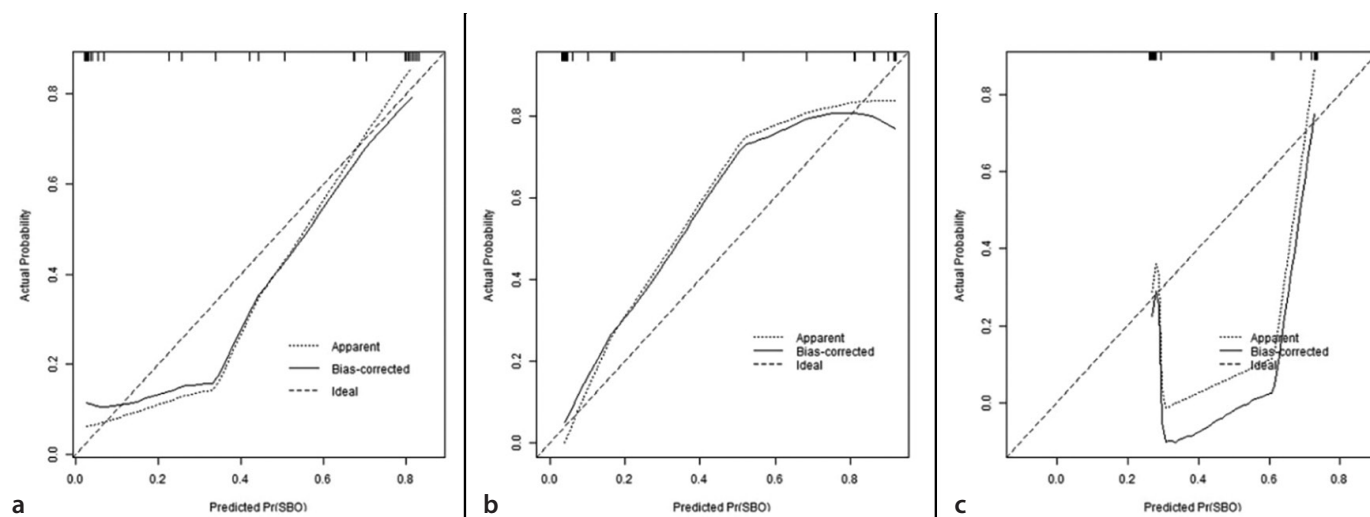


Figure 4. Calibration plots for the random forest models using radiomics features (a), semantic features (b), and apparent diffusion coefficient measurements (c). Each plot shows the calibration of predicted probabilities for the class “SBO” against the observed actual probabilities. The “Apparent” curve represents the raw calibration, and the “Bias-corrected” curve accounts for overfitting. The “Ideal” line indicates perfect calibration, where predicted probabilities exactly match observed outcomes. SBO, skull base osteomyelitis.

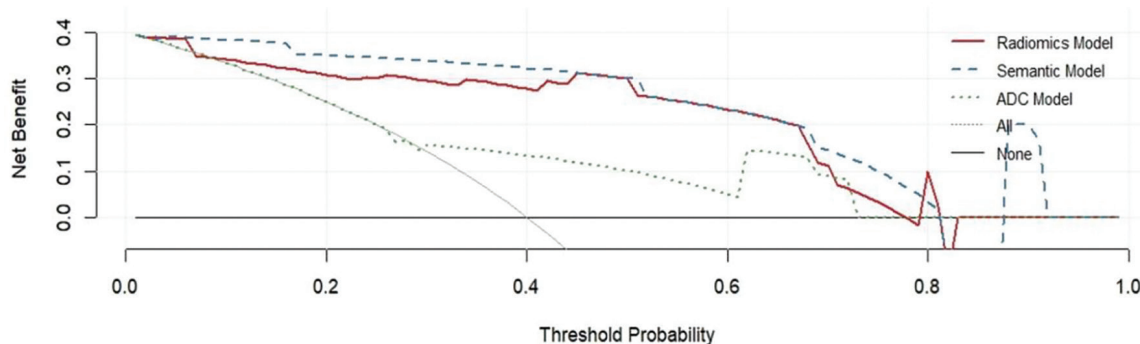


Figure 5. Decision curve analysis of the three predictive models. The plot compares the net benefit of the radiomics model (red line), the semantic model (blue dashed line), and the apparent diffusion coefficient-only model (green dotted line) across a range of threshold probabilities. The horizontal black line ("None") represents the net benefit of treating no patients, and the thin gray line ("All") represents the net benefit of treating all patients. The model with the highest curve for a given threshold probability range offers the greatest clinical utility. ADC, apparent diffusion coefficient.

SBO cases. Similarly, the prevalence of high T2 signal intensity in our SBO cohort (25.6%) was lower than in other studies.^{10,22} This likely reflects the assessment methodology and pathological heterogeneity of SBO, which

can include not only edema but also fibrosis and necrosis, both of which can result in a low T2 signal.^{7,23,24} Furthermore, the time interval between symptom onset and imaging may have influenced the T2 signal intensity.

Patients imaged at more advanced stages may exhibit a lower T2 signal due to chronic inflammation and fibrosis. Architectural distortion was noted in 14% of our SBO patients, which is higher than the previous reports in

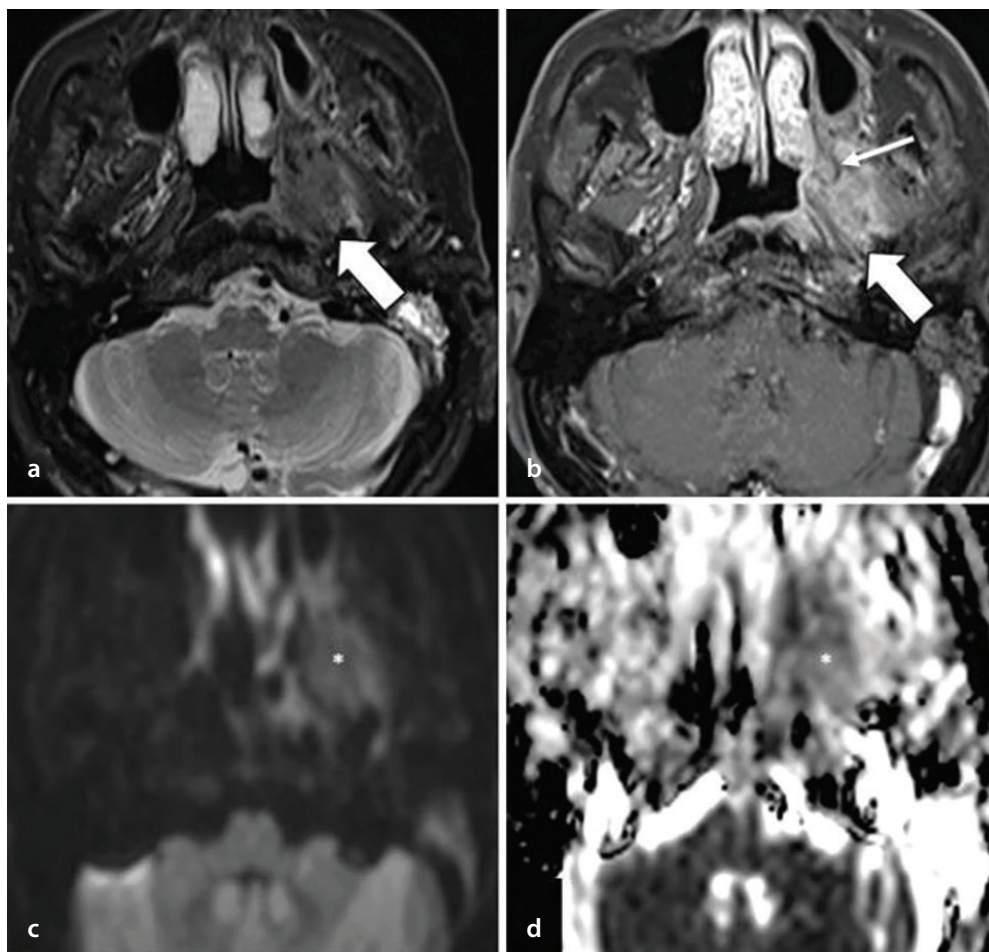


Figure 6. A patient with histologically confirmed nasopharyngeal carcinoma misclassified as skull base osteomyelitis by the semantic model but correctly identified by the radiomics model. (a) Axial fat-suppressed T2-weighted image showing a homogeneous, hypointense lesion relative to nasal mucosa with mild architectural distortion (arrow). Left mastoid cells show effusion without inflammatory enhancement (not shown). (b) Post-contrast T1-weighted image demonstrating an infiltrative, enhancing lesion in the left parapharyngeal space (thick arrow) extending into the pterygoid body (thin arrow), pterygopalatine fossa, and floor of the middle cranial fossa (not shown). The left Rosenmüller fossa is compressed without mucosal irregularity. Lateral structures, including the temporomandibular joint and parotid space, are unremarkable. The lesion enhancement pattern is similar to adjacent mucosa. (c, d) Diffusion-weighted imaging reveals relatively restricted diffusion (asterisks) with an apparent diffusion coefficient (ADC) of $0.86 \times 10^{-3} \text{ mm}^2/\text{s}$ and a normalized ADC of 0.97.

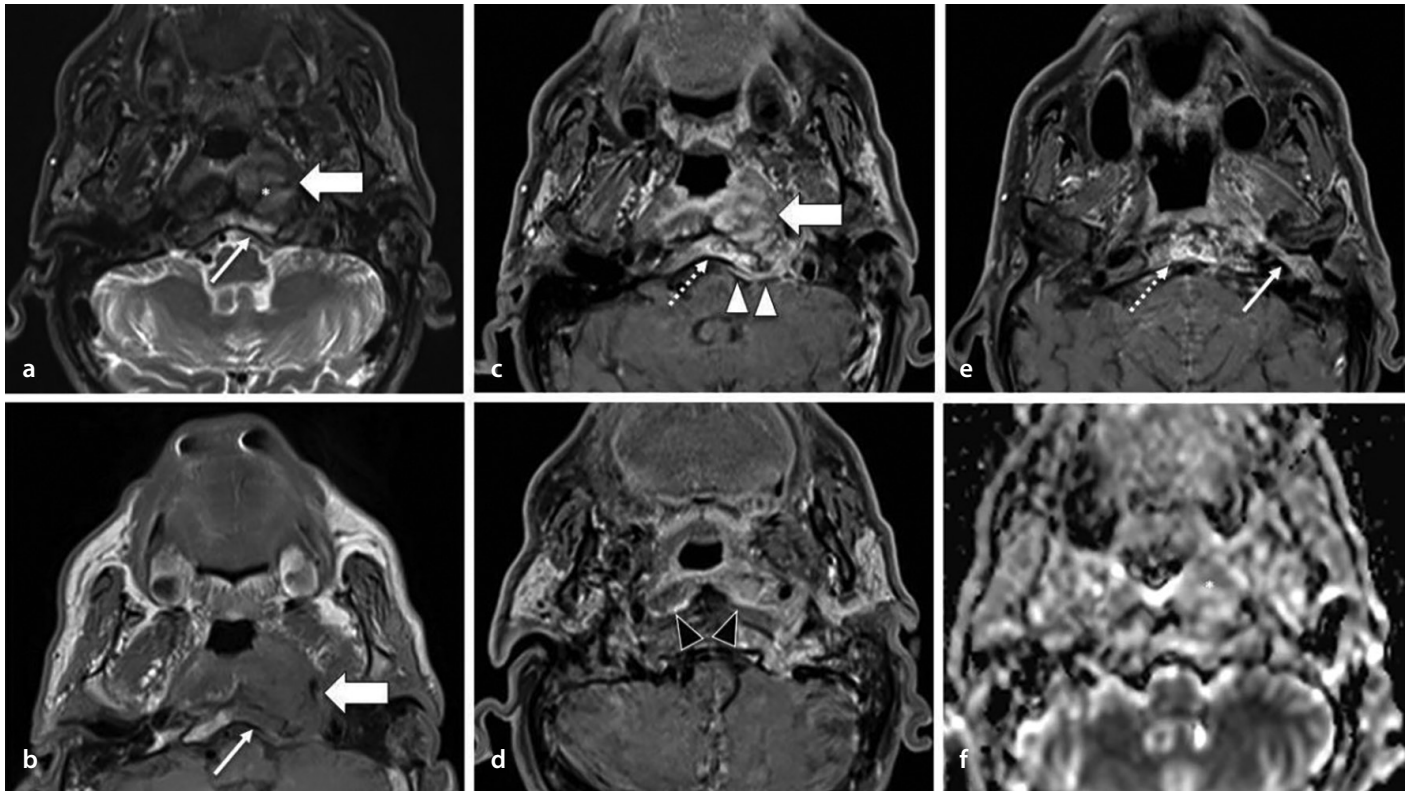


Figure 7. A patient with a confirmed diagnosis of skull base osteomyelitis (SBO) was misclassified as nasopharyngeal carcinoma by the semantic model but correctly identified as SBO by the radiomics model. (a) Axial fat-suppressed T2-weighted image demonstrates mild hyperintensity of the lesion (thick arrow) and bone marrow edema (thin arrow) in the left occipital condyle. Mild architectural distortion and asymmetric expansion of the left prevertebral muscles are also visible (asterisk). (b) Axial T1-weighted image shows deep hypointense bone marrow signal in the occipital condyle (thin arrow). (c, e) Axial post-contrast T1-weighted images show a markedly enhancing lesion involving the left prevertebral muscles, occipital condyle, and clivus (thick arrows), extending along the left Eustachian tube (e), (thin arrow). Note the linear enhancement along the retropharyngeal fascia (d), (black arrowheads), and compressed left Rosenmüller fossa without mucosal irregularities (c). Lateral structures, including the temporomandibular joint and parotid space, are preserved. Lesion enhancement is comparable with adjacent mucosa, with multifocal regions of higher enhancement within the bone (a, e); (dashed arrows). Dural thickening and enhancement are also present (d), (white arrowheads). (f) Apparent diffusion coefficient (ADC) map demonstrates facilitated diffusion (asterisk) relative to brain parenchyma (ADC: $1.24 \times 10^{-3} \text{ mm}^2/\text{s}$; normalized ADC: 1.38).

the literature.^{10,22} This discrepancy may be explained by differences in inclusion criteria. Unlike previous studies, our cohort excluded SBO cases confined to the temporal bone, instead focusing on advanced-stage patients with prominent soft tissue involvement, which can mimic NPca.

Although prior studies by Ozgen et al.⁸ and Baba et al.⁹ reported high efficacy for ADC parameters, with 96% accuracy and an AUC of 0.96, respectively, our ADC cut-off and model performance were notably lower. This discrepancy likely arises from two factors: lower mean ADC values observed in our advanced SBO cases²⁵ and variations in the time interval between symptom onset and imaging. Unlike previous studies, we applied established cohorts to an independent, unseen patient population and observed a significant decline in the diagnostic performance of ADC values. It should be acknowledged that significant differences in ADC distributions between the training and test cohorts likely contributed to the reduced

performance of the ADC-based model on external validation, rendering the comparison with radiomics not entirely equivalent. Notably, despite the application of normalization procedures in line with commonly adopted practices in the literature, ADC metrics still exhibited limited generalizability. This observation underscores the particular sensitivity of ADC values to inter-center variability and domain shifts, which may constrain their utility as stand-alone biomarkers in multicenter studies. This may be attributed to various factors, such as the selected b-values, patient-related factors, and differences between scanners. This highlights the challenge of applying a single quantitative ADC threshold across different institutions without standardized protocols.^{26,27}

Radiomics has been extensively studied in NPca for applications such as differential diagnosis, risk stratification, and the prediction of patient prognosis and treatment response.²⁸ However, to our knowledge, no radiomics studies have specifically addressed SBO.

Our SHAP analysis revealed that the most influential radiomic predictors were `t2_wavelet_LLL_firstorder_Minimum` and `t1_original_firstorder_Skewness`. The minimum intensity feature extracted from wavelet-filtered T2-weighted images was inversely associated with SBO probability, suggesting that lower signal minima were more indicative of NPca. This finding is biologically plausible when considered alongside qualitative assessment, as SBO often demonstrates a uniformly elevated T2 signal, similar to the mucosal tissue, due to edema and inflammatory exudation in the early disease phase. In contrast, NPca typically contains lower intensity foci related to dense cellularity, consistent with the observed association of lower minimum intensity values with malignancy. Positive skewness indicates a right-tailed distribution, where most voxels exhibit relatively low-to-intermediate signal intensities. In our analysis, increased skewness was positively associated with SBO, which is biologically plausible given that inflammatory lesions are typically fluid-rich and therefore

demonstrate lower baseline T1 intensities than the denser, more solid cellular architecture of NPCa.

Our semantic model, built with only five features, demonstrated strong performance. A semantic model developed on a larger dataset incorporating a broader range of features may yield even stronger results. Notably, three of the five selected radiomic predictors were first-order features, and SHAP analysis identified two of these first-order metrics as the most influential contributors to model predictions. This finding is biologically plausible, as these quantitative features are the mathematical equivalent of qualitative assessments—such as T2 signal intensity—that radiologists routinely use. Crucially, our radiomic approach achieved performance comparable with the human-led semantic evaluation but with the key advantages of being fully objective and reproducible. This suggests that with larger datasets, radiomics holds the potential to surpass purely visual assessment by quantifying subtle signal variations that are imperceptible to the human eye.

Future research should validate the robustness and generalizability of this radiomics approach in larger, prospective, multicenter cohorts. Additionally, comparative studies assessing the diagnostic performance of the model against that of expert neuroradiologists are necessary to formally establish its added clinical value. Furthermore, integration of relevant clinical risk factors, such as diabetes, inflammatory markers, and immune status, into the predictive model could enhance its diagnostic accuracy and clinical applicability. Due to the retrospective data collection, clinical history or acute phase reactants, such as erythrocyte sedimentation rate and C-reactive protein, were unavailable for a significant portion of the cohort, which prevented their inclusion in our models.

There were several limitations to our study. First, the retrospective design may have introduced selection bias. Second, the small sample size, particularly for certain subgroups, may have limited the performance of the semantic model and necessitates validation in larger cohorts. Moreover, future studies could compare these models with qualitative assessments by expert radiologists. Third, our narrow focus on skull base pathologies may limit the broader clinical applicability of the model. Finally, the significant differences in ADC parameters between the training and test sets may have contributed to the suboptimal performance of the ADC model, likely owing to the small sample

size. Although data processing may address these discrepancies, it was not applied to avoid introducing further uncertainties.

In conclusion, the radiomics model demonstrated high and robust performance in differentiating SBO from advanced NPCa. Although the semantic model based on expert assessment achieved slightly higher performance, radiomics offers a complementary, quantitative, and objective approach. In contrast, the ADC-based strategy proved highly sensitive to inter-center variability, suggesting that site-specific threshold determination may be more appropriate.

Acknowledgements

The authors gratefully acknowledge Ahmet Husrev Kösebalaban for his invaluable assistance with technical support, image pre-processing, and radiomic feature extraction.

Footnotes

Conflict of interest disclosure

The authors declared that they have no conflict of interest.

Supplementary Text: <https://d2v96fxpocvxx.cloudfront.net/beb8919b-f013-4ea1-b1c8-40332e840fe1/content-images/e7972d1b-8d19-4528-99fe-479e8b111ffd.pdf>

Supplementary Figures 1-5: <https://d2v96fxpocvxx.cloudfront.net/beb8919b-f013-4ea1-b1c8-40332e840fe1/content-images/f48b8f9f-17bb-440b-8737-fd17504852a7.pdf>

Supplementary Tables 1 and 2: <https://d2v96fxpocvxx.cloudfront.net/beb8919b-f013-4ea1-b1c8-40332e840fe1/content-images/92c210bf-5250-4c24-aa00-67b1b5e5e69d.pdf>

References

1. Chapman PR, Choudhary G, Singhal A. Skull base osteomyelitis: a comprehensive imaging review. *AJNR Am J Neuroradiol*. 2021;42(3):404-413. [\[Crossref\]](#)
2. Mahdyou P, Pulcini C, Gahide I, et al. Necrotizing otitis externa: a systematic review. *Otol Neurotol*. 2013;34(4):620-629. [\[Crossref\]](#)
3. Álvarez Jáñez F, Barriga LQ, Iñigo TR, Roldán Lora F. Diagnosis of skull base osteomyelitis. *Radiographics*. 2021;41(1):156-174. [\[Crossref\]](#)
4. Takata J, Hopkins M, Alexander V, et al. Systematic review of the diagnosis and management of necrotising otitis externa: highlighting the need for high-quality research. *Clin Otolaryngol*. 2023;48(3):381-394. [\[Crossref\]](#)

5. Başaran S, Evlice O, Benli A, et al. Skull base osteomyelitis and long term outcome. *Klimik Derg*. 2021;34(2):129-137. [\[Crossref\]](#)
6. Chawdhary G, Pankhania M, Douglas S, Bottrill I. Current management of necrotising otitis externa in the UK: survey of 221 UK otolaryngologists. *Acta Otolaryngol*. 2017;137(8):818-822. [\[Crossref\]](#)
7. van Kroonenburgh AMJL, van der Meer WL, Bothof RJP, van Tilburg M, van Tongeren J, Postma AA. Advanced imaging techniques in skull base osteomyelitis due to malignant otitis externa. *Curr Radiol Rep*. 2018;6(1):3. [\[Crossref\]](#)
8. Ozgen B, Oguz KK, Cila A. Diffusion MR imaging features of skull base osteomyelitis compared with skull base malignancy. *AJNR Am J Neuroradiol*. 2011;32(1):179-184. [\[Crossref\]](#)
9. Baba A, Kurokawa R, Kurokawa M, Ota Y, Srinivasan A. Dynamic contrast-enhanced MRI parameters and normalized ADC values could aid differentiation of skull base osteomyelitis from nasopharyngeal cancer. *AJNR Am J Neuroradiol*. 2023;44(1):74-78. [\[Crossref\]](#)
10. Goh JPN, Karandikar A, Loke SC, Tan TY. Skull base osteomyelitis secondary to malignant otitis externa mimicking advanced nasopharyngeal cancer: MR imaging features at initial presentation. *Am J Otolaryngol*. 2017;38(4):466-471. [\[Crossref\]](#)
11. See A, Tan TY, Gan EC. Atypical culture-negative skull base osteomyelitis masquerading as advanced nasopharyngeal carcinoma. *Am J Otolaryngol*. 2016;37(3):236-239. [\[Crossref\]](#)
12. Lee B, Bae YJ, Choi BS, et al. Radiologic differentiation between granulomatosis with polyangiitis and its mimics involving the skull base in humans using high-resolution magnetic resonance imaging. *Diagnostics (Basel)*. 2021;11(11):2162. [\[Crossref\]](#)
13. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—“how-to” guide and critical reflection. *Insights Imaging*. 2020;11(1):91. [\[Crossref\]](#)
14. Shur JD, Doran SJ, Kumar S, et al. Radiomics in oncology: a practical guide. *Radiographics*. 2021;41(6):1717-1732. [\[Crossref\]](#)
15. Song J, Yin Y, Wang H, Chang Z, Liu Z, Cui L. A review of original articles published in the emerging field of radiomics. *Eur J Radiol*. 2020;127:108991. [\[Crossref\]](#)
16. Kocak B, Baessler B, Bakas S, et al. CheckList for Evaluation of Radiomics Research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging*. 2023;14(1):75. [\[Crossref\]](#)
17. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging*. 2012;30(9):1323-1341. [\[Crossref\]](#)
18. Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on

- texture classification. *Magn Reson Imaging*. 2004;22(1):81-91. [\[Crossref\]](#)
19. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104-e107. [\[Crossref\]](#)
 20. Duron L, Balvay D, Vande Perre S, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. *PLoS One*. 2019;14(3):e0213459. [\[Crossref\]](#)
 21. Koçak B, Yüzkan S, Mutlu S, et al. Influence of image preprocessing on the segmentation-based reproducibility of radiomic features: *in vivo* experiments on discretization and resampling parameters. *Diagn Interv Radiol*. 2024;30(3):152-162. [\[Crossref\]](#)
 22. Brenner A, Cavel O, Shendler G, et al. CT findings in temporal bone sites in skull base osteomyelitis from malignant otitis externa. *Eur Arch Otorhinolaryngol*. 2023;280(6):2687-2694. [\[Crossref\]](#)
 23. Urbančič J, Vozel D, Battelino S, et al. Atypical skull-base osteomyelitis: comprehensive review and multidisciplinary management viewpoints. *Trop Med Infect Dis*. 2023;8(5):254. [\[Crossref\]](#)
 24. Adams A, Offiah C. Central skull base osteomyelitis as a complication of necrotizing otitis externa: imaging findings, complications, and challenges of diagnosis. *Clin Radiol*. 2012;67(10):e7-e16. [\[Crossref\]](#)
 25. Cherko M, Nash R, Singh A, Lingam RK. Diffusion-weighted magnetic resonance imaging as a novel imaging modality in assessing treatment response in necrotizing otitis externa. *Otol Neurotol*. 2016 Jul;37(6):704-707. [\[Crossref\]](#)
 26. Ogura A, Hatano I, Osakabe K, Yamaguchi N, Koyama D, Watanabe H. Importance of fractional b value for calculating apparent diffusion coefficient in DWI. *AJR Am J Roentgenol*. 2016;207(6):1239-1243. [\[Crossref\]](#)
 27. Kim SY, Lee SS, Park B, et al. Reproducibility of measurement of apparent diffusion coefficients of malignant hepatic tumors: effect of DWI techniques and calculation methods. *J Magn Reson Imaging*. 2012;36(5):1131-1138. [\[Crossref\]](#)
 28. Duan W, Xiong B, Tian T, et al. Radiomics in nasopharyngeal carcinoma. *Clin Med Insights Oncol*. 2022;16:11795549221079186. [\[Crossref\]](#)



Large-scale evaluation of multimodal large language models for pneumothorax detection

Hamza Eren Güzel¹

Cemre Özenbaş²

Ali Murat Koç³

¹University of Health Sciences Türkiye, İzmir City Hospital, Clinic of Radiology, İzmir, Türkiye

²Tınaztepe University Private Buca Hospital, Department of Radiology, İzmir, Türkiye

³İzmir Katip Çelebi University Atatürk Training and Research Hospital, Department of Radiology, İzmir, Türkiye

PURPOSE

Pneumothorax requires rapid recognition and accurate interpretation of chest X-rays (CXRs), particularly in acute settings where delays can have serious consequences. With the emergence of advanced image interpretation models capable of visual analysis, their diagnostic reliability in radiology practice remains to be determined. This study aimed to assess the diagnostic performance of three state-of-the-art systems in detecting pneumothorax using a large, well-annotated dataset.

METHODS

A total of 10,675 CXRs from the publicly available SIIM-ACR Pneumothorax Segmentation dataset were analyzed. Three multimodal models (GPT-4o, Gemini 2 Pro, and Claude 4 Sonnet) were evaluated using a uniform, image-based approach. Each model's binary outputs (presence: 1, absence: 0) were compared with reference results to determine accuracy, sensitivity, specificity, precision, and F1 scores. Additional subgroup analyses were conducted across pneumothorax size categories: small, medium, and large. Pairwise statistical comparisons were performed using McNemar's test. Sensitivity, specificity, and overall accuracy are reported with corresponding 95% confidence intervals.

RESULTS

The prevalence of pneumothorax in the dataset was 22.3% (n = 2,379). All models demonstrated high specificity (above 0.90) but consistently low sensitivity (0.16–0.36). The best overall performance was observed with Gemini 2, which achieved an accuracy of 0.79 and specificity of 0.95, whereas Claude 4 showed greater sensitivity (0.20–0.34) across lesion-size categories. Diagnostic performance improved with increasing pneumothorax size, but smaller lesions remained difficult to identify. Pairwise comparisons confirmed statistically significant differences among all evaluated systems ($P < 0.050$).

CONCLUSION

In this large-scale evaluation, the tested models exhibited strong reliability in identifying normal examinations but limited ability to detect subtle or small pneumothoraxes. Despite high specificity, low sensitivity limits the use of current multimodal large language models as rule-out tools for pneumothorax. With continued refinement, these models may eventually support radiologists by improving workflow efficiency and diagnostic confidence.

CLINICAL SIGNIFICANCE

Automated systems capable of high specificity but low sensitivity should not be relied upon to exclude pneumothorax. However, they may serve as valuable assistants for confirming positive findings and prioritizing urgent cases in busy clinical workflows.

KEYWORDS

Pneumothorax, multimodal large language models, chest X-ray, artificial intelligence, diagnostic accuracy

Handling editor: Tuğba Akıncı D'Antonoli

Corresponding author: Hamza Eren Güzel

E-mail: hamzaerenguzel@gmail.com

Received 07 November 2025; revision requested 01 January 2026; accepted 31 January 2026.



Epub: 12.02.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263696

Pneumothorax is a potentially life-threatening condition characterized by the presence of air within the pleural space, leading to partial or complete lung collapse. The diagnostic approach to pneumothorax should begin with a standard posteroanterior chest X-ray (CXR), which remains the first-line imaging modality for confirming the presence of intrapleural air and assessing its extent. Recent clinical guidelines highlight that timely and accurate interpretation of CXRs is critical, as management decisions often depend on prompt recognition and precise evaluation of pneumothorax size and severity.¹

Over the past decade, deep learning (DL) has shown strong performance in the automated detection and segmentation of pneumothorax on CXR. Several systematic reviews and large studies have reported high diagnostic accuracy across different datasets, underscoring the potential role of DL as a triage or second-reader tool.²⁻⁵ These advances were made possible not only by improvements in algorithms but also by the availability of large, open datasets that helped standardize benchmarking in pneumothorax detection.⁴

Multimodal large language models (MLLMs) are general-purpose artificial intelligence (AI) systems capable of processing both visual and textual inputs. Although prior studies have reported promising results when such models are provided with combined image and clinical context, their performance in image-only diagnostic tasks remains inconsistent.⁶⁻¹⁰ However, most current MLLMs are not specifically trained for medical imaging tasks, which may limit their spatial localization and diagnostic precision, particularly when relying solely on image-based inputs. In this context, evaluating MLLMs not only on small patient cohorts but also on large, well-curated datasets is crucial to establishing their reliability and diagnostic validity.

Main points

- Multimodal large language models (GPT-4o, Claude 4, Gemini 2) were evaluated on more than 10,000 chest X-rays for pneumothorax detection.
- All models achieved high specificity but low sensitivity, in all sizes.
- They may serve as potential rule-in aids for radiologists, but their low sensitivity makes rule-out use clinically unsafe.

In this study, we aimed to evaluate the performance of MLLMs in detecting pneumothorax on CXRs using a large-scale dataset. In addition, we assessed the ability of different MLLMs to identify pneumothoraces of varying sizes. We hypothesized that the diagnostic performance of general-purpose MLLMs for pneumothorax detection on CXRs would differ between models and across pneumothorax size categories. Given the critical importance of rapid diagnosis in emergency settings, our study also sought to explore the practical applicability and limitations of these rapidly evolving models in the detection of pneumothorax.

Methods

Dataset

We used the publicly available SIIM-ACR Pneumothorax Segmentation Dataset, originally developed for the 2019 Kaggle challenge.¹¹ The dataset provides CXRs in DICOM format, annotated with pixel-level segmentation masks for pneumothorax. No exclusion criteria were applied, as the dataset is a publicly available, pre-curated benchmark dataset with standardized annotation and quality control procedures. All DICOM images were successfully processed, and no preprocessing failures or unreadable images were encountered. After preprocessing, a total of 10,675 radiographs were included in the analysis. Each radiograph was classified as pneumothorax-positive or -negative according to the presence or absence of a segmentation mask, derived from the SIIM-ACR radiologist annotations, which served as the reference standard. For positive cases, the lesion area was calculated from the mask and used for size categorization. The overall study workflow, including data preprocessing, model inference via application programming interface (API), and comparative performance evaluation, is summarized in Figure 1.

Image preprocessing

All DICOM images were converted into 8-bit grayscale PNG format using a standardized Python-based workflow implemented with the Pydicom library (Pydicom, Boston, MA, USA) and the Pillow imaging library (Python Imaging Library, San Francisco, CA, USA). The pipeline ensured consistent windowing, normalization, and resizing across all studies to maintain compatibility and reproducibility.

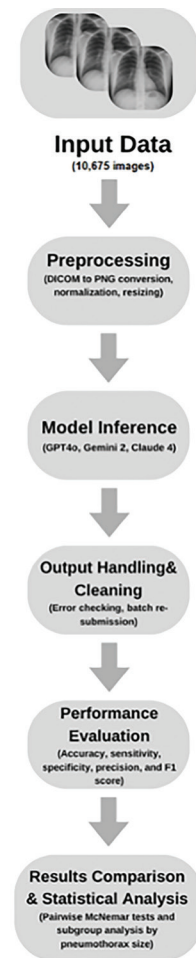


Figure 1. Overview of the study workflow, showing sequential steps from data input and preprocessing to model inference and results comparison.

Each DICOM file was decoded using Pydicom, and the corresponding modality and VOI LUT transformations were applied to preserve true intensity values.

Pixel values were normalized to the 0–255 range through percentile-based clipping (0.5th–99.5th) to reduce the influence of outliers and scanner-specific variations. All images were resized to 1024 × 1024 pixels, preserving the original aspect ratio via zero padding to avoid geometric distortion.

Each image was saved as an 8-bit grayscale PNG file suitable for MLLM inference. A conversion manifest linking each PNG to its source DICOM (SOPInstanceUID) was automatically generated for traceability (Figure 2).

Models and API access

Three MLLMs were evaluated: GPT-4o (OpenAI), Gemini 2 Pro (Google DeepMind), and Claude 4 Sonnet (Anthropic).

Each model was accessed through its official API during October and November 2025. Exact model version strings and API access dates are provided in Supplementary Table 1. All images were submitted as standardized 8-bit PNG files prepared during preprocessing. Models were used as provided by the vendors, without additional fine-tuning or training. To ensure comparability, inference was run with deterministic decoding (temperature: 0) and a capped output budget (maximum 512 tokens). For each case, the model response, image identifier, timestamp, and model name were recorded to comma-separated files and subsequently merged with the master analysis table by filename/Imageld for performance calculations. All inferences were executed via paid API accounts procured by the authors; the vendors had no role in study design, analysis, or manuscript preparation.

Prompting strategy and output handling

A single, model-agnostic prompt was used for all systems to minimize bias. The exact instruction was:

“You are a radiology assistant. Carefully analyze this CXR for pneumothorax.

Return exactly ONE character: 1 if pneumothorax is present or strongly suspected, and 0 if absent.

Answer only with 1 or 0.”

The prompt intentionally avoided examples or narrative phrasing to reduce stylistic drift between models and to standardize outputs. Returned texts were parsed with a strict rule set that extracted the first valid numeric character; if the output was non-conforming, defined as any response that did not contain a single numeric character (“1” or “0”), such as textual answers (“Yes,” “No,” “Present,” “Ab-

sent”), symbols, or blank strings, the request was repeated once to allow for correction. Persistently unparseable cases (i.e., outputs that remained non-numeric after two iterations) were flagged and subsequently re-submitted to the API as a separate batch. The new responses were manually verified and recorded by the study team to ensure completeness and consistency.

The full Python scripts used for preprocessing, API-based inference, and evaluation are publicly available.¹²

Reference standard and lesion categorization

The ground truth was defined by the segmentation masks provided in the SIIM-ACR dataset. Pneumothorax lesion area was computed in pixels. For pneumothorax-positive cases, lesion size was quantified as the total number of pixels within the segmentation mask after standardization to a 1024 × 1024 grid. Histogram analysis showed a right-skewed distribution: most cases were < 10,000 pixels, although a minority extended above 100,000 pixels. Percentile analysis revealed that the 25th, 50th (median), 75th, 90th, and 95th percentiles corresponded to 4,043, 8,666, 18,732, 34,979, and 48,550 pixels, respectively. Based on this distribution, and to ensure clinically meaningful and statistically balanced subgroups, we defined three categories: small (< 10,000 pixels), medium (10,000–35,000 pixels), and large (> 35,000 pixels). These thresholds were established a priori (before model evaluation) to avoid bias, and were chosen to reflect both the quartile distribution of lesion sizes and the expected clinical conspicuity of different pneumothorax volumes. In cases with multiple separate regions, the total combined mask area was used for categorization.

Statistical analysis

Diagnostic performance was evaluated at the case level using standard classification metrics, including sensitivity, specificity, accuracy, precision, and F1 score. Reporting followed recommended practices for diagnostic accuracy studies.¹³ Metrics were calculated for the overall dataset and separately for each lesion size category (small, medium, large). True-positive, false-positive, true-negative, and false-negative counts were derived from model predictions and reference labels. Pairwise comparisons between models were conducted using McNemar’s test to assess significant differences in classification performance. A *P* value < 0.050 was considered statistically significant. All analyses were conducted in Python (version 3.10) using pandas (version 2.2) for data management and scikit-learn (version 1.5) for metric computation.

Ethical considerations

This study was approved by the İzmir City Hospital Ethical Committee (approval number: 2025/550, date: 08.10.2025). As the SIIM-ACR dataset is publicly available and fully de-identified, the requirement for individual informed consent was waived. The study complies with the principles of the Declaration of Helsinki.

Results

Dataset characteristics

A total of 10,675 CXRs were included in the analysis. Of these, 2,379 (22.3%) were labeled as pneumothorax-positive and 8,296 (77.7%) as negative. Among the positive cases, 1,302 (54.7%) were categorized as small, 839 (35.3%) as medium, and 238 (10.0%) as large pneumothoraxes.

Model performance

All three MLLMs were evaluated against the reference standard. Their performance metrics, including accuracy, sensitivity, specificity, precision, and F1 score, are summarized in Table 1.

Although individual results varied across models, all systems achieved high specificity with comparatively lower sensitivity values.

Pairwise comparisons using McNemar’s test showed statistically significant differences among the models (*P* < 0.001). Gemini 2 demonstrated higher overall accuracy and specificity than both Claude 4 and GPT-4o (*P* < 0.001). Claude 4 exhibited higher sensitivity than Gemini 2 (*P* = 0.033). GPT-4o showed

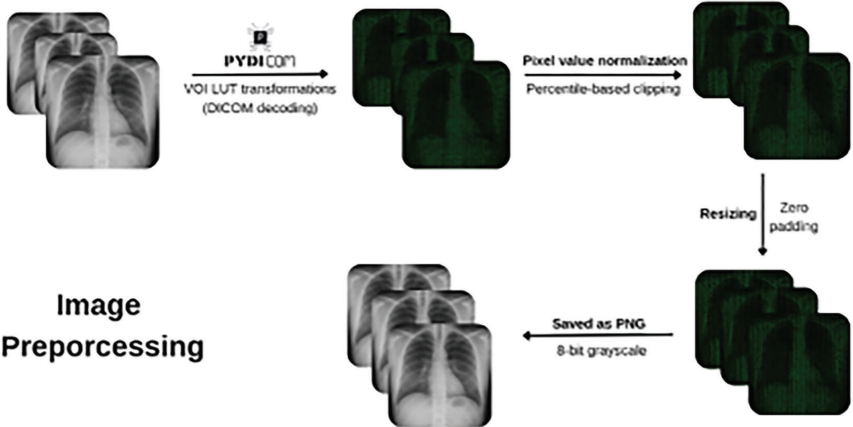


Figure 2. DICOM-to-PNG preprocessing workflow for standardized image conversion.

significantly lower performance compared with both models across all evaluated metrics ($P < 0.001$). Detailed statistical comparisons are presented in Table 2.

Performance by pneumothorax size

When performance was analyzed by pneumothorax size, metric values differed across subgroups (Table 3).

For small pneumothoraxes ($n = 1,302$), accuracies ranged from 0.90 (GPT-4o) to 0.93

(Gemini 2), sensitivities from 0.17 to 0.20, and F1 scores from 0.28 to 0.34.

For medium pneumothoraxes ($n = 839$), accuracies were between 0.92 and 0.94, sensitivities between 0.16 and 0.25, and F1 scores between 0.28 and 0.40.

For large pneumothoraxes ($n = 238$), accuracies ranged from 0.98 to 0.99, sensitivities from 0.19 to 0.36, and F1 scores from 0.32 to 0.53 across the three models.

Visual representation

The visual overview of model performance highlights how sensitivity increases with pneumothorax size and how true and false classifications are distributed across the dataset, offering an intuitive understanding of the results (Figures 3 and 4).

Table 1. Overall performance of multimodal large language models in pneumothorax detection					
Model	Accuracy	Sensitivity	Specificity	Precision	F1 score
GPT-4o	0.75	0.17	0.92	0.36	0.23
Claude 4	0.77	0.23	0.92	0.46	0.31
Gemini 2	0.79	0.22	0.95	0.56	0.31

Table 2. Pairwise McNemar test for differences in model performance				
Comparison	Accuracy (P)	Sensitivity (P)	Specificity (P)	Direction of difference
Gemini 2 vs. Claude 4	< 0.001	0.033	< 0.001	Gemini 2: higher accuracy, specificity Claude 4: higher sensitivity
Gemini 2 vs. GPT-4o	< 0.001	< 0.001	< 0.001	Gemini 2: higher in all metrics
Claude 4 vs. GPT-4o	< 0.001	< 0.001	0.21	Claude 4: higher accuracy, sensitivity No difference in specificity

Table 3. Performance stratified by pneumothorax size				
Model	Size	Accuracy	Sensitivity	F1 score
GPT-4o	Small	0.90	0.17	0.28
	Medium	0.92	0.16	0.28
	Large	0.98	0.19	0.32
Claude 4	Small	0.91	0.20	0.34
	Medium	0.93	0.25	0.40
	Large	0.98	0.34	0.50
Gemini 2	Small	0.93	0.17	0.29
	Medium	0.94	0.24	0.39
	Large	0.99	0.36	0.53

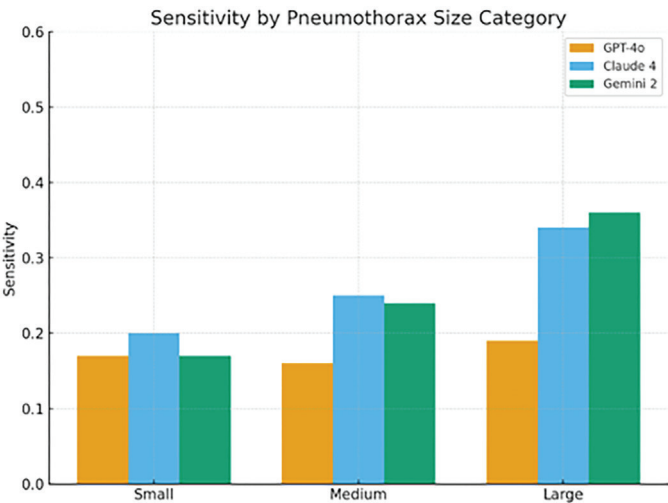


Figure 3. Sensitivity across pneumothorax size categories. All models improved with increasing lesion size, with Gemini 2 consistently achieving the highest sensitivity.

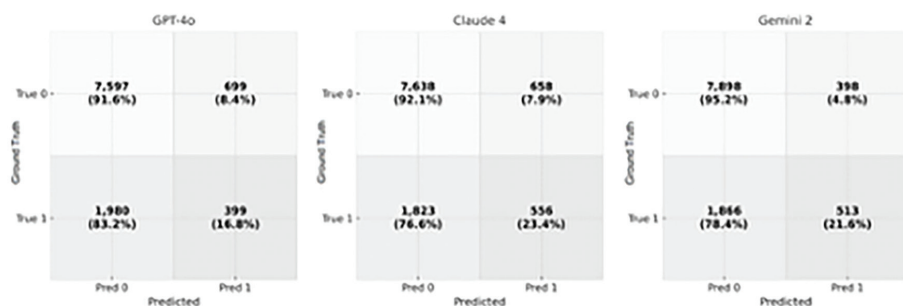


Figure 4. Combined confusion matrices for GPT-4o, Claude 4, and Gemini 2 on the full cohort (n = 10,675; positives: 2,379; negatives: 8,296). Cells display case counts and row-wise percentages within the true class (True 0/True 1); higher diagonal values indicate correct classifications (TN, TP), with axes labeled “True” (rows) and “Predicted” (columns).

TN, true negative; TP, true positive.

Discussion

This large-scale evaluation demonstrated that MLLMs achieve high specificity but consistently low sensitivity for pneumothorax detection on CXRs. This study evaluated three state-of-the-art MLLMs (GPT-4o, Claude 4, and Gemini 2) on the task of pneumothorax detection from CXRs using the SIIM-ACR dataset. The results revealed a consistent pattern: all three models achieved high specificity (> 0.90), whereas sensitivity remained low across all pneumothorax categories, ranging from 0.16 to 0.36. These findings highlight a critical limitation of current MLLMs, raising substantial concerns about their potential use in clinical scenarios such as pneumothorax, where rapid diagnosis and timely intervention are essential.

Numerous AI studies based on imaging data have been conducted to address emergency conditions such as pneumothorax, where clinicians and radiologists must make rapid diagnostic decisions. DL-based approaches, particularly convolutional neural networks, have shown remarkable potential for the automatic detection and localization of pneumothorax on CXRs. For example, Cho et al.¹⁴ investigated a DL-based method for detecting and localizing pneumothorax on CXRs, aiming to enhance diagnostic accuracy and support clinical workflows in emergency settings. They demonstrated that their approach can accurately identify pneumothorax on CXRs, reduce diagnostic delays, and support more effective clinical decision-making and patient care. Similarly, Hillis et al.¹⁵ reported that an AI model could reliably identify both pneumothorax and tension pneumothorax on CXRs. Thian et al.³ conducted a multicenter external validation study evaluating DL systems for pneumothorax detection on CXRs. Their model was tested on datasets from multiple institutions

and demonstrated robust generalizability, achieving area under the curve values ranging from 0.91 to 0.98 across external cohorts.

By contrast, there are only a limited number of studies investigating the use of LLMs for pneumothorax detection. Among them, Ostrovsky¹⁶ evaluated the performance of ChatGPT in interpreting CXRs across various thoracic pathologies. In the pneumothorax subgroup consisting of 200 patients, the model achieved a sensitivity of 0.77 and a specificity of 0.89, demonstrating moderate diagnostic capability in this emergency setting. In our study, MLLMs showed noticeably lower diagnostic performance when compared with AI tools developed specifically for image interpretation. Using a large publicly available dataset of more than 10,000 images, our evaluation indicates that current MLLMs still lack the reliability required for routine use in radiology practice.

There are inherent risks associated with integrating AI applications and MLLMs into clinical practice without fully testing their reliability and validity. In this context, the low sensitivity observed in our study could lead to serious problems in clinical practice, particularly in emergency settings. Clinicians relying entirely on these systems for decision-making may delay necessary interventions, potentially worsening the patient's condition. Conversely, the specificity rates exceeding 90% observed in our study suggest that, in their current state, MLLMs may be more suitable for rule-in purposes rather than definitive diagnosis.

In recent years, numerous studies have evaluated the performance of MLLMs in interpreting medical images in the field of radiology. In the study conducted by Sonoda et al.,⁸ the diagnostic performance of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro was systematically compared using RSNA Diag-

nosis Please cases. The results showed that Claude 3 Opus achieved the highest primary diagnostic accuracy at 54%, followed by GPT-4o at 41%, and Gemini 1.5 Pro at 33.9%. Moreover, the proportion of cases where the correct diagnosis was included within the top three differential diagnoses was 62%, 49.4%, and 41%, respectively. Similarly, a study conducted by Suh et al.¹⁷ compared GPT-4V and Gemini Pro Vision using only image inputs against board-certified radiologists. The results demonstrated that GPT-4V reached an accuracy of 49% and Gemini Pro Vision achieved 39%, both falling below the 61% accuracy observed among radiologists. Hirano et al.¹⁸ evaluated multiple MLLMs on the Japanese Diagnostic Radiology Board Examination using image-based questions and found that o3 achieved 72% accuracy, o4-mini and Gemini 2.5 Pro each reached 70%, and Claude 3.7 Sonnet received lower legitimacy scores. Similarly, Nakaura et al.¹⁹ compared several models on board examination items and reported that GPT-4o achieved the highest accuracy (45%), with Claude 3 Opus performing best in text-only tasks (46%).

Despite the varying results across different imaging modalities, one thing is clear: MLLMs are improving rapidly. Hou et al.²⁰ reported that next-generation models, such as o1 and GPT-4o, demonstrated substantial performance gains compared with their predecessors, with diagnostic accuracy approaching that of human experts in certain scenarios. Although their integration into routine radiology workflows may seem challenging at the moment, it is important to remember that MLLMs have received no training specifically focused on radiologic diagnosis. Given their rapid progress, it seems likely that they may soon perform at a level comparable to dedicated radiology AI systems.

This study has several limitations. First, the evaluated MLLMs were not specifically trained for medical imaging, reflecting a limitation inherent to model design. Second, although these models can process multimodal inputs, the evaluation in this study was limited to image-only prompts, excluding potentially informative clinical or textual context that might enhance diagnostic performance. The use of a simplified binary output format restricted the models' ability to convey diagnostic uncertainty, which may have affected their ability to detect borderline cases. Third, the SIIM-ACR Pneumothorax dataset, although large and well-curated, originates from a single competition source and may not reflect the full diversity of re-

al-world imaging conditions, acquisition parameters, or institutional variations, limiting generalizability. Fourth, the reproducibility of our findings is inherently limited by the use of API-based MLLMs. These systems are subject to ongoing backend updates and undocumented changes implemented by vendors, which may alter model behavior over time, even when identical prompts and inference parameters are used. Lastly, the findings were derived from a single dataset without external validation; future studies using independent or real-world hospital data are required to confirm robustness and clinical applicability.

In this large-scale evaluation using the SIIM-ACR Pneumothorax dataset, MLLMs demonstrated high specificity but low sensitivity in detecting pneumothorax on CXRs. From a clinical perspective, this finding indicates that current MLLMs may serve as rule-in aids for confirming pneumothorax but are unsuitable for rule-out use. Although models such as Gemini 2 and Claude 4 showed relatively better overall performance than GPT-4o, all systems missed the majority of small pneumothoraxes, limiting their clinical reliability in emergency or screening contexts. These findings indicate that current MLLMs, which are not specifically trained for radiologic interpretation, are not yet suitable for standalone diagnostic use. Nevertheless, future iterations or domain-adapted versions of MLLMs may eventually support radiologists by improving workflow efficiency and enhancing diagnostic confidence, particularly as model architectures and training strategies continue to evolve.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Table 1. <https://d2v96fx-pocvxx.cloudfront.net/90a4190a-90d9-41a4-a9c9-d78d3fa8efda/content-images/93147689-f3e8-4214-b025-72c1b0c25421.pdf>

References

1. Roberts ME, Rahman NM, Maskell NA, et al. British Thoracic Society Guideline for pleural disease. *Thorax*. 2023;78(Suppl 3):s1-s42. [\[Crossref\]](#)
2. Sugibayashi T, Walston SL, Matsumoto T, Mitsuyama Y, Miki Y, Ueda D. Deep learning for pneumothorax diagnosis: a systematic review and meta-analysis. *Eur Respir Rev*. 2023;32(168):220259. [\[Crossref\]](#)
3. Thian YL, Ng D, Hallinan JTPD, et al. Deep learning systems for pneumothorax detection on chest radiographs: a multicenter external validation study. *Radiol Artif Intell*. 2021;3(4):e200190. [\[Crossref\]](#)
4. Wang Y, Lu Z. Automatic segmentation of pneumothorax in chest radiographs based on dual-task interactive learning method. In: Proceedings of the 2023 5th International Conference on Image, Video and Signal Processing (IVSP '23); 2023 Mar 24-26; Hong Kong, China. New York: ACM; 2023. p. 51-58. [\[Crossref\]](#)
5. Güzel HE, Aşçı G, Demirbilek O, Özdemir TD, Ereklı PB. Diagnostic precision of a deep learning algorithm for the classification of non-contrast brain CT reports. *Front Radiol*. 2025;5:1509377. Erratum in: *Front Radiol*. 2025;5:1744006. [\[Crossref\]](#)
6. Zhou Y, Ong H, Kennedy P, et al. Evaluating GPT-V4 (GPT-4 with Vision) on detection of radiologic findings on chest radiographs. *Radiology*. 2024;311(2):e233270. Erratum in: *Radiology*. 2024;311(2):e249016. [\[Crossref\]](#)
7. Hayden N, Gilbert S, Poisson LM, Griffith B, Klochko C. Performance of GPT-4 with vision on text- and image-based ACR diagnostic radiology in-training examination questions. *Radiology*. 2024;312(3):e240153. [\[Crossref\]](#)
8. Sonoda Y, Kurokawa R, Nakamura Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. *Jpn J Radiol*. 2024;42(11):1231-1235. [\[Crossref\]](#)
9. Kurokawa R, Ohizumi Y, Kanzawa J, et al. Diagnostic performances of Claude 3 Opus and Claude 3.5 Sonnet from patient history and key images in Radiology's "Diagnosis Please" cases. *Jpn J Radiol*. 2024;42(12):1399-1402. [\[Crossref\]](#)
10. Büyüktoka RE, Salbas A. Multimodal large language models for pediatric bone-age assessment: a comparative accuracy analysis. *Acad Radiol*. 2025;32(11):6905-6912. [\[Crossref\]](#)
11. SIIM ACR Pneumothorax Segmentation Data [Internet]. Kaggle. [\[Crossref\]](#)
12. MLLM Pneumothorax Detection on SIIM-ACR Dataset [Internet]. GitHub. [\[Crossref\]](#)
13. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med*. 2025;31(10):3283-3289. [\[Crossref\]](#)
14. Cho Y, Kim JS, Lim TH, Lee I, Choi J. Detection of the location of pneumothorax in chest X-rays using small artificial neural networks and a simple training process. *Sci Rep*. 2021;11(1):13054. [\[Crossref\]](#)
15. Hillis JM, Bizzo BC, Mercaldo S, et al. Evaluation of an artificial intelligence model for detection of pneumothorax and tension pneumothorax in chest radiographs. *JAMA Netw Open*. 2022;5(12):e2247172. [\[Crossref\]](#)
16. Ostrovsky AM. Evaluating a large language model's accuracy in chest X-ray interpretation for acute thoracic conditions. *Am J Emerg Med*. 2025;93:99-102. [\[Crossref\]](#)
17. Suh PS, Shim WH, Suh CH, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs from diagnosis please cases. *Radiology*. 2024;312(1):e240273. [\[Crossref\]](#)
18. Hirano Y, Miki S, Yamagishi Y, et al. Assessing accuracy and legitimacy of multimodal large language models on Japan Diagnostic Radiology Board Examination. *Jpn J Radiol*. 2026;44(1):209-217. [\[Crossref\]](#)
19. Nakaura T, Yoshida N, Kobayashi N, et al. Performance of multimodal large language models in Japanese Diagnostic Radiology Board Examinations (2021-2023). *Acad Radiol*. 2025;32(5):2394-2401. [\[Crossref\]](#)
20. Hou B, Mukherjee P, Batheja V, Wang KC, Summers RM, Lu Z. One year on: assessing progress of multimodal large language model performance on RSNA 2024 Case of the Day Questions. *Radiology*. 2025;316(2):e250617. [\[Crossref\]](#)



©Copyright 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Prospective quantitative analysis of hyperparameter and input optimization in GPT-5: comparative contribution to radiologist performance in abdominal radiology

Eren Çamur¹

Turay Cesur²

Yasin Celal Güneş³

Muhammed Batuhan Gökhan⁴

Rıza Sarper Ökten⁵

¹Ankara 29 Mayıs State Hospital, Clinic of Radiology, Ankara, Türkiye

²Ankara Mamak State Hospital, Clinic of Radiology, Ankara, Türkiye

³Yüksek İhtisas Hospital, Clinic of Radiology, Kırıkkale, Türkiye

⁴Yozgat Yerköy State Hospital, Clinic of Radiology, Yozgat, Türkiye

⁵Ankara Bilkent City Hospital, Clinic of Radiology, Ankara, Türkiye

Handling editor: Tuğba Akıncı D'Antonoli

Corresponding author: Eren Çamur

E-mail: eren.camur@outlook.com

Received 21 November 2025; revision requested 21 December 2025; accepted 31 January 2026.



Epub: 16.02.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263762

PURPOSE

This study aims to evaluate the effect of input format and hyperparameter settings on GPT-5 and explore the contribution of GPT-5 assistance to radiologists' performance in abdominal cases.

METHODS

In this prospective study, 86 abdominal cases were evaluated, with GPT-5 evaluated in two deployment contexts: browser-based GPT-5 (default, non-configurable sampling settings) and GPT-5 accessed via the OpenAI application programming interface (API) with different temperature and top-p settings. First, the diagnostic and differential diagnosis performance of browser-based GPT-5 in these cases was assessed using two different input formats: "only visual" and "visual with imaging findings and clinical presentation." Subsequently, its performance was evaluated at varying temperature (0, 0.5, 1, 1.5) and top-p (0, 0.5, 1) values; the values at which the model performs best are considered "optimal settings." Finally, two junior radiologists evaluated the same cases with and without GPT-5 assistance with washout periods. Their performances were compared internally and with that of an abdominal radiologist.

RESULTS

With the "only visual" input format, browser-based GPT-5 achieved a diagnostic accuracy of 12%, increasing to 58% when imaging findings and clinical presentations were provided ($P < 0.001$). Hyperparameter optimization further improved GPT-5 performance, with diagnostic accuracy increasing to 73% at the optimal settings (temperature: 1.5, top-p: 1) and mean differential diagnosis scores improving from 3.44 to 3.84. The radiologists' diagnostic accuracy increased from 73% and 71% without assistance to 87% and 86% with browser-based GPT-5 assistance and further to 94% with GPT-5 with optimal settings assistance. Differential diagnosis performance similarly improved from median scores of 4 (range: 3–5) without assistance to 5 (range: 4–5) with GPT-5 (with optimal settings) assistance.

CONCLUSION

Using hyperparameter and input optimization settings with GPT-5 could improve its clinical utility.

CLINICAL SIGNIFICANCE

This study evaluates GPT-5 performance in a single-source, open-access abdominal case set. In this study, GPT-5 performance improved with structured text inputs and API-based hyperparameter optimization, and large language model (LLM) assistance was associated with improved diagnostic and differential diagnosis performance among junior radiologists. These findings suggest that documenting and standardizing hyperparameter settings (e.g., temperature and top-p) may be important for future LLM-based decision-support applications.

KEYWORDS

ChatGPT, abdomen, artificial intelligence, temperature, hyperparameter, optimization

The development of large language models (LLMs) and, more recently, large multimodal models (LMMs) and vision-language models (VLMs) marks a major advance in artificial intelligence (AI), with the potential to transform various fields.¹ LLMs are developed using extensive datasets and sophisticated algorithms, enabling them to generate human-like text accurately. This capability has garnered considerable interest from both academia and industry, particularly for complex decision-making in data-rich fields such as medicine.² Within healthcare, LLMs hold promise for enhancing clinical workflows, offering diagnostic insights, supporting documentation, and facilitating patient education.^{3,4} Unlike text-only LLMs, LMMs can jointly process multiple input

modalities such as text and images. A major subgroup of LMMs are VLMs, which integrate image understanding with natural language processing to enable multimodal reasoning.⁵

Radiology, a specialty that relies heavily on precision and detail to establish a correct diagnosis, stands to benefit immensely from integrating multimodal models (LMMs/VLMs). These models have already demonstrated notable potential across radiologic subspecialties, excelling in tasks such as generating multiple-choice questions and craftinging patient-friendly report impressions.^{6,7} Previous studies have underscored their high performance in analyzing publicly available cases across subspecialties, especially when detailed patient histories and imaging findings are provided.⁸⁻¹³

For clinical applications, optimizing the performance of these models requires adjustment of the hyperparameters, such as temperature and top-p, which influence response variability and creativity.^{14,15} Temperature controls the randomness and creativity of model responses—a higher temperature results in more diverse and creative outputs, whereas a lower value makes the output more focused and deterministic.^{16,17} Top-p limits the model by considering only the most probable tokens whose cumulative probability meets a specified threshold.¹⁵ Together, these parameters help balance creativity and precision in generating outputs.¹⁴⁻¹⁶ Unlike deterministic models (e.g., convolutional neural networks), LLMs are inherently variable, producing diverse outputs even for identical inputs. This variability can affect reliability, a critical factor in radiology.⁹ Adjusting these hyperparameters enables the model to balance creativity with consistency, fostering dependable LLM-based applications in imaging interpretation.¹⁸

In a recent study, Suh et al.¹⁹ investigated how temperature settings (0, 0.5, and 1) affect the diagnostic accuracy of multimodal LLMs, specifically ChatGPT-4V and Gemini Pro Vision, across radiological cases spanning multiple subspecialties. They reported modest improvements in diagnostic accuracy with increasing temperature for both models (GPT-4V: 41% to 49%; Gemini Pro Vision: 29% to 39%).¹⁹ This study underscores the influence of temperature settings on potential improvements in the diagnostic performance of models.

To the best of our knowledge, no prior study has evaluated the effect of top-p and temperature settings on GPT-5 performance in abdominal radiology. This study addresses

this gap and explores the impact of GPT-5 assistance on radiologists' performance.

Methods

Study design

This experimental prospective study used a cross-sectional design with three steps and was conducted between July 2024 and November 2025. The cases used in this study were obtained from the Eurorad database, established by the European Society of Radiology in April 2000.²⁰ Cases with multisystemic etiologies were included if abdominal imaging constituted the primary diagnostic focus, reflecting the spectrum of conditions routinely encountered in abdominal radiology practice. Each included case had "imaging findings," "clinical presentation," "correct diagnosis," and "differential diagnosis" sections.

The performance of GPT-5 was evaluated in two deployment contexts: browser-based GPT-5 (<https://chatgpt.com>), which uses default and undisclosed hyperparameter settings, and GPT-5 accessed via the OpenAI application programming interface (API; <https://platform.openai.com>), which allows explicit control of the temperature and top-p settings. The performances of browser-based GPT-5 and GPT-5 at different temperature and top-p settings using the OpenAI API were compared using the same cases. The temperature and top-p settings at which GPT-5 performs best are called the "optimal settings." The responses of browser-based GPT-5 and GPT-5 with the optimal settings for each case were noted. Finally, to demonstrate the contribution of GPT-5 assistance to radiologists' performance, Radiologist 1 (R1) and Radiologist 2 (R2), both European Diploma in Radiology certified with 7 years of experience in general radiology, evaluated the same cases with and without GPT-5 assistance, and their performances were compared with that of an abdominal radiologist (AR) with 25 years of experience in abdominal radiology. Radiologist 3 (R3) and Radiologist 4 (R4), both also EDiR-certified radiologists with 7 years of experience in general radiology, were responsible for case selection, the verification of input quality, and consensus-based evaluation of the model and radiologist responses throughout the study.

The study methodology adhered to the Standards for Reporting Diagnostic Accuracy Studies statement and was based on relevant items from the TRIPOD-LLM reporting framework for studies involving LLMs and

Main points

- This study evaluated the performance of GPT-5 in abdominal radiology by assessing its diagnostic accuracy and differential diagnosis performance with two different input formats (only visual vs. visual with imaging findings and clinical presentation) using browser-based GPT-5 (default settings) and GPT-5 via the OpenAI application programming interface with controlled hyperparameters (temperature/top-p), demonstrating the potential contribution of GPT-5 assistance to abdominal radiology.
- The diagnostic accuracy of GPT-5 improved markedly when supplemented with imaging findings and clinical presentations (from 12% to 58%). Hyperparameter optimization further enhanced the diagnostic and differential diagnosis performance of the model [from 58% to 73% and from 3.44 ± 1.10 , 4 (4–3) to 3.84 ± 0.81 , 4 (4–3)].
- Systematic hyperparameter optimization revealed that the optimal settings significantly enhanced both diagnostic and differential diagnosis performance compared with different settings and browser-based settings, demonstrating the critical role of hyperparameter optimization in large language model (LLM) performance.
- Radiologists' performance was markedly improved with GPT-5 assistance; unassisted diagnostic accuracy rates of 73% and 71% increased to 87% and 86% with browser-based GPT-5 assistance and further to 94% with optimal settings assistance, aligning their performance with that of an experienced abdominal radiologist [92%, Likert score: 5 (5–49)].
- These findings underscore the potential of optimized LLMs as effective decision-support tools in abdominal radiology while emphasizing the need for further research to validate their utility in diverse clinical settings.

LMMs/VLMs, including explicit reporting of the model version, input structure, hyperparameter configuration, and radiologist–AI interaction.²¹ This study was not registered; the protocol, TRIPOD-LLM checklist, and analysis plan were finalized prior to model querying and are provided in Supplementary Table 1. All data supporting the findings of this study are available within the text and Supplementary Table 2.

The flowchart of the study is illustrated in Figure 1.

Data collection

From the “Abdominal Imaging” section of the Eurorad database (<https://www.eurorad.org/advanced-search?filter%5B0%5D=section%3A37>), 100 cases were randomly selected by R3 and R4, who assessed the responses together. A total of 14 cases were excluded based on predefined criteria to ensure the integrity and diagnostic neutrality of the dataset, including 4 cases excluded due to the absence of a specified differential diagnosis and 6 cases removed because the correct diagnosis was explicitly mentioned within the imaging findings section, potentially biasing model interpretation. An additional three cases were excluded owing to the lack of descriptive findings in the imaging findings section, and one case was omitted because it contained only a video file without an accompanying static radiologic image

suitable for model analysis. All the included cases were considered abdominal radiology cases, comprising both primary abdominal pathologies and multisystemic diseases with predominant abdominal organ involvement (the diagnostic reasoning was driven primarily by abdominal imaging findings).

For each case, multiple radiological images were available online. From these, R3 selected the two most diagnostically relevant images, excluding those with visual cues (e.g., arrows or annotations). Images were captured as unmodified JPEG screenshots and uploaded to GPT-5. The clinical presentation and imaging findings were transcribed verbatim without alteration or preprocessing, as all cases were already anonymized and publicly accessible. No additional enhancement or formatting was performed. None of the models were pretrained with any specific information by authors prior to the study. To minimize memory contamination, all cases were entered on the same day using seven independent accounts across browser-based environments, each with distinct temperature and top-p settings. In addition, to assess potential data contamination, we screened model outputs for verbatim case-title matches and for near-exact reproduction of the reference diagnosis phrasing. No verbatim title matches were identified, and near-exact phrasing overlap was observed in 2/86 cases (2.3%), which were retained but flagged in the sensitivity analyses.

Prompting and the hyperparameter tuning process

A single structured, zero-shot following prompt was used uniformly across all GPT-5 models:

“As a highly experienced radiology professor with 25 years of expertise in abdominal radiology, you will help me to solve abdominal cases. I will give you radiological images and sometimes give their imaging findings and clinical presentations of the cases. Your task is to analyze the images, imaging findings, and clinical presentations and then combine them to obtain the most likely diagnosis for the patient. Give me also the best four differential diagnoses.”

The prompt employed role-based contextualization to emulate senior radiologist reasoning, aiming to enhance clinical relevance and foster detailed differentials. To prevent bias from variable session prompts, one standardized format was applied without iterative refinement. Prior to the study, the GPT-5 models received no additional training or author-provided information that could influence outcomes.

As the internal settings of browser-based GPT-5 are undisclosed and non-configurable, its performance was assessed as a baseline. To overcome this limitation, additional experiments were conducted via OpenAI’s API, enabling manual adjustment of the tempera-

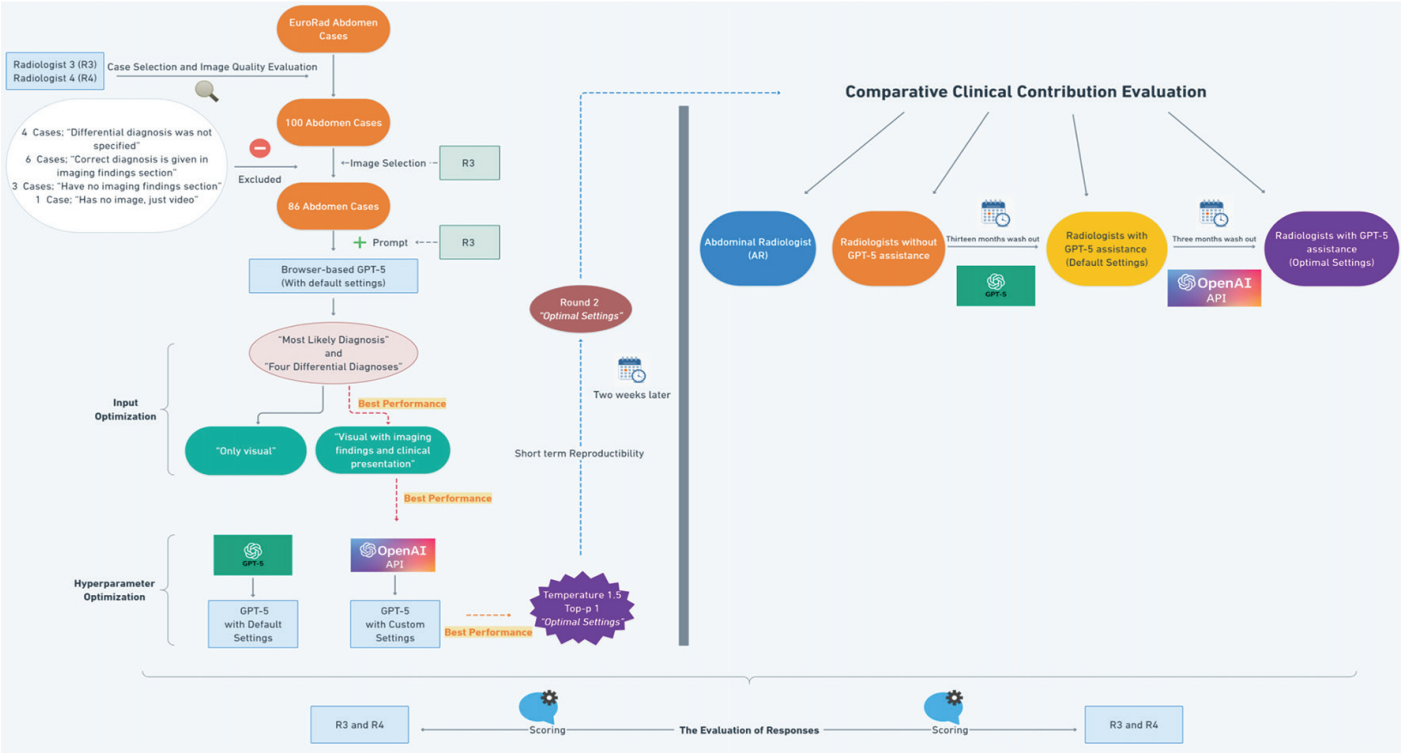


Figure 1. Study flowchart. API, application programming interface.

ture and top-p settings, which influence response determinism and diversity. Separate accounts were used for each setting to avoid contamination or memorization. The models were first tested across temperature values (0, 0.5, 1, 1.5) with a fixed top-p value—higher temperature values produced unevaluable outputs. The optimal temperature was then combined with the three top-p values (0, 0.5, 1) for further performance optimization.

Browser-based GPT-5

In the first phase of the study, browser-based GPT-5 (<https://chatgpt.com>; GPT-5-2025-08-08), which does not allow user control over the hyperparameter settings, was evaluated in two input formats: “only visual” and “visual with imaging findings and clinical presentation.” To mitigate version drift, all browser-based queries were performed on the same day using the same displayed model build (GPT-5-2025-08-08), and API queries were executed using a fixed model identifier. For each case, the model was asked to pro-

vide the most likely diagnosis and four differentials (Figure 2). All cases were entered by R3 in a single tab and session to avoid memory contamination. Although this minimized inter-session variability, it may have introduced contextual carryover effects between cases due to the continuous chat structure.

OpenAI's application programming interface (GPT-5 with different hyperparameter settings)

In the second step, using OpenAI's API (<https://platform.openai.com>) and allowing the explicit adjustment of hyperparameters, the same cases were uploaded again to GPT-5 at different temperature and top-p settings, which are the two most important hyperparameters affecting the randomness, creativity, and precision of GPT-5 responses. In this step, cases were uploaded to GPT-5 in “visual with imaging findings and clinical presentation” format. Temperature was evaluated at settings of 0, 0.5, 1, and 1.5. At temperature: 2, GPT-5 does not respond correctly as text,

often generating codes instead of text and producing answers that cannot be evaluated; therefore, no evaluation was performed in this setting. The temperature setting at which GPT-5 has the highest diagnostic accuracy was determined, and the cases were uploaded again at this temperature setting with different top-p settings (0, 0.5, and 1). Although these settings were selected to optimize diagnostic performance by balancing consistency and creativity, the same cases were uploaded again 2 weeks later using the optimal settings and the same format to assess short-term response stability and reproducibility; this was referred to as “Round 2.”

Contribution of GPT-5 assistance to radiologist performance

In July 2024, R1 and R2 independently evaluated all cases offline using R3's internet-isolated computer without access to ChatGPT. For each case, they recorded the most likely diagnosis and four differential diagnoses. All evaluations by R1, R2, and AR were performed under blinded conditions, using personal computers without internet access. Each case included clinical history and corresponding radiological images.

Following a 13-month washout, in August 2025, R1 and R2 reassessed the cases in randomized case order but this time reviewing anonymized responses from browser-based GPT-5 (comprising the model's most likely and differential diagnoses) without knowing the source model. After a 3-month washout, in November 2025, they repeated the evaluation with GPT-5 outputs generated under the optimal settings, again blinded to the model identity.

At no step were R1 or R2 informed of the correct diagnoses. All model outputs were formatted into standardized digital folders containing anonymized clinical data, radiological images, and GPT-5 responses. These were securely transferred via encrypted drives and reviewed on offline systems by R3, ensuring complete blinding and data isolation. Finally, to better demonstrate the contribution of GPT-5's assistance to radiologists, AR evaluated the same cases without GPT-5 assistance in July 2024, and AR's performance was compared with those of R1 and R2.

Fundamentals of response evaluation

Through consensus, R3 and R4 assessed the responses of all the models and radiologists (R1, R2, and AR), categorizing the “most likely diagnosis” as either correct (1) or incor-

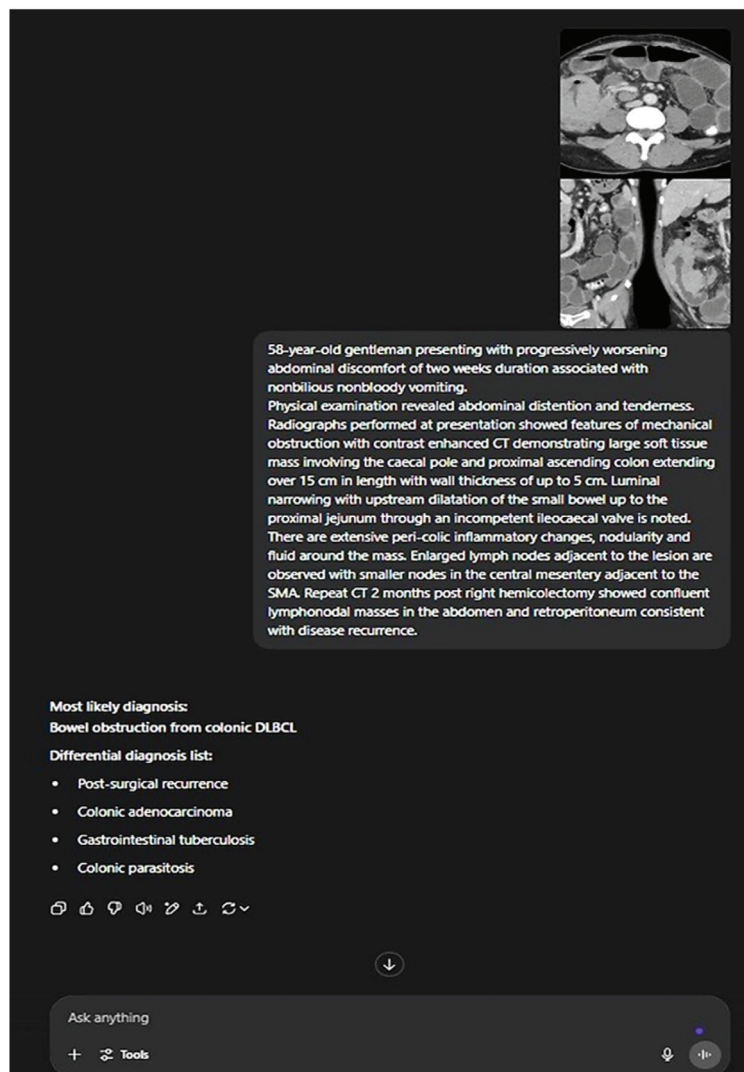


Figure 2. Chat session example with browser-based GPT-5. CT, computed tomography.

rect (0) and the “differential diagnosis” according to a 5-point Likert score:

- 1: 0/4 differentials are correct
- 2: 1/4 differentials are correct
- 3: 2/4 differentials are correct
- 4: 3/4 differentials are correct
- 5: 4/4 differentials are correct

The order of the differential diagnoses was not considered in the scoring. The responses were evaluated by R3 and R4 through consensus using only the differentials listed in the Supplementary Tables 1 and 2, and R3 and R4 independently scored a randomly selected subset (25/86 cases; 29.1%) prior to reaching consensus. Inter-assessor agreement was substantial for diagnostic correctness (Cohen’s κ : 0.82) and moderate to substantial for the 5-point differential diagnosis score (weighted κ : 0.84). No additional diagnoses were accepted as “correct” beyond those provided in the answers. However, synonyms of the terms used (e.g., “celiac disease” vs. “gluten-sensitive enteropathy” or “volvulus” vs. “intestinal torsion”) were scored as correct when medically equivalent.

This binary scoring (correct or incorrect) approach was chosen because each case had a clearly defined “most likely” (correct) diagnosis derived from the dataset, allowing objective evaluation. By contrast, a differential diagnosis inherently reflects a graded spectrum of alternatives and cannot be categorized only through binary scoring; therefore, a 5-point Likert scale was used to assess the degree of overlap between the responses and the reference differential list. The model inputs, responses, scoring criteria, and reader–AI interaction workflows were predefined and documented in accordance with key TRIPOD-LLM recommendations to enhance reproducibility and interpretability.

Statistical analysis

Descriptive statistics included mean, median, standard deviation, range, and frequencies with percentages. Normality was assessed using the Kolmogorov–Smirnov test. Diagnostic accuracy differences were evaluated using the McNemar test, and the Likert scores for GPT-5 and radiologists were compared using the Wilcoxon signed-rank test. Because comparisons across multiple inference settings were exploratory, we report unadjusted *P* values and interpret findings in the context of potential multiplicity. “Most likely diagnosis” and “differential diagnosis” Likert scores were not normally dis-

tributed (Kolmogorov–Smirnov, *P* < 0.001). Agreement between GPT-5 with the optimal settings and the Round 2 results was assessed using Cohen’s kappa. Analyses were performed using SPSS software (version 26.0; IBM Corp., Armonk, NY, USA). Statistical significance was set at *P* < 0.05.

Results

Diagnostic and differential diagnosis performance of browser-based GPT-5 and across different hyperparameter settings

Browser-based GPT-5 demonstrated limited diagnostic accuracy in the “only visual” format, correctly answering 12% of cases [10/86; 95% confidence interval (CI), 6%–21%]. When the imaging findings and clinical presentations were added, performance markedly improved to 58% (50/86; 95% CI: 48%–68%) (*P* = 0.0006; Table 1). Differential diagnosis performance similarly improved from a mean of 1.85 (95% CI: 1.65–2.05) to 3.44 (95% CI: 3.25–3.63) (*P* = 0.0004).

At temperature settings of 0, 0.5, 1, and 1.5, diagnostic accuracies were 53% (46/86; 95% CI: 42%–63%), 64% (55/86; 95% CI: 54%–74%), 66% (57/86; 95% CI: 56%–76%), and 73% (63/86; 95% CI: 63%–82%), respectively. A significant improvement over browser-based performance was observed only at temperature: 1.5 (*P* = 0.007; Table 2). In the differential diagnoses, higher temperatures correlated with improved performance, with scores increasing from 3.05 (95% CI: 2.88–3.22; temperature: 0) to 3.84 (95% CI: 3.66–4.02; temperature: 1.5). Similarly, tuning the top-*p* parameter revealed a trend toward improved differential diagnoses. In addition, GPT-5 achieved a mean Likert score of 3.19 (95% CI: 3.01–3.37) at top-*p*: 0, 3.45 (95% CI: 3.27–3.63) at top-*p*: 0.5, and 3.84 (95% CI: 3.66–4.02) at top-*p*: 1, with top-*p*: 1 significantly outperforming top-*p*: 0 (*P* = 0.043; Table 3).

Overall, the optimal settings were defined as temperature: 1.5 and top-*p*: 1, which enabled GPT-5 to achieve its best diagnostic and differential diagnosis performance. Although there was a minor difference in the

Table 1. Diagnostic accuracy rates of browser-based GPT-5 with different evaluation formats		
Format	Accuracy rate (%)	<i>P</i>
Only visual	12	<i>P</i> 0.0006
Visual with imaging findings and clinical presentation	58	
<i>P</i> values obtained from McNemar test.		

Table 2. Comparison of the diagnostic performances of GPT-5 at different temperature settings					
	Temperature 0	Temperature 0.5	Temperature 1	Temperature 1.5	Diagnostic accuracy (%)
Temperature 0	–	0.004	0.013	0.0009	53
Temperature 0.5	0.004	–	0.727	0.057	64
Temperature 1	0.013	0.727	–	0.070	66
Temperature 1.5	0.0009	0.057	0.070	–	73
<i>P</i> values obtained from Wilcoxon test.					

Table 3. Comparison of the diagnostic performances of GPT-5 at different top- <i>p</i> settings				
	Top- <i>p</i> 0	Top- <i>p</i> 0.5	Top- <i>p</i> 1	Mean Likert score
Top- <i>p</i> 0	–	0.004	0.043	3.19
Top- <i>p</i> 0.5	0.004	–	0.048	3.45
Top- <i>p</i> 1	0.043	0.048	–	3.84
<i>P</i> values obtained from Wilcoxon test.				

GPT-5 responses with the optimal settings in Round 2, no significant difference was observed in diagnostic accuracy ($P = 0.41$) or differential diagnosis performance ($P = 0.36$). Agreement between GPT-5 responses generated with the optimal settings at baseline and in the Round 2 evaluation was high, with Cohen's κ : 0.882 (95% CI: 0.80–0.96) for the most likely diagnosis and κ : 0.816 for differential diagnosis performance, indicating strong agreement.

The performance of the browser-based GPT-5 and across different hyperparameter settings are provided in Figures 3 and 4.

Diagnostic and differential diagnosis performance of radiologists with and without GPT-5 assistance

Without GPT-5 assistance, R1 and R2 demonstrated diagnostic accuracies of 73% (63/86; 95% CI: 63%–82%) and 71% (61/86; 95% CI: 61%–80%), respectively. With browser-based GPT-5 assistance, their accuracy significantly improved to 87% (75/86; 95% CI: 79%–94%) and 86% (74/86; 95% CI: 78%–93%), respectively ($P = 0.001$, $P = 0.001$). Using GPT-5 with the optimal settings further improved their accuracy to 94% (81/86; 95% CI: 88%–99%) ($P = 0.031$, $P = 0.028$). Comparatively, AR achieved an accuracy of 92% (79/86; 95% CI: 85%–97%), which was superior to the accuracy of R1 and R2 without GPT-5 assistance ($P < 0.001$) but not significantly different from their performances with both browser-based GPT-5 and GPT-5 with the optimal settings (Table 4).

Likert Scores for Differential Diagnosis

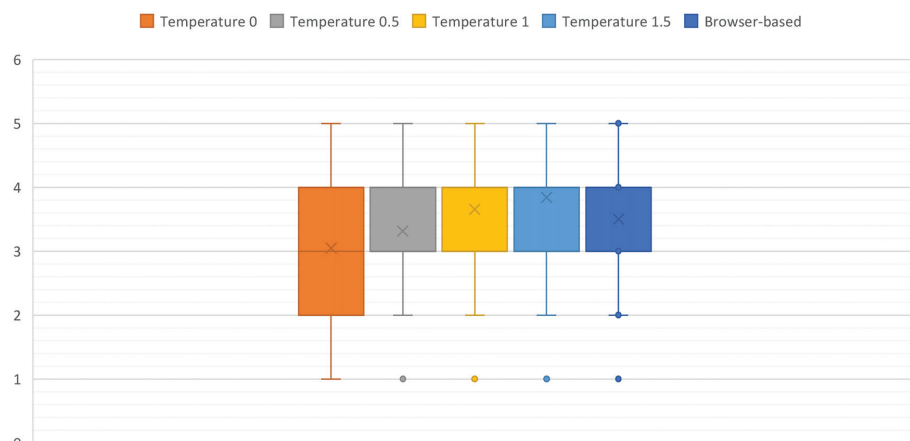


Figure 3. Differential diagnosis performances at different temperature settings and with browser-based GPT-5.

Likert Scores for Differential Diagnosis

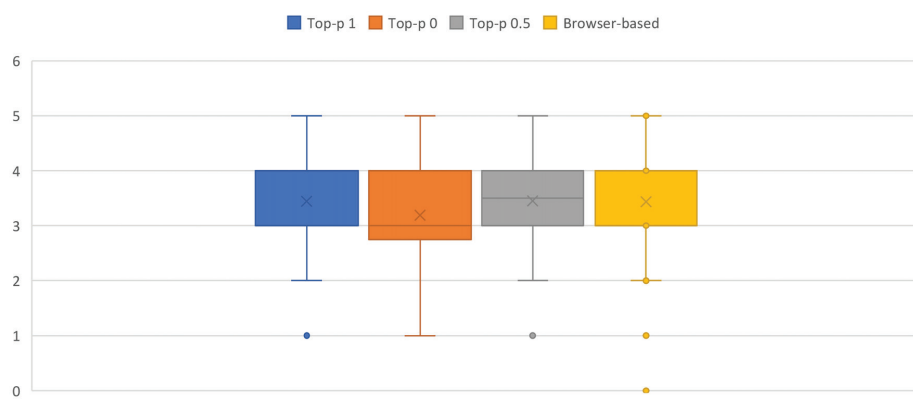


Figure 4. Differential diagnosis performances at different top-p settings and with browser-based GPT-5.

Table 4. Comparison of the diagnostic performances of radiologists with/without GPT-5 assistance and with an abdominal radiologist

	Abdominal radiologist	R1 without GPT-5 assistance	R1 with browser-based GPT-5 assistance	R1 with GPT-5 optimal setting assistance	R2 without GPT-5 assistance	R2 with browser-based GPT-5 assistance	R2 with GPT-5 optimal setting assistance	Diagnostic accuracy (%)
Abdominal radiologist	–	0.0006	0.219	0.687	0.0005	0.180	0.125	92
R1 without GPT-5 assistance	0.0006	–	0.0008	0.0005	0.227	0.001	0.0006	73
R1 with browser-based GPT-5 assistance	0.219	0.0008	–	0.031	0.0007	1	0.004	87
R1 with GPT-5 optimal setting assistance	0.687	0.0005	0.031	–	0.0002	0.016	0.250	94
R2 without GPT-5 assistance	0.0005	0.227	0.0007	0.0002	–	0.0006	0.0002	71
R2 with browser-based GPT-5 assistance	0.180	0.001	1	0.016	0.0006	–	0.002	86
R2 with GPT-5 optimal setting assistance	0.125	0.0006	0.004	0.250	0.0002	0.002	–	94

P values obtained from Wilcoxon test. R1, radiologist 1; R2, radiologist 2.

The differential diagnosis performance (mean Likert score) of R1 and R2 improved from 3.85 (95% CI: 3.67–4.03) and 3.90 (95% CI: 3.72–4.08) to 4.20 (95% CI: 4.05–4.35) and 4.24 (95% CI: 4.09–4.39), respectively, with browser-based GPT-5 assistance and further to 4.56 (95% CI: 4.43–4.69) and 4.49 (95% CI: 4.35–4.63), respectively, with the assistance of the optimal settings ($P < 0.001$, $P < 0.001$) (Table 5). Although AR achieved a mean Likert score of 4.40 (95% CI: 4.25–4.55), both radiologists outperformed AR in terms of differential diagnosis performance when assisted by GPT-5 with the optimal settings ($P = 0.002$, $P = 0.001$; Table 6).

Discussion

The most striking result of our study was that optimizing GPT-5's hyperparameters significantly enhanced both diagnostic accuracy and differential diagnosis performance in abdominal radiology. This improvement has important clinical implications, as a well-structured differential diagnosis list directly aids radiologists and clinicians by narrowing diagnostic considerations, reducing uncertainty, and potentially helping patient management decisions. The improved model performance following hyperparameter optimization likely reflects enhancements in the model's diagnostic reasoning processes. Rather than simply generating more creative

responses, fine-tuning the temperature and top-p settings may encourage the model to explore a broader but still clinically plausible range of diagnostic possibilities, improving its capacity to generate comprehensive differential diagnoses.^{14–16} This is particularly beneficial in scenarios involving ambiguous imaging findings, overlapping disease presentations, or rare pathologies, where rigid, pattern-based outputs may fall short. When calibrated appropriately, these settings help the model prioritize salient features, propose relevant alternatives, and support diagnostic reasoning akin to expert human thinking—ultimately translating into increased diagnostic confidence for the radiologist.

Table 5. Distribution of differential diagnosis Likert scores across GPT-5 conditions and radiologists

	1 Point (%)	2 Point (%)	3 Point (%)	4 Point (%)	5 Point (%)
Browser-based GPT-5 (only visual)	62	20	3	4	11
Browser-based GPT-5 (visual with imaging findings and clinical presentation)	7	7	32	40	14
GPT-5 with optimal settings	3	5	20	55	17
R1 without GPT-5 assistance	0	8	21	45	26
R1 with browser-based GPT-5 assistance	0	8	7	38	47
R1 with GPT-5 optimal setting assistance	0	0	8	22	70
R2 without GPT-5 assistance	0	11	15	46	27
R2 with browser-based GPT-5 assistance	0	9	6	36	49
R2 with GPT-5 optimal setting assistance	0	0	8	28	74
Abdominal radiologist	0	0	12	37	51

Likert score definition; 1: 0/4 correct differentials, 2: 1/4, 3: 2/4, 4: 3/4, 5: 4/4. R1, radiologist 1; R2, radiologist 2.

Table 6. Comparison of the differential diagnosis performances of radiologists with/without GPT-5 assistance and with an abdominal radiologist

	Abdominal radiologist	R1 without GPT-5 assistance	R1 with browser-based GPT-5 assistance	R1 with GPT-5 optimal setting assistance	R2 without GPT-5 assistance	R2 with browser-based GPT-5 assistance	R2 with GPT-5 optimal setting assistance	Mean Likert score
Abdominal radiologist	–	0.0006	0.002	0.002	0.0005	0.036	0.117	4.40
R1 without GPT-5 assistance	0.0006	–	0.0008	0.0002	0.285	0.0005	0.0003	3.85
R1 with browser-based GPT-5 assistance	0.006	0.0008	–	0.0004	0.0004	0.046	0.0006	4.20
R1 with GPT-5 optimal setting assistance	0.002	0.0002	0.0004	–	0.0002	0.0007	0.109	4.56
R2 without GPT-5 assistance	0.0005	0.285	0.0004	0.0002	–	0.0006	0.0003	3.90
R2 with browser-based GPT-5 assistance	0.036	0.0005	0.046	0.0007	0.0006	–	0.001	4.24
R2 with GPT-5 optimal setting assistance	0.117	0.0003	0.0006	0.109	0.0003	0.001	–	4.49

P values obtained from Wilcoxon test. R1, radiologist 1; R2, radiologist 2.

The study by Suh et al.¹⁹ suggested that temperature may influence multimodal diagnostic outputs, although reported improvements were modest and not consistently significant. Our study extends these findings by systematically evaluating both the temperature and top-p settings by focusing specifically on abdominal radiology and assessing impacts on diagnostic accuracy, differential diagnosis quality, and the contribution to radiologists' performance.

Another important result of our study is that GPT-5 provides more accurate responses for the most likely diagnoses and differentials when imaging findings and clinical presentations are provided in addition to radiological images. Previous studies have focused on the visual performance of ChatGPT.²²⁻²⁶ Dehdab et al.²⁴ reported a diagnostic accuracy of 56% for ChatGPT-4V in the interpretation of chest computed tomography (CT); however, performance improved markedly to 83.3% in cases of diffuse Coronavirus Disease 2019 involvement, likely due to more conspicuous imaging features. Similarly, Kuzan et al.²⁶ demonstrated that the model exhibited high accuracy in identifying magnetic resonance imaging (MRI) sequences (approximately 89%) and reasonable sensitivity (79.6%) for detecting diffusion restriction in acute stroke imaging. Conversely, Ren et al.²³ observed limited diagnostic performance in the detection of osteosarcoma on radiographs, with an overall accuracy of only 20%, underscoring the current limitations of LLMs in direct image interpretation. Horiuchi et al.²⁵ further compared the diagnostic capabilities of ChatGPT-4 (text-based input) and ChatGPT-4V (image-based input) in musculoskeletal imaging, concluding that the text-based model demonstrated superior diagnostic accuracy. Our study is unique in that it reveals how the performance of ChatGPT changes when supported by imaging findings and clinical presentations. LLMs use natural language processing as a starting point; because of their nature, it is likely that when imaging findings are described and clinical presentations provided, LLMs are better able to evaluate and analyze this information in a text-based manner.

Previous studies have evaluated the diagnostic performance of ChatGPT in various sections. Horiuchi et al.²⁵ evaluated the diagnostic performance of ChatGPT-4 on 100 consecutive cases from the American Journal of Neuroradiology "Case of the Week," reporting that the diagnostic accuracy of ChatGPT-4 in these cases was 50%.¹³ Moreover, Kahalian et al.²⁷ uploaded 52 radiological images to ChatGPT-4 in their study on

oral and maxillofacial pathologies, reporting that the model had a diagnostic accuracy of 56.9% when given a hint of an imaging finding in addition to these images. Similarly, our results demonstrate that browser-based GPT-5 has a diagnostic accuracy of 58%.

As a diagnostic adjunct, ChatGPT's recommendations may reinforce radiologist confidence by serving as an AI-driven cognitive checklist, potentially reducing diagnostic omissions. Alignment between the model's differential output and the radiologist's impression encourages broader deliberation while allowing final synthesis within the clinical, laboratory, and imaging context. Future research should assess radiologist performance under varying conditions—receiving only the top-ranked diagnosis, only the full differential list, or both—to clarify the respective contributions of broad versus focused AI support. Such comparative analyses will guide the optimal integration of LLMs into radiologic practice, determining whether expansive reasoning or targeted guidance provides the greatest diagnostic benefit.

We noted that the diagnostic performance of radiologists improved with GPT-5 assistance both with browser-based and optimal settings. Similarly, Siepmann et al.²⁸ evaluated the influence of ChatGPT-4 assistance on radiological interpretation by asking six radiologists with varying levels of experience to assess 40 different radiological images—including X-ray, CT, MRI, and angiographic examinations—in both unassisted and ChatGPT-4-assisted sessions. ChatGPT-4 assistance slightly increased the diagnostic accuracy of the radiologists, as evidenced by an improvement from 75.4% to 78.3%, but this difference was not statistically significant.²⁸ This study is the first to demonstrate that GPT-5 assistance not only enhances diagnostic accuracy but also facilitates more comprehensive differential diagnosis formulation, thereby contributing to improved diagnostic reasoning.

This study has several limitations. First, it was based on a single open-access dataset, which raises concerns about data contamination and limits the generalizability of the results to the broader, heterogeneous patient populations encountered in actual radiologic practice. Additionally, no real patient data were used, which further constrains the applicability of the findings to clinical settings that involve complex comorbidities and subspecialty-specific nuances.

Second, since this study focuses on hyperparameter optimization, evaluation with

different prompts/prompt engineering effects was not tested. Different prompts may affect the success of GPT-5 by providing different results in this regard. Further studies evaluating the effect of different prompts are needed.

Third, GPT-5 responded to the same cases in different time periods in this study. This may have caused differences in the responses due to both the nature of ChatGPT itself and the different responses resulting from stochasticity and updates made at different times within the model; models can improve themselves over time with new knowledge and updates. We did not conduct controlled experiments or probing to infer the default hyperparameter settings used by browser-based GPT-5. As such, potential variability stemming from undocumented or evolving internal configurations may cause performance differences at different times. Therefore, the browser-based GPT-5 responses included in the study reflect the responses at the time of the study. It should be noted that these responses may improve and change in the future.

Fourth, despite long washout periods and a randomized case order, R1 and R2 continued routine clinical practice during the 13-month interval, and a natural learning curve may have contributed to improved performance. Therefore, the observed gains cannot be attributed solely to GPT-5 assistance, and residual confounding from time-related improvement cannot be fully excluded.

Finally, although the radiologists were blinded to the model generating each response, the same cases were evaluated at different stages of the study. Despite implementing a 3-month washout period to minimize recall bias, the potential for learning bias cannot be fully excluded. Future studies should consider alternative designs or longer washout intervals to further reduce this risk.

In conclusion, GPT-5 performance varied strongly with different input formats and hyperparameter settings. Adding imaging findings and clinical presentations increased browser-based diagnostic accuracy from 12% to 58%, and API tuning improved it further to 73% at the optimal setting (temperature: 1.5, top-p: 1) with better differential diagnosis quality. Radiologist accuracy improved from 73%/71% unaided to 87%/86% with browser-based assistance and to 94%/94% with optimized assistance, approaching the AR benchmark (92%). These findings indicate a need for further studies with standardized,

documented hyperparameter configurations to develop LLM-assisted decision-support systems and improve their contribution to radiologist performance.

Acknowledgments

Eurorad cases used in this study have been approved for use in this study by “Ms. Stefanie Muzik,” who has granted the necessary permission and confirmed that these cases may be used.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Table 1 Link: <https://d2v96fxpocvxx.cloudfront.net/90a4190a-90d9-41a4-a9c9-d78d3fa8efda/content-images/591c3248-ae15-4ca8-bb66-1924aff4d612.pdf>

Supplementary Table 2 Link: <https://d2v96fxpocvxx.cloudfront.net/90a4190a-90d9-41a4-a9c9-d78d3fa8efda/documents/DIR-2026.263762-supplement-table-2.xlsx>

References

- Kasneci E, Seßler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ*. 2023;103:102274. [\[Crossref\]](#)
- Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. [\[Crossref\]](#)
- Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res*. 2023;25:e48659. [\[Crossref\]](#)
- Kuckelman IJ, Yi PH, Bui M, Onuh I, Anderson JA, Ross AB. Assessing AI-powered patient education: a case study in radiology. *Acad Radiol*. 2024;31(1):338-342. [\[Crossref\]](#)
- Nakaura T, Ito R, Ueda D, et al. The impact of large language models on radiology: a guide for radiologists on the latest innovations in AI. *Jpn J Radiol*. 2024;42(7):685-696. [\[Crossref\]](#)
- Mistry NP, Saeed H, Rafique S, Le T, Obaide H, Adams SJ. Large language models as tools to generate radiology board-style multiple-choice questions. *Acad Radiol*. 2024;31(9):3872-3878. [\[Crossref\]](#)
- López-Úbeda P, Martín-Noguerol T, Díaz-Angulo C, Luna A. Evaluation of large language models performance against humans for summarizing MRI knee radiology reports: a feasibility study. *Int J Med Inform*. 2024;187:105443. [\[Crossref\]](#)
- Almeida LC, Farina EMJM, Kuriki PEA, Abdala N, Kitamura FC. Performance of ChatGPT on the Brazilian radiology and diagnostic imaging and mammography board examinations. *Radiol Artif Intell*. 2024;6(1):e230103. [\[Crossref\]](#)
- Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*. 2024;310(1):e232756. [\[Crossref\]](#)
- Toyama Y, Harigai A, Abe M, et al. Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. *Jpn J Radiol*. 2024;42(2):201-207. [\[Crossref\]](#)
- Ariyaratne S, Jenko N, Mark Davies A, Iyengar KP, Botchu R. Could ChatGPT pass the UK radiology fellowship examinations? *Acad Radiol*. 2024;31(5):2178-2182. [\[Crossref\]](#)
- Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative performance of ChatGPT and bard in a text-based radiology knowledge assessment. *Can Assoc Radiol J*. 2024;75(2):344-350. [\[Crossref\]](#)
- Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. *Neuroradiology*. 2024;66(1):73-79. [\[Crossref\]](#)
- Lee JH, Shin J. How to optimize prompting for large language models in clinical research. *Korean J Radiol*. 2024;25(10):869-873. [\[Crossref\]](#)
- Cheat Sheet: Mastering Temperature and Top_p in ChatGPT API [Internet]. OpenAI Developer Forum. [\[Crossref\]](#)
- Davis J, Van Bulck L, Durieux BN, Lindvall C. The temperature feature of ChatGPT: modifying creativity for clinical research. *JMIR Hum Factors*. 2024;11:e53559. [\[Crossref\]](#)
- Akamine A, Hayashi D, Tomizawa A, et al. Effects of temperature settings on information quality of ChatGPT-3.5 responses: a prospective, single-blind, observational cohort study. *medRxiv*. 2024 Jun 12. Preprint. [\[Crossref\]](#)
- Suh CH, Yi J, Shim WH, Heo H. Insufficient transparency in stochasticity reporting in large language model studies for medical applications in leading medical journals. *Korean J Radiol*. 2024;25(11):1029-1031. [\[Crossref\]](#)
- Suh PS, Shim WH, Suh CH, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs from diagnosis please cases. *Radiology*. 2024;312(1):e240273. [\[Crossref\]](#)
- Eurorad [Internet]. Vienna: European Society of Radiology. [\[Crossref\]](#)
- Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology*. 2015;277(3):826-32. [\[Crossref\]](#)
- Kaba E, Solak M, Beyazal M. Evaluating ChatGPT-4o in diffusion-weighted imaging interpretation: is it useful? *Acad Radiol*. 2025;32(1):591-593. [\[Crossref\]](#)
- Ren Y, Guo Y, He Q, et al. Exploring whether ChatGPT-4 with image analysis capabilities can diagnose osteosarcoma from X-ray images. *Exp Hematol Oncol*. 2024;13(1):71. [\[Crossref\]](#)
- Dehdab R, Brendlin A, Werner S, et al. Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment. *Jpn J Radiol*. 2024;42(10):1168-1177. [\[Crossref\]](#)
- Horiuchi D, Tatekawa H, Oura T, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. *Eur Radiol*. 2025;35(1):506-516. [\[Crossref\]](#)
- Kuzan BN, Meşe İ, Yaşar S, Kuzan TY. A retrospective evaluation of the potential of ChatGPT in the accurate diagnosis of acute stroke. *Diagn Interv Radiol*. 2025;31(3):187-195. [\[Crossref\]](#)
- Kahalian S, Rajabzadeh M, Öçbe M, Medisoglu MS. ChatGPT-4.0 in oral and maxillofacial radiology: prediction of anatomical and pathological conditions from radiographic images. *Folia Med (Plovdiv)*. 2024;66(6):863-868. [\[Crossref\]](#)
- Siepmann R, Huppertz M, Rastkhiz A, et al. The virtual reference radiologist: comprehensive AI assistance for clinical image reading and interpretation. *Eur Radiol*. 2024;34(10):6652-6666. [\[Crossref\]](#)



Keros-Net: a convolutional block attention module–squeeze-and-excitation-integrated hybrid learning framework for olfactory fossa depth classification

Zülküf Akdemir¹
 Murat Canayaz²
 Abdulaziz Yalınkılıç³
 Suat Orak⁴

¹Van Yüzüncü Yıl University Faculty of Medicine,
Department of Radiology, Van, Türkiye

²Van Yüzüncü Yıl University Faculty of Engineering,
Department of Computer Engineering, Van, Türkiye

³Van Yüzüncü Yıl University Faculty of Medicine,
Department of Otorhinolaryngology, Van, Türkiye

⁴Lokman Hekim Van Hospital, Clinic of Radiology,
Van, Türkiye

PURPOSE

This study aimed to develop a hybrid decision support framework combining deep learning (DL) and machine learning (ML) to automatically classify olfactory fossa depth on paranasal computed tomography (CT) images according to the Keros classification. The goal was to enhance accuracy, reduce observer variability, and support safer endoscopic sinus surgery.

METHODS

A retrospective dataset of 481 individuals (1,549 cropped coronal CT slices) was analyzed and labeled as Keros types I–III by an experienced radiologist. Deep features were extracted using DenseNet121, DenseNet169, and DenseNet201 architectures enhanced with convolutional block attention modules (CBAMs) and squeeze-and-excitation (SE) blocks. Four feature selection techniques—SHapley Additive exPlanations, recursive feature elimination (RFE), principal component analysis, and SelectKBest—were applied to reduce dimensionality. Selected features were classified using support vector machines (SVMs), random forest, XGBoost, logistic regression, and Naive Bayes. Five-fold cross-validation was used to assess accuracy, precision, recall, and F1-score metrics.

RESULTS

Among the baseline models, DenseNet169 achieved the highest accuracy (88.37%). After feature selection, the RFE + SVM and RFE + logistic regression combinations yielded the best performance with an accuracy of 97.90%, demonstrating a substantial improvement over DL models alone. The most effective feature selection technique was RFE, and SVMs consistently produced well-balanced classification results.

CONCLUSION

Integrating CBAM–SE-enhanced DenseNet architectures with optimized feature selection and classic ML classifiers enables highly accurate and reliable automatic classification of Keros types. The proposed hybrid approach outperforms conventional DL models and provides a robust framework for objective radiological assessment.

CLINICAL SIGNIFICANCE

Accurate preoperative identification of olfactory fossa depth is essential for preventing complications such as cribriform plate injury and cerebrospinal fluid leakage during endoscopic sinus surgery. The proposed system offers an efficient, reproducible, and objective tool that may enhance surgical planning, reduce operator dependency, and increase patient safety.

KEYWORDS

Keros classification, olfactory fossa, computed tomography, deep learning, convolutional block attention modules, squeeze-and-excitation blocks, feature selection, machine learning

Handling editor: Tuğba Akıncı D'Antonoli

Corresponding author: Zülküf Akdemir

E-mail: za.radyoloji@hotmail.com

Received 20 November 2025; revision requested 16
January 2026; accepted 25 February 2026.



Epub: 18.03.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263749

You may cite this article as: Akdemir Z, Canayaz M, Yalınkılıç A, Orak S. Keros-Net: a convolutional block attention module–squeeze-and-excitation-integrated hybrid learning framework for olfactory fossa depth classification. *Diagn Interv Radiol.* 2026;32(5):600-607.

Endoscopic surgery of the sinonasal region has become increasingly common in recent years. Because of its close relationship with the skull base and neighboring structures, this area carries considerable risk in terms of anatomical variations and complications. Surrounding landmarks include the anterior skull base, cribriform plate, lamina papyracea, optic nerve, and internal carotid artery. To reduce complications, surgeons must have detailed anatomical knowledge and make effective use of advanced imaging techniques.^{1,2} Thus, radiological assessment before and during surgery plays a key role in identifying complex variations and localizing pathology with precision.

Among the important structures in this region, the anterior ethmoidal artery (AEA) is particularly vulnerable during anterior skull base procedures.³ Injury to the AEA may lead to severe outcomes, including epistaxis, vision loss, intracranial hemorrhage, or cerebrospinal fluid (CSF) leakage.⁴ Therefore, careful evaluation of anatomical landmarks and classification of the AEA's course are crucial for improving surgical safety.⁵⁻⁷

The Keros classification is especially relevant in this context because it describes the depth between the ethmoid roof and the cribriform plate: type I: 1–3 mm, type II: 4–7 mm, and type III: 8–16 mm. Type III is considered the most critical variation, as a deeper cribriform plate position increases the risk of serious complications.⁸ Preoperative assessment using this classification helps reduce the likelihood of intracranial penetration and supports safer, more controlled interventions.^{3,9-13}

Main points

- This study introduces the first hybrid framework that integrates convolutional block attention module- and squeeze-and-excitation block-enhanced DenseNet architectures with multiple feature selection strategies for automatic Keros classification from paranasal computed tomography scans.
- The proposed model achieved a substantial performance improvement, increasing baseline DenseNet accuracy from 83%–88% to up to 97.9% using recursive feature elimination combined with support vector machines or logistic regression.
- By reducing inter-observer variability and providing objective, automated assessment of olfactory fossa depth, the framework offers strong potential as a clinical decision support tool for safer endoscopic sinus surgery.

Computed tomography (CT) of the paranasal sinuses is routinely used not only for evaluating sinonasal disease before surgery but also for mapping anatomical structures and their relationships, thereby improving surgical planning and reducing complications.⁹ Accurate classification of olfactory fossa depth remains essential for preventing complications such as cribriform plate damage and CSF leakage; however, manual interpretation of CT images can vary between radiologists and over time, introducing subjectivity into the assessment.

In parallel with these developments, artificial intelligence (AI) has rapidly gained importance in medicine. AI-based decision support systems are increasingly used in medical imaging to improve diagnostic accuracy, reduce intraoperative risks, and support physicians by easing workloads. Machine learning (ML) and deep learning (DL), particularly deep convolutional neural networks (CNNs), enable automatic feature extraction and robust classification from CT, magnetic resonance imaging, and ultrasound, facilitating more objective and reproducible imaging-based assessments.

Motivations

Given that manual Keros assessment is time-consuming and subject to inter-observer variability—especially in cases with subtle anatomical differences—there is a clear need for automatic, objective, and reproducible classification methods. In this context, AI offers an opportunity to reduce subjectivity, increase efficiency, and, ultimately, improve surgical outcomes. Our motivation was to address these challenges by developing a reliable decision support framework that can integrate seamlessly into clinical practice.

Contributions

To the best of our knowledge, this is the first study to apply a hybrid DL and ML approach for automatic olfactory fossa depth classification based on the Keros system. We propose DenseNet architectures enhanced with convolutional block attention modules (CBAMs) and squeeze-and-excitation (SE) blocks for deep feature extraction, followed by feature optimization using SHapley Additive exPlanations (SHAP), recursive feature elimination (RFE), principal component analysis (PCA), and SelectKBest and classification using supervised algorithms including support vector machines (SVMs), random forest, logistic regression, Naive Bayes, and XGBoost. This integrated pipeline achieved an

accuracy of 97.9% and demonstrated strong performance in anatomically critical regions, supporting its robustness and potential for clinical application as an explainable and efficient decision support tool for surgical planning in sinonasal disease.

Related research

In their review, Alsalama et al.¹⁴ provided an overview of imaging and modeling techniques based on paranasal sinus anatomy, emphasizing automatic segmentation and classification approaches. They highlighted that sinus morphology can support individual identification and demographic profiling (e.g., age and sex) when conventional methods are unavailable and discussed the role of three-dimensional (3D) modeling in predicting demographic traits such as age, sex, and ethnicity, underlining the reliability of sinus structures for forensic identification and demographic inference.¹⁴

Wang et al.¹⁵ created a CT dataset of 242 patients with chronic rhinosinusitis with nasal polyps (CRSwNP) and applied a customized 3D nnU-Net v2 (Division of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany model for segmentation and deep feature extraction. They reported strong performance (Dice 92.8%, Intersection over Union 86.64%, accuracy 99.69%, precision 92.63%, recall 93.22%) while noting that the dataset consisted only of surgical CRSwNP cases, which may limit generalizability to broader clinical populations and across centers.¹⁵

In our study, we used high-resolution paranasal CT scans from a larger and more diverse population, with careful cropping to standardize inputs. To improve sensitivity to subtle or low-contrast regions, DenseNet was augmented with CBAMs and SE attention blocks, enabling more effective focus on anatomically subtle structures and improving classification performance across patient subgroups.

Similarly, Lee et al.¹⁶ proposed a CNN-based model to automatically compute the Lund–Mackay score (LMS) by segmenting sinus regions on CT and directly generating LMS values. Using 1,399 patients and manual segmentations of 77 scans for training, they achieved an average Dice score of 0.85, with LMS prediction accuracy across regions ranging from 86% to 99%, suggesting that automation can standardize scoring and support clinical decision-making.¹⁶

Beyond paranasal sinus analysis, Gogus et al.¹⁷ introduced a hybrid model combining CBAMs with Swin Transformer blocks for femoral stem implant classification, outperforming architectures such as DenseNet201, VGG19, and InceptionV3. Their MSFT-Net addressed inter-class morphological similarity through attention mechanisms and transformer layers, paralleling our approach in leveraging attention to improve feature extraction from complex medical images. In our framework, CBAMs and SE blocks similarly enhance focus on subtle regions such as the lateral lamella in paranasal CT, improving sensitivity and classification accuracy; together, these findings support the value of attention-enhanced designs for reliable performance in anatomically challenging scenarios.

Methods

We retrospectively reviewed the CT images of 4,427 patients who underwent non-contrast paranasal sinus scans between April 2020 and January 2024, retrieved from the Picture Archiving and Communication System. Based on the study criteria, a total of 481 individuals aged 18–67 years (mean age: 29.1 ± 10.5 years) were included in the analysis. Of these, 255 (53%) were men and 226 (47%) women.

Routine paranasal sinus CT examinations were performed with patients in the supine position, the head placed in hyperextension. Axial images were obtained using 64-slice SOMATOM go.Up and 128-slice SOMATOM Definition AS multidetector CT scanners (Siemens Healthineers, Erlangen, Germany). The imaging protocol was as follows; tube voltage: 100 kV, effective mAs: 67, slice thickness: 3.0 mm, field of view: 170 mm, pitch: 0.8–1 mm, rotation time: 1 s, and collimation: 128×0.6 mm. The bone window settings were WW: 2,200 and WL: 475. Reconstructions were performed in the coronal plane perpendicular to the hard palate, with 0.75-mm slice thickness, using the J70h very sharp bone filter as standard.

For evaluation, only coronal reformatted images were used. Classification followed the Keros system, which defines the depth of the olfactory fossa as type I (1–3 mm), type II (4–7 mm), and type III (8–16 mm). In cases of asymmetry, the side with greater depth was considered. Ultimately, 1,549 images were prepared, consisting of 330 for type I, 677 for type II, and 542 for type III. Images were saved in JPEG format, averaging 3.2 slices per patient.

All images were uniformly cropped according to predefined anatomical reference points. The superior border included the crista galli, the lateral borders encompassed the medial walls of the maxillary sinuses, and the inferior limit was set at the superior surface of the palatine process of the maxilla.

Inclusion criteria: Patients over 18 years of age who underwent paranasal sinus CT imaging.

Exclusion criteria: Cases with ethmoidal or frontal mucosal thickening, sinonasal polyps, poor positioning, image artifacts, inverted papilloma, septal or cribriform fractures, nasolacrimal duct pathologies, osteoma, foreign bodies, intranasal masses, crista galli lesions, septal perforation, prior surgery, or low-quality scans.

All image selection and cropping were performed by a radiologist with 14 years of experience.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

The de-identified participant data and analysis outputs are available from the corresponding author upon reasonable request and subject to institutional/ethical approvals. This study was approved by Van Yüzüncü Yıl University Non-Interventional Clinical Research Ethics Committee (protocol number: 2023/05-03, date: 12.05.2023).

Proposed approach

We developed a three-stage hybrid framework for the automatic classification of olfactory fossa depth. First, a labeled dataset of CT images was created. These images were then processed using a DenseNet-based fea-

ture extractor augmented with CBAMs and SE blocks to extract high-dimensional deep feature vectors. The CBAMs improved the model's ability to attend to spatial and channel-wise information, whereas the SE blocks rescaled feature channels to strengthen discriminative representations. The overall workflow of the proposed method is illustrated in Figure 1.

In the second stage, the extracted deep features were optimized using four different feature selection methods: SHAP, RFE, PCA, and SelectKBest. These methods identified the most meaningful subsets of features for classification and reduced model complexity. In the third stage, the selected features were fed into five different traditional classifiers (SVMs, random forest, XGBoost, logistic regression, and Naive Bayes). Each combination was tested using cross-validation and evaluated with metrics such as accuracy, precision, recall, and F1-score. Using this framework, high accuracy in Keros classification was achieved by combining the powerful feature extraction capabilities of DL with the flexibility of traditional ML methods.

Dataset and preprocessing

The image data used in this study consist of high-resolution CT scans acquired in the coronal plane of the paranasal sinuses, labeled as Keros types 1, 2, and 3 by an experienced radiologist. The data were obtained in anonymized form and used within the scope of the study after receiving ethical approval. Basic preprocessing steps such as normalization and resizing (e.g., to 256×256 pixels) were applied to prepare the images for classification. Some examples from the dataset are shown in Figure 2.

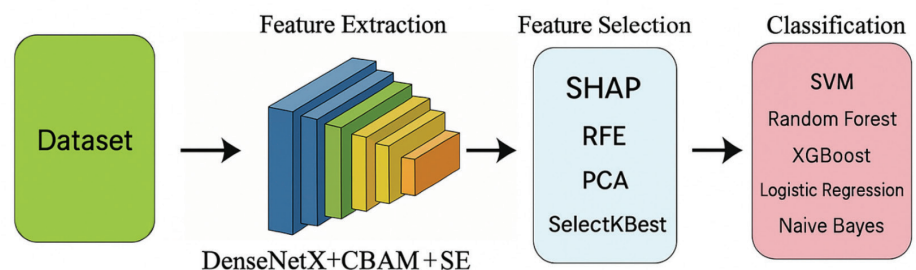


Figure 1. Structure of proposed model. Schematic overview of the proposed three-stage hybrid framework for automatic classification of olfactory fossa depth. A labeled computed tomography image dataset is first constructed and used as input to a DenseNet-based feature extractor augmented with CBAMs and SE blocks to obtain high-dimensional deep feature vectors. These features are then subjected to feature selection (SHAP, RFE, PCA, or SelectKBest), and the selected features are finally fed into machine learning classifiers (SVMs, random forest, XGBoost, logistic regression, or Naive Bayes) for final category prediction. CBAM, convolutional block attention module; SE block, squeeze-and-excitation block; RFE, recursive feature elimination; PCA, principal component analysis; SVM, support vector machine; SHAP, SHapley Additive exPlanations.

Deep feature extraction

In the feature extraction stage, the DenseNet architecture was used as the backbone. However, this base architecture was enhanced with attention mechanisms. The CBAMs optimized the attention distribution of the input images based on both spatial and channel features, whereas the SE block reinforced channel-wise key information. This design enabled the model to focus more on anatomically critical regions (e.g., the lateral lamella area). As output, a high-dimensional deep feature vector was obtained from each image.

Statistical analysis

Statistical analyses were conducted to evaluate the performance, reliability, and generalizability of the proposed hybrid classification framework. To prevent data leakage and avoid artificially inflated performance due to patient-specific anatomical similarity, all data partitioning procedures were performed at the patient level rather than at the slice level. All CT slices belonging to a given patient were assigned exclusively to either the training or testing set. An 80–20 stratified split was applied at the patient level to

preserve the distribution of Keros types, and model evaluation was performed using five-fold stratified cross-validation, ensuring strict patient-level independence across all folds. Hyperparameters for the classic ML classifiers were selected based on standard configurations and maintained across experiments. Model performance was evaluated using cross-validation on the training data. Due to the relatively limited dataset size and the primary focus on comparative performance evaluation, exhaustive hyperparameter optimization and nested cross-validation were not applied. Class imbalance was addressed at multiple stages of the pipeline. During DL model training, stratified data splitting was applied to preserve class proportions, and class weights were computed from the training labels using a balanced weighting scheme and incorporated into the loss function. For the classic ML classifiers trained on selected deep features, five-fold stratified cross-validation was employed to maintain class distribution across folds. In addition, weighted-averaged precision, recall, and F1-score metrics were reported to ensure balanced performance evaluation across all Keros types. For each experiment, accuracy, precision, recall, and F1-score were computed

and averaged across folds to obtain stable performance estimates.

To reduce the dimensionality of the deep feature vectors and enhance classifier performance, four statistical feature selection methods were applied: RFE, PCA, SelectKBest, and SHAP. RFE was used to iteratively eliminate less informative features, PCA reduced multicollinearity by transforming features into orthogonal principal components, SelectKBest ranked features based on statistical significance, and SHAP values were used to quantify feature importance and enhance model interpretability. The dimensionality of the deep feature vectors before and after each feature selection method is summarized in Table 1.

Comparative statistical evaluation was performed between baseline DL models and feature-selected ML classifiers by examining improvements in accuracy and metric consistency. Since the objective was predictive performance rather than inferential hypothesis testing, no parametric or non-parametric statistical tests were applied. All analyses were conducted using Python (v3.x), scikit-learn, SHAP library, XGBoost, NumPy, and related DL frameworks.

Machine learning classifiers

The selected features were evaluated using five classic classifiers with different characteristics:

- **Support vector machines:** Effective in constructing decision boundaries in high-dimensional spaces.
- **Random forest:** An ensemble model composed of decision trees, often successful in handling class imbalance.
- **XGBoost:** A tree-based gradient boosting method.
- **Logistic regression:** Creates linear decision boundaries and is commonly used as a baseline comparison model.
- **Naive Bayes:** A probabilistic model that assumes features are independent of each other.

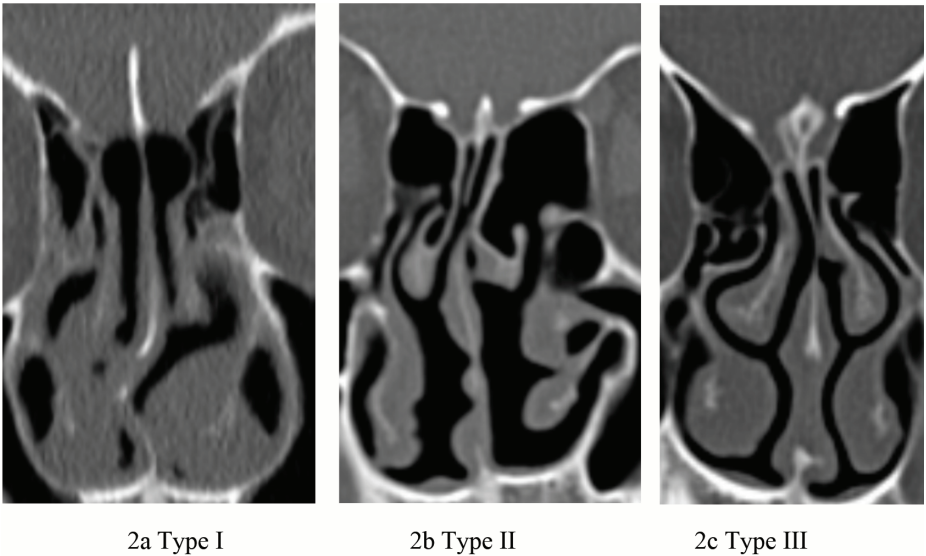


Figure 2. Representative examples of high-resolution coronal computed tomography images of the paranasal sinuses used in this study, labeled according to the Keros classification: (2a) type 1, (2b) type 2, and (2c) type 3.

Table 1. Feature dimensionality before and after feature selection						
Backbone (CBAMs + SE DenseNet)	Deep feature source layer	Before selection (n × d)	SelectKBest (k = 100)	PCA (n = 100)	RFE (n = 100)	SHAP (top 100)
DenseNet121	Dense(128) (penultimate layer)	n × 128	n × 100	N × 100	n × 100	n × 100
DenseNet169	Dense(128) (penultimate layer)	n × 128	n × 100	n × 100	n × 100	n × 100
DenseNet201	Dense(128) (penultimate layer)	n × 128	n × 100	n × 100	n × 100	n × 100
CBAM, convolutional block attention module; SE, squeeze-and-excitation; PCA, principal component analysis; RFE, recursive feature elimination; SHAP, SHapley Additive exPlanations.						

Findings and results

In the experimental findings of the study, classification was first performed directly using the DenseNet121, DenseNet169, and DenseNet201 architectures without applying any feature extraction or selection methods. The models were trained and tested for Keros classification (types 1, 2, and 3). Table 2 presents the comparative results of the baseline models in terms of accuracy, precision, recall, and F1-score metrics.

DenseNet169 achieved the highest performance across all metrics, making it the best baseline model. With an accuracy of 88.37%, it demonstrated a strong generalization ability in classifying olfactory fossa depths.

DenseNet201, with an accuracy of 86.05%, ranked second and exhibited competitive performance, particularly in terms of precision (0.8645).

DenseNet121 showed the lowest performance, with an accuracy of 83.33%. This result indicates that its shallower architecture limited its classification capacity.

The accuracy, precision, recall, and F1-score values of the models are similar. This indicates that the class distribution is balanced and that the models generally demonstrate consistent performance across classes. In particular, the minimal difference between metrics in the DenseNet169 model shows that it produces balanced and reliable outputs.

Second, deep features extracted using DenseNet architectures enhanced with CBAMs and SE blocks were optimized through different feature selection algorithms (RFE, PCA, SHAP, SelectKBest) and used for Keros classification with traditional ML classifiers (SVMs, logistic regression, Naive Bayes, etc.). Table 3 presents the comparative results of the models in terms of accuracy, precision, recall, and F1-score metrics.

DenseNet169 emerged as the highest-performing model. In particular, the combinations of RFE + SVMs and RFE + logistic regression achieved an accuracy of 97.90%, indicating a very high generalization capacity.

RFE proved to be the most effective feature selection method overall, achieving the highest accuracies with almost all models.

SVMs and logistic regression produced high and consistent results. The fact that the accuracy, precision, recall, and F1-Score values were similar indicates that the models performed balanced classification.

Comparative evaluation

Table 4 compares the accuracy values of only the baseline models (without feature selection or different classifiers) with those of the best feature selection + ML classifier combinations.

Models supported with feature selection and traditional classifiers achieved an accuracy improvement of 10%–14% compared with the baseline models.

This increase highlights that feature engineering and the choice of an appropriate classifier are critical, especially for models operating on complex and imbalanced classes.

The simplest architecture, DenseNet121, achieved 83.33% accuracy as a baseline model while reaching 97.05% accuracy with feature selection and SVMs. This demonstrates that even simpler models can deliver strong performance when supported with appropriate enhancement methods. The graphical representation of the results is presented in Figure 3.

These findings demonstrate that the integration of CBAMs + SE-enhanced deep feature extraction, feature selection, and ML classifiers is a powerful and effective approach for anatomically critical tasks that influence surgical decisions, such as Keros classification. In particular, the DenseNet169 + RFE + SVMs combination yielded the best results in terms of both accuracy and metric consistency.

Clinical usability and interface integration

To facilitate real-world clinical adoption, the proposed Keros classification model was deployed within a dedicated web-based interface designed for radiologists and endoscopic sinus surgeons. The interface, developed using the Flask framework, enables clinicians to upload cropped coronal CT slices of the olfactory fossa and instantly obtain Keros type I–III predictions with their corresponding probability estimates. The clean, task-oriented design supports multi-image uploading, automatic preprocessing, model selection, and real-time visualization of results.

The interface displays explanatory information regarding the clinical relevance of each Keros type, allowing the prediction to be contextualized alongside anatomical detail and potential surgical risk. This design contributes to decision support by present-

Table 2. Performance comparison of the baseline models

Model	Accuracy	Precision	Recall	F1 score
DenseNet121	0.8333	0.8342	0.8333	0.8326
DenseNet169	0.8837	0.8902	0.8837	0.8839
DenseNet201	0.8605	0.8645	0.8605	0.8606

The table summarizes the performance of the baseline DenseNet architectures for the classification task. DenseNet169 achieved the best overall performance among the baseline models (accuracy, 0.8837; precision, 0.8902; recall, 0.8837; F1 score, 0.8839). Accuracy, recall, precision, and F1-score are reported as proportions, with F1-score representing the harmonic mean of precision and recall.

Table 3. Performance comparison (feature-selected models)

Model	Feature selection	Classifier	Accuracy	Precision	Recall	F1 score
DenseNet169	RFE	SVM	0.9790	0.9792	0.9790	0.9791
DenseNet169	RFE	Logistic Regression	0.9790	0.9791	0.9790	0.9791
DenseNet169	PCA	SVM	0.9783	0.9784	0.9783	0.9783
DenseNet201	RFE	Naive Bayes	0.9775	0.9775	0.9775	0.9775
DenseNet201	SHAP	SVM	0.9767	0.9768	0.9767	0.9767

This table summarizes the performance of feature-selected models based on DenseNet169 and DenseNet201 backbones combined with different feature selection methods (RFE, PCA, SHAP) and classifiers (SVMs, logistic regression, Naive Bayes). The highest overall performance was achieved by RFE-based DenseNet169 models with SVMs and logistic regression (accuracy, 0.9790; precision, 0.9792 and 0.9791; recall, 0.9790; F1-score, 0.9791). Accuracy, precision, recall, and F1-score are reported as proportions, with F1-score representing the harmonic mean of precision and recall. RFE, recursive feature elimination; PCA, principal component analysis; SHAP, SHapley Additive exPlanations; SVM, support vector machine.

Table 4. Comparison of base and optimized classification accuracies for DenseNet models

Model	Base model accuracy	Best combination accuracy	Increase (%)
DenseNet121	0.8333	0.9705 (RFE + SVM)	+13.7
DenseNet169	0.8837	0.9790 (RFE + SVM)	+10.8
DenseNet201	0.8605	0.9775 (RFE + Naive Bayes)	+13.4

This table summarizes the effect of combining each DenseNet backbone with the best-performing feature selection and classifier configuration on classification accuracy. “Base Model Accuracy” indicates the performance of the convolutional neural network alone, without feature selection or an external classifier, whereas “Best Combination Accuracy” denotes the highest-performing feature selection–classifier pipeline for that backbone. “Increase (%)” indicates the percentage improvement in accuracy relative to the corresponding base model. Accuracy values are reported as proportions.

RFE, recursive feature elimination; SVM, support vector machine.

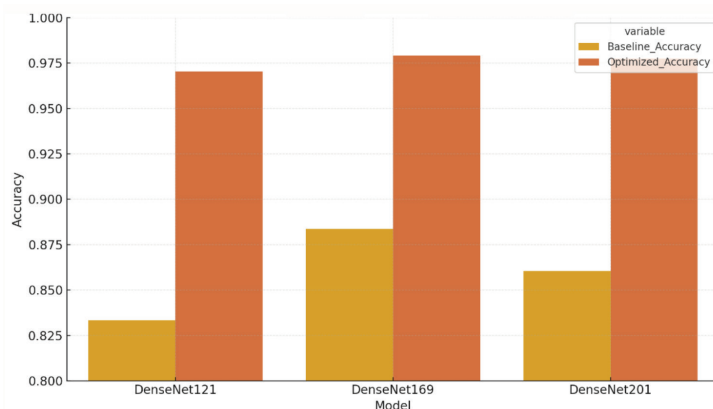


Figure 3. Comparison of baseline and optimized accuracy values for three different DenseNet architectures. The yellow bars indicate the baseline accuracy of the deep networks used without additional processing, and the orange bars represent the optimized accuracy obtained after applying feature selection and a machine learning classifier. The most prominent improvement is observed for DenseNet121, where the accuracy increases from 83.33% in the baseline model to 97.05% in the optimized configuration using feature selection and a support vector machine, indicating that even the simplest architecture can achieve high performance when appropriate enhancement methods are applied.

ing the model output together with radiological slice quality, patient history, and relevant anatomical variants. The system stores previously analyzed scans in a “History” module, enabling clinicians to revisit earlier evaluations for comparison or follow-up planning.

Importantly, the interface is optimized for routine clinical workflow without requiring graphics processing unit (GPU) hardware, as the deep feature extraction and model optimization steps are embedded within the pretrained hybrid model. This ensures rapid inference and seamless integration into standard radiology workstations. By providing reproducible, objective Keros grading with an intuitive user experience, the interface demonstrates strong potential for use in pre-operative planning, reducing observer variability, and improving surgical risk assessment. Taken together, this system represents a practical and clinician-friendly implementation of the hybrid classification framework, supporting its feasibility for real-world deployment in otolaryngology and radiology settings. A graphical representation of the interface is presented in Figure 4.

Discussion

In this study, we proposed a hybrid model to automate Keros classification of olfactory fossa depth, combining DenseNet with attention mechanisms and multiple feature selection techniques. Unlike earlier studies, our approach integrated CBAMs and SE blocks into the DenseNet framework, improving the model’s focus on critical anatomical regions such as the lateral lamella and strengthening the representational quality of the extracted features. By applying feature selection methods such as SHAP, RFE, PCA, and SelectKBest, we were able to filter deep features before feeding them into conventional ML algorithms. This not only boosted classification accuracy but also enhanced interpretability and computational efficiency. The strongest outcomes were achieved with the RFE + SVMs and RFE + logistic regression pipelines, reaching an accuracy of 97.9%, underlining the robustness of the chosen strategy and its clinical potential.

Beyond the accuracy values, the study contributes to the field by showing how au-

tomated approaches can support surgical planning. Traditional methods, which rely on manual interpretation, are time-consuming and subject to variation between observers, making the assessment of risky anatomical variations difficult. Our findings demonstrate that automated feature extraction and classification can help reduce surgical risks and improve safety. Moreover, the performance gain—rising from approximately 83%–88% in baseline DL models to nearly 98% when combined with feature selection and classic classifiers—is a substantial step forward in clinical applicability. These results confirm that model success does not depend only on network depth but also on careful feature engineering and algorithm choice.

In addition to the technical contributions of the proposed framework, we developed a web-based clinical interface to ensure its applicability in real-world clinical environments. This interface allows clinicians to upload cropped coronal CT slices and instantly receive Keros type I–III predictions along with probability estimates, enabling rapid and objective assessment. Its intuitive design, support for multiple image uploads, automated preprocessing, and integrated history management make it well suited for routine radiology workflows. Importantly, the system operates without requiring GPU hardware, as all deep feature extraction and optimization steps are embedded within the pretrained hybrid model. By reducing observer variability and providing reproducible grading of olfactory fossa depth, the interface highlights the clinical relevance and practical translational potential of the proposed hybrid decision support system.

Nevertheless, several limitations should be acknowledged. First, image labeling and cropping were performed by a single radiologist with 14 years of experience. Although this ensured consistent anatomical localization and annotation, it may introduce potential observer bias, as inter-observer variability is a known issue in Keros classification. Due to the retrospective nature of the dataset and institutional constraints, the inclusion of a second independent rater and quantitative inter-rater agreement analysis was not feasible. Future studies will aim to incorporate multi-reader annotations and inter-observer agreement metrics, such as Cohen’s kappa, to further strengthen reliability. Second, all CT images were obtained from a single center using Siemens scanners with relatively uniform acquisition parameters. Although this consistency reduces variability related to imaging protocols, it may limit the general-

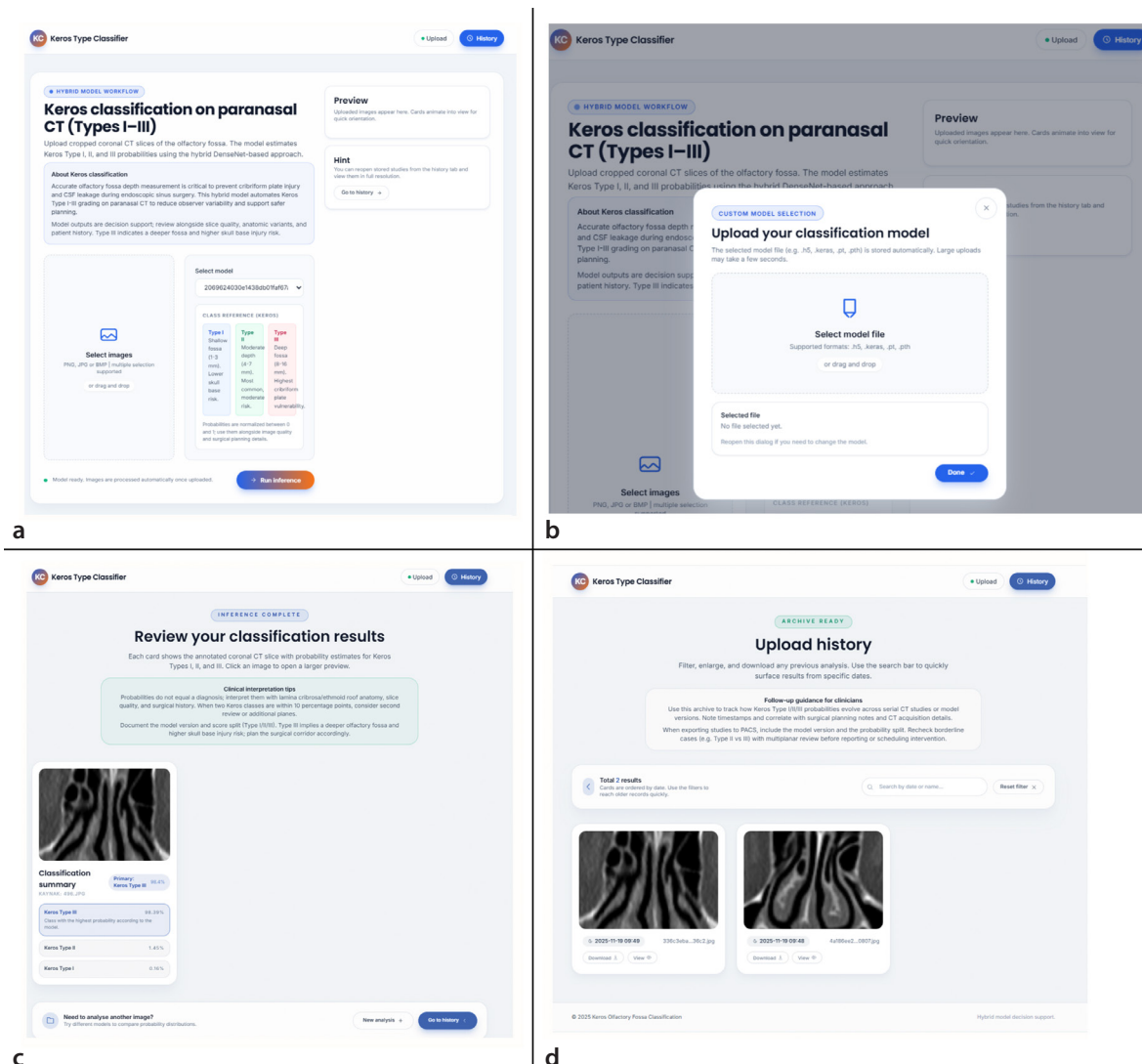


Figure 4. Clinical interface of the proposed decision support system. Overview of the clinical application interface developed for automated Keros classification. (a) Interface of application. The main interface of the application, allowing clinicians to upload coronal computed tomography slices for analysis. (b) Different models loading screen. The model selection screen, enabling users to choose among different pretrained hybrid models. (c) Show the results. The results display panel showing the predicted Keros type and corresponding probability values. (d) Show the history. The history section, where previously analyzed scans are stored for review and comparison.

izability of the proposed framework to data acquired from different institutions, scanner vendors, or acquisition settings. In addition, the dataset size and single-center design led to underrepresentation of some categories, particularly Keros type I. Future multicenter studies with larger and more heterogeneous datasets are therefore required to better assess model robustness under real-world clinical conditions. Third, nested cross-validation was not employed, which may limit the precision of generalization error estimation. This aspect should be addressed in future investigations with larger datasets that allow for more computationally intensive validation strategies. Furthermore, as a fixed DenseNet + CBAMs + SE architecture was used, the potential benefits of alternative attention mechanisms or transformer-based networks could not be explored. Future research may

investigate these architectures and incorporate interactive or explainable visualization tools to further improve interpretability and user confidence. Despite these limitations, the proposed framework demonstrates strong methodological rigor and clinical relevance and represents a meaningful contribution to the literature on AI-assisted radiological assessment of anatomically critical regions.

We presented a comprehensive modeling framework for Keros classification by integrating DL-based feature extraction, diverse feature selection methods, and classic ML algorithms. DenseNet, enhanced with CBAMs and SE blocks, was used to generate high-quality deep features, which were then optimized with selection techniques such as SHAP, RFE, PCA, and SelectKBest. These

features were classified using SVMs, random forest, logistic regression, Naive Bayes, and XGBoost.

Among the tested models, DenseNet169 combined with RFE and SVMs achieved the highest accuracy at 97.9%. This highlights the importance of both the architectural depth of the network and the choice of feature selection strategy. It is also significant that models relying solely on DL (approx. 88% accuracy) were markedly improved by the integration of feature selection and traditional classifiers. Other performance indicators, including precision, recall, and F1-score metrics, further confirmed the stability and reliability of the proposed method.

These findings suggest that combining deep feature extraction with effective fea-

ture selection and appropriate ML classifiers can enable the development of highly accurate clinical decision support systems for medical image analysis.

Although the results are promising, some areas remain for further improvement. First, the dataset was relatively small and drawn from a single center, limiting generalization. Larger, multicenter studies are needed to confirm robustness. Second, this work relied on a fixed DenseNet + CBAMs + SE design. Future research may explore other attention mechanisms, transformer-based architectures, or multi-scale feature extraction. Hybrid feature selection methods, such as SHAP combined with RFE or genetic algorithm-based selection, could also be investigated.

Another important direction concerns explainability. Although SHAP was employed, further efforts are needed to enhance interpretability and develop interfaces that clearly communicate the model's decision process to clinicians. Testing the system within real-time decision support platforms and assessing its usability in surgical planning will also be essential for successful clinical translation. Furthermore, the clinical interface developed in this study demonstrates that the proposed hybrid model can be effectively integrated into routine radiology workflows, enabling rapid, objective, and reproducible Keros classification to support surgeons in preoperative planning and risk assessment.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Snyderman CH, Pant H, Carrau RL, Prevedello D, Gardner P, Kassam AB. What are the limits of endoscopic sinus surgery?: the expanded endonasal approach to the skull base. *Keio J Med*. 2009;58(3):152-160. [\[Crossref\]](#)
2. O'Brien WT Sr, Hamelin S, Weitzel EK. The preoperative sinus CT: avoiding a "CLOSE" call with surgical complications. *Radiology*. 2016;281(1):10-21. [\[Crossref\]](#)
3. Özdemir A, Bayar Muluk N. The important adjacent structures for anterior ethmoidal artery in FESS: anterior ethmoidal artery canal angle, supraorbital ethmoid cells and Keros classification. *J Clin Neurosci*. 2022;98:207-212. [\[Crossref\]](#)
4. Ding J, Sun G, Lu Y, et al. Evaluation of anterior ethmoidal artery by 320-slice CT angiography with comparison to three-dimensional spin digital subtraction angiography: initial experiences. *Korean J Radiol*. 2012;13(6):667-673. [\[Crossref\]](#)
5. Souza SA, Souza MM, Gregório LC, Ajzen S. Anterior ethmoidal artery evaluation on coronal CT scans. *Braz J Otorhinolaryngol*. 2009;75(1):101-106. [\[Crossref\]](#)
6. Başak S, Karaman CZ, Akdilli A, Mutlu C, Odabaşı O, Erpek G. Evaluation of some important anatomical variations and dangerous areas of the paranasal sinuses by CT for safer endonasal surgery. *Rhinology*. 1998;36(4):162-167. [\[Crossref\]](#)
7. Pandolfo I, Vinci S, Salamone I, Granata F, Mazziotti S. Evaluation of the anterior ethmoidal artery by 3D dual volume rotational digital subtraction angiography and native multidetector CT with multiplanar reformations. Initial findings. *Eur Radiol*. 2007;17(6):1584-1590. [\[Crossref\]](#)
8. Abdullah B, Chew SC, Aziz ME, et al. A new radiological classification for the risk assessment of anterior skull base injury in endoscopic sinus surgery. *Sci Rep*. 2020;10:4600. [\[Crossref\]](#)
9. VAM, Santosh B. A study of clinical significance of the depth of olfactory fossa in patients undergoing endoscopic sinus surgery. *Indian J Otolaryngol Head Neck Surg*. 2017;69(4):514-522. [\[Crossref\]](#)
10. Yağmur AR, Çıvgın E, Özcan KM, Yurtsever Kum N, Karakuş MF, Dere HH. Analysis of the correlation of the lamina papyracea-to-midline distance with the location of anterior ethmoidal artery and Keros classification. *Indian J Otolaryngol Head Neck Surg*. 2023;75(4):3146-3151. [\[Crossref\]](#)
11. Şahan MH, Inal M, Muluk NB, Şimşek G. Cribriform plate, crista galli, olfactory fossa and septal deviation. *Curr Med Imaging Rev*. 2019;15(3):319-325. [\[Crossref\]](#)
12. Paudel S, Sah R, Budhathoki T, Pandey G. Evaluation of olfactory fossa depth using computed tomography. *J Nepal Health Res Counc*. 2025;22(4):707-711. [\[Crossref\]](#)
13. Babu AC, Nair MRPB, Kuriakose AM. Olfactory fossa depth: CT analysis of 1200 patients. *Indian J Radiol Imaging*. 2018;28(4):395-400. [\[Crossref\]](#)
14. Alsalama A, Harous S, Elnagar A. Paranasal sinus analysis based on deep learning and machine learning techniques: a comprehensive survey. *Intelligent Systems with Applications*. 2025;27:200559. [\[Crossref\]](#)
15. Wang Y, Zhang X, Du W, Dai N. Deep learning-based fully automatic segmentation of the paranasal sinuses in chronic rhinosinusitis patients using computed tomographic images. *IEEE Access*. 2025;13:16444-16454. [\[Crossref\]](#)
16. Lee DJ, Hamghalam M, Wang L, et al. The use of a convolutional neural network to automate radiologic scoring of computed tomography of paranasal sinuses. *Biomed Eng Online*. 2025;24(1):49. [\[Crossref\]](#)
17. Gogus E, Yılmaz A, Enercan M. A novel deep hybrid model for automatic femoral stem classification in hip arthroplasty from radiographs: MSFT-Net with CBAM and transformer modules. *IEEE Access*. 2025;13:102564-102577. [\[Crossref\]](#)



Performance of multimodal large language models for the detection and characterization of bone lesions on radiographs

Raşit Eren Büyüktoka¹

Ali Salbas²

Atilla Hikmet Cilengir³

Asli Dilara Buyuktoka²

Murat Sürücü⁴

Zehra Hilal Adibelli^{1,5}

¹University of Health Sciences Türkiye, İzmir City Hospital, Department of Radiology, İzmir, Türkiye

²İzmir Katip Celebi University, Atatürk Training and Research Hospital, Department of Radiology, İzmir, Türkiye

³İzmir Democracy University, Buca Seyfi Demirsoy Training and Research Hospital, Department of Radiology, İzmir, Türkiye

⁴Burdur Mehmet Akif Ersoy University, Bucak Faculty of Computer and Informatics, Department of Software Engineering, Burdur, Türkiye

⁵University of Health Sciences Türkiye, İzmir Faculty of Medicine, Department of Radiology, İzmir, Türkiye

Handling editor: Tuğba Akıncı D'Antonoli

Corresponding author: Raşit Eren Büyüktoka

E-mail: rasiterenbuyuktoka@hotmail.com

Received 14 February 2026; revision requested 11 March 2026; last revision requested 26 April 2026; accepted 28 April 2026.



Epub: 11.05.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263945

PURPOSE

Multimodal large language models (LLMs) offer emerging capabilities in medical image interpretation; however, their efficacy in orthopedic oncology remains unverified. This study aimed to evaluate and benchmark the performance of five contemporary LLMs—ChatGPT 5.2, Gemini 3 Flash, MedGemma 4B, Claude Sonnet 4.6, and DeepSeek-VL2—in detecting and characterizing bone lesions on plain radiographs without task-specific fine-tuning.

METHODS

A retrospective analysis was conducted using 3,746 anonymized images from the Bone Tumor X-ray Radiograph Dataset (BTXRD), comprising normal, benign, and malignant cases. Reference standard annotations were provided directly by the BTXRD dataset. Models were evaluated on two tasks: lesion detection and lesion characterization. Diagnostic performance metrics, including accuracy, precision, sensitivity, specificity, and Cohen's kappa, were calculated and compared with reference-standard annotations.

RESULTS

ChatGPT 5.2 demonstrated the highest overall accuracy (0.803) and specificity among the models (0.916) for lesion detection, although its sensitivity (0.689) was comparatively low. MedGemma 4B showed relatively low performance, with an overall accuracy of 0.677. Claude Sonnet 4.6 and Gemini 3 Flash had the highest sensitivities among the models (0.991 and 0.972, respectively) but low specificities (0.038 and 0.201, respectively), resulting in excessive false positives. In the characterization task, ChatGPT 5.2 consistently achieved the highest performance among the models, with an accuracy of 0.758 and a weighted F1 score of 0.745. DeepSeek-VL2 achieved high specificity but very low sensitivity for malignancy (0.714 and 0.022, respectively). Gemini 3 Flash provided high sensitivity for malignancy (0.711) but low overall accuracy.

CONCLUSION

Multimodal LLMs demonstrated heterogeneous performance in the evaluation of bone lesions on plain radiographs, with substantial differences across models and tasks. Although some models achieved high accuracy in lesion detection and overall classification, performance was inconsistent across tasks, particularly in identifying malignant lesions and balancing sensitivity and specificity. These findings suggest that, despite their potential, current multimodal LLMs are not yet sufficiently reliable for diagnostic use in orthopedic oncology and should be considered investigational until further development and validation.

CLINICAL SIGNIFICANCE

Multimodal LLMs currently lack the diagnostic reliability required for bone lesion assessment, often exhibiting excessive false positives or failing to detect malignancy. Although generalist models show promise, expert radiologist oversight remains essential to ensure patient safety and oncologic accuracy in musculoskeletal imaging.

KEYWORDS

Bone neoplasms, large language models, artificial intelligence, radiography

Accurate diagnosis of bone lesions is a critical step in orthopedic oncology because management strategies can vary from nonoperative monitoring to surgical removal.^{1,2} X-rays remain the primary imaging modality for assessing tumors and tumor-like lesions in bones.¹⁻³ Diagnostic decision-making, particularly the differentiation of benign tumors from aggressive bone tumor subtypes, relies heavily on the interpretation of radiologic features and requires substantial expertise and experience.^{4,5} The complexity of bone tumor classification, along with the rarity and overlapping characteristics of subtypes, can create challenges in diagnosing bone lesions.

Recent advances in artificial intelligence (AI), especially the rise of large language models (LLMs), have opened new possibilities across various medical fields, including radiology.⁶⁻⁸ AI-based tools have been developed to improve musculoskeletal radiology processes, varying from the automatic detection of pathologies on magnetic resonance images (MRIs) to the detection and classification of bone lesions on radiographs. Traditionally, these image-centric classification tasks have relied heavily on convolutional neural network (CNN)-based architectures, which excel at extracting hierarchical structural features from medical images.⁹⁻¹¹ However, the introduction of LLMs has brought a new perspective to image analysis. Initial uses of unimodal LLMs focused on text-based tasks; however, recent advances in multimodal LLMs enable the interpretation of both text and visual data.^{12,13} Early studies using these multimodal LLMs in musculoskeletal radiology for tasks such as anatomical question answering or clinical decision support have shown promising but

inconsistent results compared with human experts.^{14,15} However, evidence regarding their performance in image-centric diagnostic tasks, particularly in musculoskeletal oncology, remains limited. Importantly, it is unclear how these general-purpose models perform when tasked with interpreting plain radiographs for complex diagnostic tasks, such as bone lesion detection and subtype classification, without task-specific fine-tuning or in-context learning.

Against this background, the present study aims to evaluate five available multimodal LLMs (ChatGPT 5.2, Gemini 3 Flash, MedGemma 4B, Claude Sonnet 4.6, and DeepSeek-VL2) for the assessment of bone lesions on plain radiographs. Specifically, we compare their performance across complex tasks: lesion detection and lesion characterization, using a publicly available reference dataset. To our knowledge, this is the first study to extensively cover different subtypes of bone tumors while evaluating multimodal LLM performance. Rather than establishing clinical readiness, this study seeks to define a baseline evaluation for current multimodal LLM capabilities and to identify limitations that must be addressed before such models can be considered for clinical or decision-support applications in orthopedic oncology.

Methods

The retrospective study was approved by the Ethics Committee of İzmir Bakırçay University on October 15, 2025 (decision number: 2476). Informed consent was waived by the committee due to the study's design. This study was conducted and reported in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis-LLM guidelines.¹⁶

Study design

Five independent LLMs—ChatGPT 5.2, Gemini 3 Flash, MedGemma 4B, Claude Sonnet 4.6, and DeepSeek-VL2—were evaluated across two diagnostic tasks: lesion detection (Task 1) and lesion characterization (Task 2). For each task, model predictions were compared with reference-standard annotations. For the classification of bone tumors in this study, the label definitions and expert annotations from the publicly available Bone Tumor X-ray Radiograph Dataset (BTXRD) introduced by Yao et al. were adopted.¹⁷ All labels in the BTXRD dataset were assigned based on histopathological diagnoses recorded in the source data or consensus reviews by

expert radiologists—particularly for benign tumors—during dataset curation. Figure 1 visualizes the study flowchart.

Dataset preparation

This retrospective study utilized the publicly available BTXRD dataset.¹⁷ The radiographs within the BTXRD cohort were acquired from multiple institutions. The dataset comprises radiographic cohorts of malignant bone tumors, benign bone tumors, and normal bone cases. Each case was evaluated using a single plain radiograph; no additional projections, serial images, or cross-sectional imaging were provided to the models.

All images were obtained directly from the dataset and provided in 8-bit grayscale JPEG format; therefore, no additional resizing or geometric modifications were applied beyond the dataset's original release. All data were fully anonymized before inclusion in the dataset, and no patient-identifiable information was accessible to the investigators.

Model prompting

All models were evaluated using a zero-shot prompting strategy with a fixed, optimized prompt. Before the main analysis, a pilot study for prompt refinement was conducted using a stratified random subset of 100 radiographs, fully independent of the final test set. All models processed this identical subset to ensure comparability. Prompt consistency was assessed by performing 10 repeated inference runs per case, and the final prompt was selected based on achieving complete structural adherence to the predefined output schema across all runs. These 100 cases were strictly excluded from the final performance analysis to prevent data leakage. The final study population consisted of 3,646 cases.

The finalized prompt instructed the models to analyze each radiograph using only the provided image, patient age, and patient sex to determine lesion presence and characterization.

To ensure reproducibility, the evaluated models were divided into two deployment strategies based on their architectures. For the cloud-based proprietary models, specific application programming interface (API) snapshots were used to prevent inconsistencies caused by model updates over time. Specifically, OpenAI's GPT-5.2-2025-12-11, Google's Gemini-3-Flash-preview (snapshot dated January 21, 2026), and Anthropic's Claude-Sonnet-4-6 (snapshot dated March

Main points

- Current multimodal large language models (LLMs) demonstrate variable and overall limited performance in bone lesion assessment, with inconsistent differentiation between benign and malignant lesions that prevents reliable independent diagnostic use.
- Current multimodal LLMs exhibit different diagnostic profiles: For example, Claude Sonnet 4.6 and Gemini 3 Flash showed high sensitivity but suffered from very high false-positive rates, whereas other models failed to reliably identify malignant features.
- Despite being tailored for medical applications, the domain-specific model MedGemma 4B currently exhibits critical diagnostic errors, underscoring the need for further development.

13, 2026) were employed. All inferences were performed between December 15, 2025, and March 13, 2026, via these APIs, with each radiograph processed independently. In contrast, the open-weight models (MedGemma 4B and DeepSeek-VL2) were downloaded and executed locally. The inference process was fully automated using a custom pipeline, through which images were analyzed either via APIs or local model deployments, and outputs were systematically recorded for subsequent analysis. The zero-shot inference pipelines for these models, as well as the subsequent statistical data analysis, were executed locally on a workstation equipped with an NVIDIA RTX 3080 GPU (16 GB VRAM) (NVIDIA, Santa Clara, CA, USA) in a Linux environment using Python 3.10 and the Hugging Face Transformers library. No quantization techniques were applied to preserve full model performance. Furthermore, no fine-tuning, in-context examples, or iterative feedback was used. The temperature was fixed at 0.00, and the maximum number of generated tokens was set to 300. The finalized prompt and inference parameters were applied uniformly across all cases.

The final prompt was as follows:

"This task is for research purposes only. You are not responsible for any medical decision.

You are a radiologist with expertise in musculoskeletal imaging.

Analyze the following X-ray image. The patient is a [AGE]-year-old [GENDER].

Answer the questions below.

Do not include explanations or any additional text.

Rules:

- Q1 corresponds to the question: "Is there any bone lesion present?"

- Q2 corresponds to the question: "If a lesion is present, classify its most likely nature."

- If Q1_bone_lesion_present is "No", Q2_lesion_nature must be "Not applicable".

- If Q1_bone_lesion_present is "Yes", Q2_lesion_nature must be either "benign" or "malignant".

Return the output strictly in the specified format.

Q1_bone_lesion_present: Yes | No

Q2_lesion_nature: benign | malignant | Not applicable"

Statistical analysis

All analyses were performed in Python (v3.10) using the scikit-learn and statsmodels libraries. Descriptive statistics were used to characterize the study cohort. Continuous variables were summarized using the median and interquartile range (IQR), and categorical variables were reported as frequencies and percentages.

All analyses were conducted separately for each task. For both tasks, ground-truth labels served as the reference standard. Because all models were evaluated on the same radiographs, statistical analyses were performed using paired methods.

For Task 1 (lesion detection), performance was assessed as a binary classification problem. Accuracy, sensitivity, specificity, F1 score, and Cohen's kappa (κ) were calculated. Pairwise differences in accuracy were evaluated using the McNemar test. Because the F1 score is a dataset-level metric, pairwise differences in F1 were assessed using paired stratified bootstrap resampling, with the paired difference ($\Delta F1$) reported with 95% confidence intervals (CIs).

For Task 2 (lesion characterization), a three-class analysis was performed across the categories of lesion absent, benign, and malignant. Overall accuracy, weighted F1 score, and Cohen's κ were calculated. In addition, class-wise performance metrics—including precision, sensitivity, F1 score, and support—were computed for each category. Overall model differences were assessed using Cochran's Q test, with outputs converted to correct versus incorrect classifications. Significant results were followed by post hoc McNemar tests, and differences in F1 score ($\Delta F1$) were estimated using paired bootstrap resampling.

Bootstrap CIs were derived from 2,000 stratified resamples. The Holm-Bonferroni correction was applied for multiple comparisons, and a two-sided adjusted P value < 0.05 was considered statistically significant. Kappa values were interpreted according to Landis and Koch: slight ≤ 0.20 ; fair: 0.21–0.40; moderate: 0.41–0.60; substantial: 0.61–0.80; and almost perfect agreement ≥ 0.80 .¹⁸

Results

Study cohort

After the exclusion of cases used for prompt optimization, a total of 3,646 cases were included in the analysis. The median age of the study population was 35.0 years

(IQR: 16.0–53.0). Of the total cohort, 2,041 (56.0%) were men and 1,605 (44.0%) were women. According to the reference standard, 1,817 cases (49.8%) were classified as "tumor present" and 1,829 cases (50.2%) as "no tumor." The tumor-present cases included 1,589 benign tumors (43.6%) and 228 malignant tumors (6.3%).

Lesions were most frequently located in the tibia ($n = 617$, 16.9%) and femur ($n = 570$, 15.6%), together accounting for nearly one-third of all cases, followed by the fibula ($n = 256$, 7.0%) and humerus ($n = 233$, 6.4%). Distal extremity bones, such as the hand ($n = 95$, 2.6%), foot ($n = 90$, 2.5%), radius ($n = 71$, 1.9%), and ulna ($n = 62$, 1.7%), were less commonly involved. Joint-based locations were rare overall; the knee, hip, ankle, elbow, wrist, and shoulder joints collectively represented a small proportion of cases ($< 2\%$). Table 1 summarizes the study cohort characteristics.

Task 1: lesion detection

Task 1 evaluated tumor detection across the five models. ChatGPT 5.2 demonstrated the highest overall performance among the models, achieving an accuracy of 0.803 (95% CI: 0.791–0.815) with moderate agreement with the reference standard (κ : 0.605, 95% CI: 0.582–0.630). Sensitivity was 0.689 (95% CI: 0.669–0.710), whereas specificity reached 0.916 (95% CI: 0.903–0.927), resulting in an F1 score of 0.777 (95% CI: 0.762–0.792).

MedGemma 4B showed relatively low performance, with an accuracy of 0.677 (95% CI: 0.662–0.691) and fair agreement (κ : 0.356, 95% CI: 0.325–0.384) for lesion detection. The model achieved a sensitivity of 0.823 (95% CI: 0.804–0.840) and a specificity of 0.533 (95% CI: 0.510–0.556), corresponding to an F1 score of 0.718 (95% CI: 0.705–0.729).

Gemini 3 Flash exhibited an accuracy of 0.585 (95% CI: 0.575–0.595) and only slight agreement beyond chance (κ : 0.172; 95%

Table 1. Study cohort characteristics

Characteristic	Value
Total cases, n	3,646
Age, median (IQR) (years)	35.0 (16.0–53.0)
Male, n (%)	2,041 (56.0%)
Female, n (%)	1,605 (44.0%)
Normal cases, n (%)	1,829 (50.2%)
Cases with tumors, n (%)	1,817 (49.8%)
Benign tumors, n (%)	1,589 (43.6%)
Malignant tumors, n (%)	228 (6.3%)

IQR, interquartile range.

CI: 0.153–0.192). Although sensitivity was very high at 0.972 (95% CI: 0.965–0.979), specificity was markedly reduced at 0.201 (95% CI: 0.182–0.219), yielding an F1 score of 0.700 (95% CI: 0.694–0.706).

Claude Sonnet 4.6 showed an accuracy of 0.513 (95% CI: 0.508–0.518) and slight agreement (κ : 0.028, 95% CI: 0.019–0.038). The model achieved a sensitivity of 0.991 (95% CI: 0.987–0.995) and a specificity of 0.038 (95% CI: 0.030–0.047), corresponding to an F1 score of 0.670 (95% CI: 0.667–0.672).

DeepSeek-VL2 demonstrated an accuracy of 0.577 (95% CI: 0.560–0.592) and slight agreement (κ : 0.153, 95% CI: 0.121–0.184). The model achieved a sensitivity of 0.620 (95% CI: 0.598–0.642) and a specificity of 0.533 (95% CI: 0.510–0.556), corresponding to an F1 score of 0.593 (95% CI: 0.577–0.609).

For Task 1, there were statistically significant differences in accuracy rates among the models. ChatGPT 5.2 was statistically more accurate than Gemini 3 Flash, MedGemma 4B, Claude Sonnet 4.6, and DeepSeek-VL2 ($p < 0.001$ for all comparisons). Furthermore, MedGemma 4B statistically outperformed Gemini 3 Flash, DeepSeek-VL2, and Claude Sonnet 4.6 ($p < 0.001$ for all comparisons). Table 2 summarizes all performance metrics for the models. Confusion matrices illustrating true-positive, false-positive, true-negative, and false-negative counts for the lesion detection task are shown in Figure 2.

Task 2: lesion characterization (benign vs. malignant)

In Task 2, we initially analyzed all class labels, including the no-lesion, benign, and malignant categories.

For this three-class classification task, ChatGPT 5.2 consistently achieved the highest performance among the models. The model reached an accuracy of 0.758 (95%

CI: 0.745–0.770), a weighted F1 score of 0.745 (95% CI: 0.732–0.758), and moderate agreement with the reference standard (κ : 0.546, 95% CI: 0.522–0.569). MedGemma 4B demonstrated lower performance than ChatGPT 5.2, with an accuracy of 0.621 (95% CI: 0.607–0.635), a weighted F1 score of 0.608 (95% CI: 0.593–0.623), and fair agreement (κ :

0.307, 95% CI: 0.281–0.332). DeepSeek-VL2 followed with an accuracy of 0.529 (95% CI: 0.512–0.544), a weighted F1 score of 0.514 (95% CI: 0.499–0.530), and slight agreement (κ : 0.118, 95% CI: 0.089–0.149). Gemini 3 Flash showed an accuracy of 0.484 (95% CI: 0.470–0.498), a weighted F1 score of 0.457 (95% CI: 0.441–0.473), and fair agreement

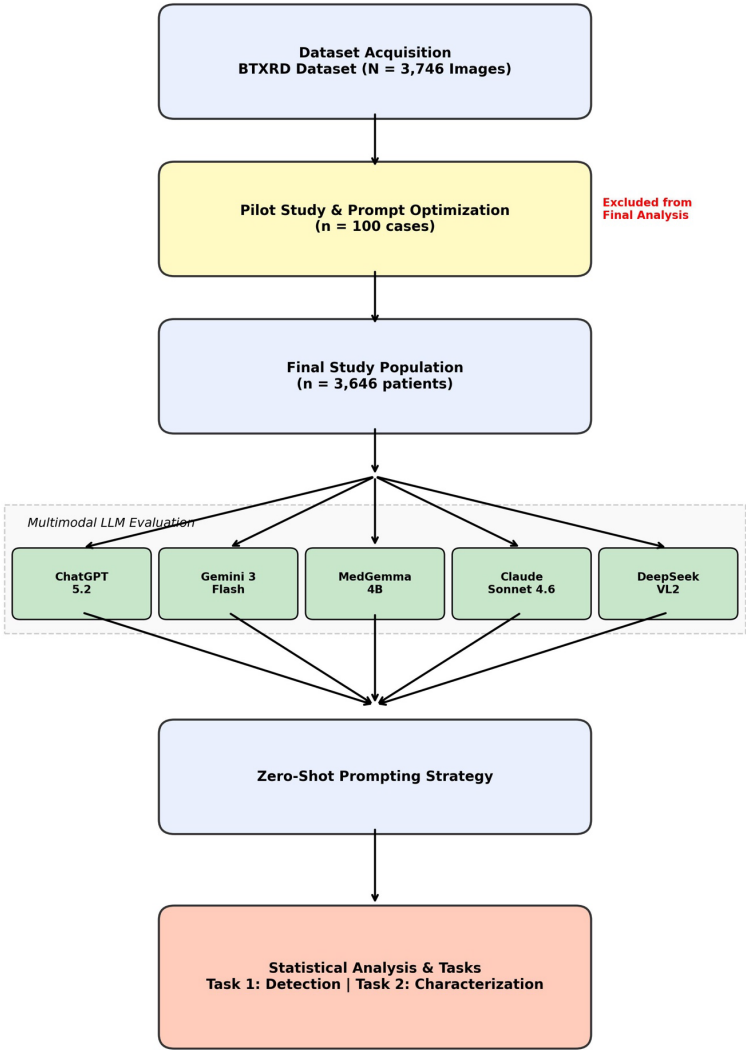


Figure 1. Flowchart of the study. BTXRD, Bone Tumor X-ray Radiograph Dataset; LLM, large language model.

Table 2. Performance metrics for lesion detection across the models					
Model	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	F1 score (95% CI)	Cohen's kappa (95% CI)
ChatGPT 5.2	0.803 (0.791–0.815)	0.689 (0.669–0.710)	0.916 (0.903–0.927)	0.777 (0.762–0.792)	0.605 (0.582–0.630)
Gemini 3 Flash	0.585 (0.575–0.595)	0.972 (0.965–0.979)	0.201 (0.182–0.219)	0.700 (0.694–0.706)	0.172 (0.153–0.192)
MedGemma 4B	0.677 (0.662–0.691)	0.823 (0.804–0.840)	0.533 (0.510–0.556)	0.718 (0.705–0.729)	0.356 (0.325–0.384)
Claude Sonnet 4.6	0.513 (0.508–0.518)	0.991 (0.987–0.995)	0.038 (0.030–0.047)	0.670 (0.667–0.672)	0.028 (0.019–0.038)
DeepSeek-VL2	0.577 (0.560–0.592)	0.620 (0.598–0.642)	0.533 (0.510–0.556)	0.593 (0.577–0.609)	0.153 (0.121–0.184)

CI, confidence interval.

(κ : 0.204, 95% CI: 0.186–0.223). Claude Sonnet 4.6 demonstrated the lowest performance, with an accuracy of 0.303 (95% CI: 0.291–0.332), a weighted F1 score of 0.268 (95% CI: 0.256–0.280), and slight agreement (κ : 0.068, 95% CI: 0.055–0.081). The distribution of predictions across the three categories (no lesion, benign, and malignant) for each model is visualized in the confusion matrices presented in Figure 3. Table 3 summarizes the performance metrics of all models across the three class-subtyping tasks.

For Task 2, comparisons based on accuracy rates revealed a statistically significant difference among the five multimodal LLMs

($P < 0.001$). In pairwise analyses, ChatGPT 5.2 demonstrated significantly better performance than Gemini 3 Flash, MedGemma 4B, Claude Sonnet 4.6, and DeepSeek-VL2 ($P < 0.001$ for all comparisons). Additionally, MedGemma 4B showed significantly higher accuracy than DeepSeek-VL2, Gemini 3 Flash, and Claude Sonnet 4.6 ($P < 0.001$). Class-wise performance metrics for Task 2 are summarized in Table 4. For the no-lesion class, ChatGPT 5.2 achieved a sensitivity of 0.916, precision of 0.748, and F1 score of 0.824, whereas MedGemma 4B, DeepSeek-VL2, Gemini 3 Flash, and Claude Sonnet 4.6 showed sensitivities of 0.533, 0.533, 0.201, and 0.038, respectively. In the benign class,

sensitivities were 0.796 for MedGemma 4B, 0.778 for Gemini 3 Flash, 0.642 for ChatGPT 5.2, 0.596 for DeepSeek-VL2, and 0.530 for Claude Sonnet 4.6, with corresponding F1 scores of 0.652, 0.630, 0.705, 0.532, and 0.504. For malignant lesions, Claude Sonnet 4.6 yielded the highest sensitivity (0.851), followed by Gemini 3 Flash (0.711), whereas ChatGPT 5.2, MedGemma 4B, and DeepSeek-VL2 showed lower sensitivities of 0.289, 0.114, and 0.022, respectively. Malignant F1 scores were 0.407 for ChatGPT 5.2, 0.289 for Gemini 3 Flash, 0.191 for Claude Sonnet 4.6, 0.180 for MedGemma 4B, and 0.043 for DeepSeek-VL2.

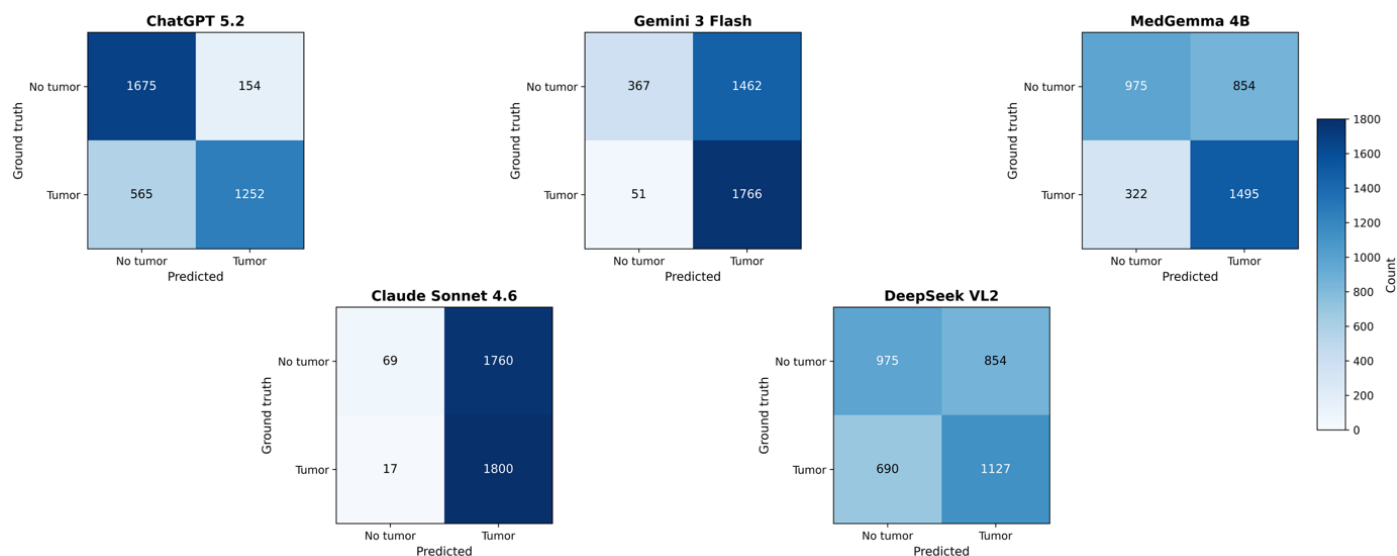


Figure 2. Confusion matrices for Task 1 (lesion detection). The panels display the performance of LLMs in detecting the presence of bone lesions. LLM, large language model.

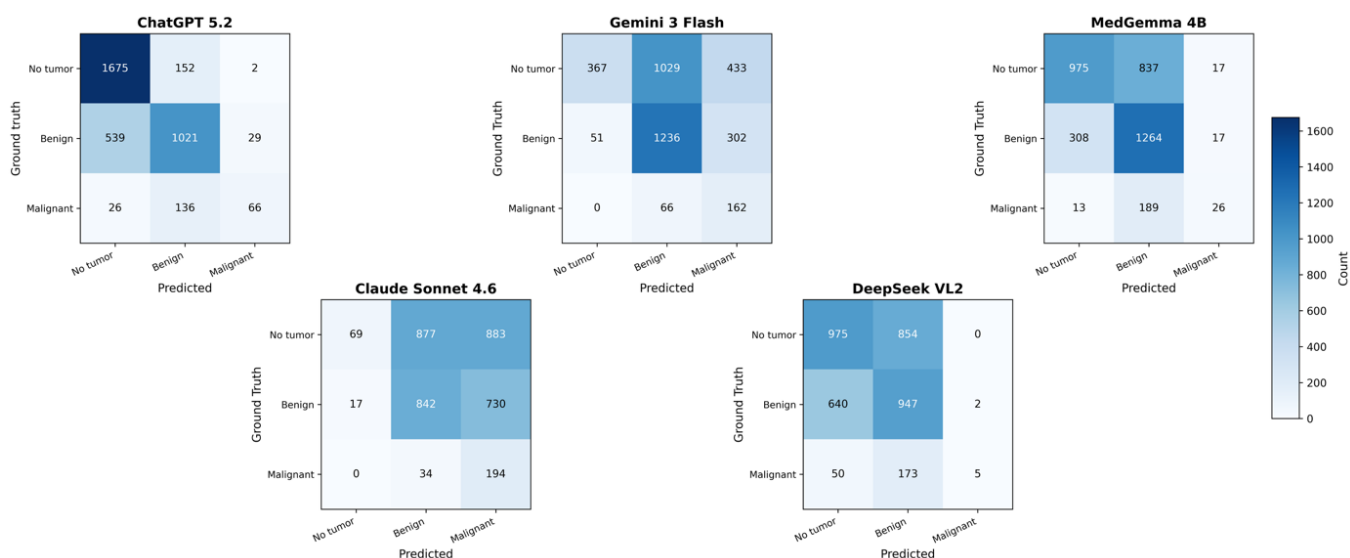


Figure 3. Confusion matrices for the three-class classification task. The matrices illustrate the classification performance for distinguishing between “no lesion,” “benign,” and “malignant” categories across the five evaluated multimodal LLMs. LLM, large language model.

Model	Accuracy (95% CI)	Weighted F1 score (95% CI)	Cohen's kappa (95% CI)
ChatGPT 5.2	0.758 (0.745–0.770)	0.745 (0.732–0.758)	0.546 (0.522–0.569)
Gemini 3 Flash	0.484 (0.470–0.498)	0.457 (0.441–0.473)	0.204 (0.186–0.223)
MedGemma 4B	0.621 (0.607–0.635)	0.608 (0.593–0.623)	0.307 (0.281–0.332)
Claude Sonnet 4.6	0.303 (0.291–0.332)	0.268 (0.256–0.280)	0.068 (0.055–0.081)
DeepSeek-VL2	0.529 (0.512–0.544)	0.514 (0.499–0.530)	0.118 (0.089–0.149)

CI, confidence interval.

Model	Class	Precision	Sensitivity	F1 score	Number of cases
MedGemma 4B	No lesion	0.752	0.533	0.624	1.829
	Benign	0.552	0.796	0.652	1.589
	Malignant	0.433	0.114	0.180	228
Gemini 3 Flash	No lesion	0.878	0.201	0.327	1.829
	Benign	0.530	0.778	0.630	1.589
	Malignant	0.181	0.711	0.289	228
ChatGPT 5.2	No lesion	0.748	0.916	0.824	1.829
	Benign	0.780	0.642	0.705	1.589
	Malignant	0.680	0.289	0.407	228
Claude Sonnet 4.6	No lesion	0.802	0.038	0.072	1.829
	Benign	0.480	0.530	0.504	1.589
	Malignant	0.107	0.851	0.191	228
DeepSeek-VL2	No lesion	0.586	0.533	0.558	1.829
	Benign	0.480	0.596	0.532	1.589
	Malignant	0.714	0.022	0.043	228

Discussion

In this head-to-head exploratory study of five multimodal LLMs, we observed that although all models demonstrated the capacity to generate diagnostic outputs and ChatGPT 5.2 provided the most robust baseline performance, the discordance between models highlights the volatility of zero-shot inference in complex diagnostic tasks. Although we initially hypothesized that medical pretraining might confer an advantage, ChatGPT 5.2 outperformed the medical-specific MedGemma 4B, whereas the remaining general-purpose models did not demonstrate superior performance. Notably, MedGemma 4B exceeded several other general-purpose models, suggesting that both model scale and domain-specific training contribute to performance and require further optimization.

Although ChatGPT 5.2 achieved moderate agreement (κ : 0.605) and 80.3% accuracy in lesion detection, its zero-shot capabilities remain inferior to benchmarks established by supervised deep learning algorithms. Across prior studies, diagnostic performance varied by both model architecture and imaging modality. High accuracy and area under the curve values exceeding 0.79 were reported for X-ray–based deep learning approaches, particularly when combined with clinical or demographic features. Furthermore, computed tomography or MRI-based models achieved similar discrimination with or without added patient-level information.^{11,19–23}

It is also noteworthy that while prior applications of LLMs in orthopedic oncology have predominantly focused on textual report analysis or data extraction, this study represents one of the first evaluations of their direct visual capabilities on radiographs

in this field.^{24,25} However, despite the nascent stage of this technology, the observed performance deficits suggest considerable limitations in current architectures. Consequently, these findings serve as a cautionary signal for radiologists, clinicians, and patients, indicating that general-purpose multimodal LLMs are not currently reliable for diagnostic decision-making and should be approached with skepticism until rigorous validation is achieved.

An analysis of model performance reveals distinct diagnostic profiles characterized by an imbalance between sensitivity and specificity, potentially limiting their clinical integration. Two divergent interpretative patterns were observed. First, Gemini 3 Flash and Claude Sonnet 4.6 demonstrated a high-sensitivity, low-specificity profile in lesion detection, resulting in a high rate of false positives. This tendency toward

over-identification in Gemini models is consistent with reported performance in other clinical domains, such as emergency triage and cervical cytology.^{26,27} In contrast, MedGemma 4B systematically misclassified malignant lesions as benign during the characterization task. Although this model maintained high specificity, its low sensitivity for malignancy poses a substantial diagnostic risk, as malignant lesions are frequently categorized as nonaggressive. This discrepancy between models—ranging from excessive false positives to the failure to identify malignant features—underscores that current zero-shot multimodal architectures lack the calibrated decision thresholds necessary for autonomous oncologic risk stratification.

Despite having lower performance than dedicated task-specific AI models, multimodal LLMs may offer complementary value in radiology in the future. Unlike conventional deep learning systems that require local deployment and integration into imaging infrastructure, multimodal LLMs are broadly accessible through web-based interfaces, allowing for use across institutions with varying technical resources, including resource-limited settings.²⁸ In addition, recent research suggests that the most promising role for LLMs might not be as standalone diagnostic classifiers, but as components of agentic systems in which the LLM functions as a reasoning engine coordinating specialized tools, such as task-specific CNN pipelines.²⁹⁻³¹ For instance, an agentic approach in orthopedic oncology could structure the diagnostic process through a series of specialized, interoperable models. In the first stage, a CNN-based model performs lesion detection and segmentation. In the second stage, a separate model accesses relevant clinical data—including prior malignancy history and laboratory findings—and integrates these with the imaging outputs to generate a differential diagnosis and preliminary assessment. In the final stage, a third model synthesizes all available information into a structured radiology report and presents it to the radiologist for review and validation.

Our findings indicate that this accessibility currently comes with unacceptable diagnostic risks. The inability of the highest-performing models to reliably detect malignancy or exclude pathology renders them unsuitable for diagnostic decision-making. Consequently, these tools should currently be viewed as experimental rather than as diagnostic alternatives. Radiologists, clinicians, and patients must be cautioned that the fluent textual output of these models does not equate to

radiologic competence; thus, human-in-the-loop verification remains non-negotiable.

Our study has limitations inherent to its exploratory design. First, the use of a zero-shot prompting strategy—without providing reference exemplars (few-shot learning) or chain-of-thought reasoning—may have underestimated the models' true potential. Second, the reliance on single-view radiographs contrasts with clinical practice, which typically incorporates orthogonal views and cross-sectional imaging. Third, the nature of commercial closed-source models precludes an analysis of the specific visual features driving their decisions due to their proprietary, non-interpretable architectures. Furthermore, model outputs may vary over time with undocumented updates.³² Fourth, the direct comparison between massive, proprietary frontier models and the significantly smaller MedGemma 4B and DeepSeek-VL2 models is intrinsically unbalanced. The observed performance differences are likely driven by the vast discrepancy in parameter counts rather than strictly by the difference between general and medical-specific training. Additionally, the MedGemma 27B model was not included due to the lack of multimodal capability at the time of study design, which limits our ability to assess performance scaling within the same model family.

Future research should prioritize prompt engineering strategies to mitigate the observed biases and investigate whether fine-tuning methods can realign these generalist models with the sensitivity requirements of orthopedic oncology. In addition, future studies should evaluate larger multimodal variants within the same model family; higher-parameter models, such as the MedGemma 27B version, may demonstrate improved performance and provide further insight into scaling effects. Finally, prospective reader studies evaluating radiologist performance with and without LLM assistance are needed to determine clinical utility while accounting for ongoing model updates and version changes. Despite these limitations, our study provides key insights into the use of multimodal LLMs in orthopedic oncology and highlights the current limitations of these models within clinical decision algorithms, such as the interpretation of direct radiographs in routine practice.

This study establishes a baseline for the zero-shot capabilities of multimodal LLMs in orthopedic oncology. Our findings demonstrate heterogeneous performance across models: ChatGPT 5.2 achieved the highest

overall results, whereas the domain-specific MedGemma 4B outperformed several other general-purpose models. However, despite these results, all evaluated models exhibited critical performance deficits compared with established supervised deep-learning benchmarks. The distinct operating profiles observed—ranging from high sensitivity and low specificity to the dangerous overcalling of benign lesions and low malignancy detection—indicate that current multimodal LLMs lack the calibrated decision boundaries required for safe oncologic decision support. Consequently, although these models offer potential for future workflow integration and increased accessibility, they are not currently suitable for independent diagnostic decision-making. To ensure oncologic safety before clinical adoption, future development should pivot from utilizing LLMs as standalone classifiers toward integrating them into agentic workflows.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Miller TT. Bone tumors and tumorlike conditions: analysis with conventional radiography. *Radiology*. 2008;246(3):662-674. [\[Crossref\]](#)
2. Lalam R, Bloem JL, Noebauer-Huhmann IM, et al. ESSR consensus document for detection, characterization, and referral pathway for tumors and tumorlike lesions of bone. *Semin Musculoskelet Radiol*. 2017;21(5):630-647. [\[Crossref\]](#)
3. Mehta K, McBee MP, Mihal DC, England EB. Radiographic analysis of bone tumors: a systematic approach. *Semin Roentgenol*. 2017;52(4):194-208. [\[Crossref\]](#)
4. Chang CY, Garner HW, Ahlawat S, et al. Society of Skeletal Radiology- white paper. Guidelines for the diagnostic management of incidental solitary bone lesions on CT and MRI in adults: bone reporting and data system (Bone-RADS). *Skeletal Radiol*. 2022;51(9):1743-1764. [\[Crossref\]](#)
5. Park C, Azhideh A, Pooyan A, et al. Diagnostic performance and inter-reader reliability of bone reporting and data system (Bone-RADS) on computed tomography. *Skeletal Radiol*. 2025;54:209-217. [\[Crossref\]](#)
6. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*. 2024;30(2):80-90. [\[Crossref\]](#)

7. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*. 2024;310(1):e232756. [\[Crossref\]](#)
8. Salbas A, Buyuktoka RE. Performance of large language models in recognizing brain MRI sequences: a comparative analysis of ChatGPT-4o, Claude 4 Opus, and Gemini 2.5 Pro. *Diagnostics (Basel)*. 2025;15(15):1919. [\[Crossref\]](#)
9. von Schackey CE, Wilhelm NJ, Schäfer VS, et al. Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology*. 2021;301(2):398-406. [\[Crossref\]](#)
10. Engin O, Durmaz Engin C, Çilengir AH, Dirim Mete B. Detection and classification of supraspinatus pathologies on shoulder magnetic resonance images using a code-free deep learning application. *Asia Pac J Sports Med Arthrosc Rehabil Technol*. 2025;42:1-7. [\[Crossref\]](#)
11. Li J, Li S, Li X, et al. Primary bone tumor detection and classification in full-field bone radiographs via YOLO deep learning model. *Eur Radiol*. 2023;33(6):4237-4248. [\[Crossref\]](#)
12. OpenAI Team. Introducing GPT-5. Accessed January 2025. [\[Crossref\]](#)
13. Comanici G, Bieber E, Schaekermann M, et al. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv*. 2025;arXiv:2507.06261. [\[Crossref\]](#)
14. Salbas A, Baysan EK. Assessment of large language models in musculoskeletal radiological anatomy: a comparative study with radiologists. *Jt Dis Relat Surg*. 2026;37(1):190-199. [\[Crossref\]](#)
15. Zamora T, Salas P, Zuñiga S, Botello E, Andia ME. Generative artificial intelligence, large language models and ChatGPT in musculoskeletal oncology: current applications and future potential. *J Clin Orthop Trauma*. 2025;69:103161. [\[Crossref\]](#)
16. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. 2025;31(1):60-69. [\[Crossref\]](#)
17. Yao S, Huang Y, Wang X, et al. A radiograph dataset for the classification, localization, and segmentation of primary bone tumors. *Sci Data*. 2025;12(1):88. [\[Crossref\]](#)
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174. [\[Crossref\]](#)
19. von Schackey CE, Wilhelm NJ, Schäfer VS, et al. Development and evaluation of machine learning models based on X-ray radiomics for the classification and differentiation of malignant and benign bone tumors. *Eur Radiol*. 2022;32(9):6247-6257. [\[Crossref\]](#)
20. Liu R, Pan D, Xu Y, et al. A deep learning-machine learning fusion approach for the classification of benign, malignant, and intermediate bone tumors. *Eur Radiol*. 2022;32(2):1371-1383. [\[Crossref\]](#)
21. He Y, Pan I, Bao B, et al. Deep learning-based classification of primary bone tumors on radiographs: a preliminary study. *EBioMedicine*. 2020;62:103121. [\[Crossref\]](#)
22. Eweje FR, Bao B, Wu J, et al. Deep learning for classification of bone lesions on routine MRI. *EBioMedicine*. 2021;68:103402. [\[Crossref\]](#)
23. Yildiz Potter I, Yeritsyan D, Mahar S, et al. Automated bone tumor segmentation and classification as benign or malignant using computed tomographic imaging. *J Digit Imaging*. 2023;36(3):869-878. [\[Crossref\]](#)
24. Kanemaru N, Yasaka K, Fujita N, Kanzawa J, Abe O. The fine-tuned large language model for extracting the progressive bone metastasis from unstructured radiology reports. *J Imaging Inform Med*. 2025;38(2):865-872. [\[Crossref\]](#)
25. Yang F, Yan D, Wang Z. Large-scale assessment of ChatGPT's performance in benign and malignant bone tumors imaging report diagnosis and its potential for clinical applications. *J Bone Oncol*. 2024;44:100525. [\[Crossref\]](#)
26. Lee S, Jung S, Park JH, Cho H, Moon S, Ahn S. Performance of ChatGPT, Gemini and DeepSeek for non-critical triage support using real-world conversations in emergency department. *BMC Emerg Med*. 2025;25(1):176. [\[Crossref\]](#)
27. Laohawetwanit T, Apornvirat S, Asaturova A, Li H, Lami K, Bychkov A. Evaluation of general-purpose large language models as diagnostic support tools in cervical cytology. *Pathol Res Pract*. 2025;274:156159. [\[Crossref\]](#)
28. Stogiannos N, Cuocolo R, Akinci D'Antonoli T, et al. Recognising errors in AI implementation in radiology: a narrative review. *Eur J Radiol*. 2025;191:112311. [\[Crossref\]](#)
29. Tzanis E, Adams LC, Akinci D'Antonoli T, et al. Agentic systems in radiology: principles, opportunities, privacy risks, regulation, and sustainability concerns. *Diagn Interv Imaging*. 2026;107(1):7-16. [\[Crossref\]](#)
30. Tzanis E, Klontzas ME. ReclAI: a multi-agent framework for degradation-aware performance tuning of medical imaging AI. *arXiv*. 2025. [\[Crossref\]](#)
31. Tzanis E, Klontzas ME. mAIstro: an open-source multi-agent system for automated end-to-end development of radiomics and deep learning models for medical imaging. *Eur J Radiol Artif Intell*. 2025;4:100044. [\[Crossref\]](#)
32. Gu Y, Fu J, Liu X, et al. The illusion of readiness: stress testing large frontier models on multimodal medical benchmarks. *arXiv*. 2025;arXiv:2509.18234. [\[Crossref\]](#)



Quality of reporting of studies on artificial intelligence in radiology: what is the current state of the field?

Ahmed Maiter¹⁻³

¹Sheffield Teaching Hospitals NHS Foundation Trust,
Department of Radiology, Sheffield, United Kingdom

²NIHR Sheffield Biomedical Research Centre,
Sheffield, United Kingdom

³University of Sheffield, School of Medicine and
Population Health, Sheffield, United Kingdom

High-quality research reporting is of paramount importance. Transparent descriptions of study design, methods, and limitations are essential for reproducibility and underpin trust in results. Furthermore, consistent reporting of methods and outcomes is necessary for meaningful comparisons across similar studies and effective evidence synthesis. In contrast, incomplete reporting limits the interpretability of studies, increases the risk of misinterpreting findings, and raises the likelihood of research waste.

Promoting, implementing, and maintaining high-quality reporting practices in the field of artificial intelligence (AI) in radiology can be challenging. Numerous factors contribute to model performance and require detailed descriptions, particularly the data used for model training and testing. Seemingly minor variations in data, such as scanner vendor or study participant age distribution, may result in dramatic shifts in performance, making transparent reporting essential.

Additionally, the ability to explain how AI systems function and reach decisions is valuable, yet often not straightforward. This so-called “black box” problem is exacerbated by the development of increasingly complex models and the proliferation of proprietary commercial AI devices. Moreover, AI studies in radiology encompass both clinical and technical dimensions, each requiring transparent reporting. This duality can be a hurdle for research teams whose expertise is predominantly rooted in one of these domains.

Over the past two decades, a variety of reporting guidelines have been developed for different types of medical research.^{1,2} These guidelines specify and standardize the components of a study that should be reported by manuscript authors. Their overarching goals are to improve the transparency and reproducibility of research, facilitate peer review, and support the comparison of findings across publications. Several of these guidelines are now well established and are required by journals to be completed and submitted alongside manuscripts.

The Checklist for Artificial Intelligence in Medical Imaging (CLAIM) was the first reporting guideline specifically developed for studies on AI in medical imaging. Initially published in 2020 and updated in 2024, CLAIM takes a broad approach, covering general information applicable to most studies in the field.³⁻⁵ Since its introduction, a growing number of systematic and other literature reviews have used CLAIM to evaluate publications, yielding insights into deficiencies in reporting quality in specific areas of AI in radiology.⁶ In addition, several other reporting guidelines focused on specific study designs in AI-based radiology research have also been developed.⁷

In this issue of Diagnostic and Interventional Radiology, Koçak et al.⁸ present an umbrella review of AI studies in radiology, conducting a comprehensive two-level assessment of adherence to CLAIM. Review articles published before August 2024 that evaluated studies using the original version of CLAIM were eligible for inclusion. Thirty-three review articles were assessed at the review level, encompassing a total of 1,458 studies. Of these, 421 studies were assessed individually at the study level, identified from 15 reviews, and CLAIM adherence was extracted at both levels as score and/or compliance values. Univariate and multivariate logistic regression analyses were performed to identify predictors of CLAIM adherence and critiques of CLAIM within the included reviews were also appraised.

Handling editor: Tuğba Akıncı D'Antonoli

Corresponding author: Ahmed Maiter

E-mail: a.maiter@sheffield.ac.uk

Received 04 April 2025; revision requested 08 April 2025;
accepted 13 April 2025.



Epub: 05.05.2025

Publication date: 01.09.2026

DOI: 10.4274/dir.2025.253315

You may cite this article as: Maiter A. Quality of reporting of studies on artificial intelligence in radiology: what is the current state of the field? *Diagn Interv Radiol.* 2026;32(5):616-617.

The results convey several important messages for the field of AI in radiology. First, there has been a clear dichotomy in how adherence to CLAIM has been measured in the literature, which poses challenges for the direct comparison of studies. Twenty-six reviews summarised CLAIM adherence as a score, indicating the total number of items reported out of all 42 in the checklist. In contrast, 18 reviews summarised adherence as compliance, referring to the proportion of applicable items that were reported. As Koçak et al.⁸ highlight, CLAIM is a broad reporting guideline, and not all items will be applicable to every study—for example, details about model parameter initialization are unlikely to apply to studies evaluating commercial devices in clinical settings. By accounting for item applicability, compliance is likely to be a more meaningful metric than score, and this work supports its adoption as the standard for measuring CLAIM compliance moving forward.

Second, Koçak et al.⁸ identified variability in the adherence of publications to CLAIM, reflecting inconsistencies in the quality of reporting across the field. The median score was 26 at both the review and study levels, whereas median compliance was 63% at the review level and 68% at the study level. In other words, around one-third of the CLAIM items were unreported in the included studies. Furthermore, the score was below 21 in 11% of studies, and compliance was below 50% in 10%, indicating that approximately one in ten studies reported only a minority of checklist items. Three variables were associated with higher CLAIM adherence: more recent year of publication, journal impact factor quartile, and specific radiology subfields. The link between impact factor quartiles and adherence suggests variable enforcement of CLAIM and reporting standards by journal editors and peer reviewers. Overall, the findings support greater use and application of CLAIM by researchers, journals, and peer reviewers.

Third, some items in CLAIM have been reported more frequently than others, suggesting systemic issues in the quality of reporting within the field. Eleven checklist items were reported in fewer than 50% of studies despite covering information that is crucial

for understanding the performance, generalizability, and scope of use of AI models. These frequently underreported items mostly relate to details about the data used or the methods and metrics of model evaluation. The inadequate reporting of certain items raises particularly serious concerns: a clear description of case inclusion and exclusion is necessary to identify selection biases in training and testing datasets; demographic and clinical characteristics of cases are essential for understanding model generalizability; measures of significance and uncertainty reflect the internal validity of model performance; and failure analysis provides insights into model limitations. Peer reviewers and editors may wish to pay particular attention to these 11 frequently underreported items when appraising manuscripts.

The comprehensive umbrella review by Koçak et al.⁸ highlights the variation in how CLAIM has been applied and the general variability in CLAIM adherence across studies on AI in radiology. Multiple shortcomings were identified, offering actionable insights for authors, editors, peer reviewers, and readers of publications in the field. There are, of course, limitations to the work, particularly the reliance on the quality and consistency of CLAIM evaluations in previous reviews and the inclusion of studies published before the original version of CLAIM. However, the two-level analysis, the consideration of both checklist score and compliance, and the identification of predictors of adherence together constitute a rigorous methodological approach.

Looking ahead, there is clear potential for implementing automated approaches to evaluating studies using CLAIM, which could help improve consistency across the field. Similar analyses could also be extended to other reporting guidelines relevant to AI in radiology. Lastly, it would be valuable to explore the relationship between CLAIM adherence and the downstream impact of studies, such as citation rates or clinical translation.

Footnotes

Funding

The author is funded by the National Institute for Health and Care Research (NIHR) Sheffield Biomedical Research Centre (NIHR203321). The views expressed are those of the author and not necessarily those of the NIHR or the Department of Health and Social Care.

Conflict of interest disclosure

The author is a member of the trainee editorial board for Radiology: Artificial Intelligence.

References

1. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. *Lancet*. 2008;371(9619):1149-1150. [\[Crossref\]](#)
2. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Med*. 2010;8:24. [\[Crossref\]](#)
3. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2(2):e200029. [\[Crossref\]](#)
4. Tejani AS, Klontzas ME, Gatti AA, et al. Updating the checklist for artificial intelligence in medical imaging (CLAIM) for reporting AI research. *Nat Mach Intell*. 2023;5(9):950-951. [\[Crossref\]](#)
5. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;6(4):e240300. [\[Crossref\]](#)
6. Maiter A, Salehi M, Swift AJ, Alabed S. How should studies using AI be reported? lessons from a systematic review in cardiac MRI. *Front Radiol*. 2023;3:1112841. [\[Crossref\]](#)
7. Klontzas ME, Gatti AA, Tejani AS, Kahn CE Jr. AI reporting guidelines: how to select the best one for your research. *Radiol Artif Intell*. 2023;5(3):e230055. [\[Crossref\]](#)
8. Koçak B, Köse F, Keleş A, Şendür A, Meşe İ, Karagülle M. Adherence to the checklist for artificial intelligence in medical imaging (CLAIM): an umbrella review with a comprehensive two-level analysis. *Diagn Interv Radiol*. 2025. [\[Crossref\]](#)



Radiomics-based differentiation of benign and malignant breast masses on contrast-enhanced mammography: a reproducible workflow using open-source tools

Aykut Teymur¹
 Sibel Kul²
 Ramazan Özgür Doğan³
 Hayati Türe⁴
 Temel Kayıkçıoğlu⁵
 Onur Bektaş²
 Sinan Erdemi⁶
 Cemal Aydoğan⁷

¹University of Health Sciences Türkiye, Antalya City Hospital, Clinic of Radiology, Antalya, Türkiye

²Karadeniz Technical University Faculty of Medicine, Department of Radiology, Trabzon, Türkiye

³Trabzon University Faculty of Computer and Information Sciences, Department of Artificial Intelligence Engineering, Trabzon, Türkiye

⁴Trabzon University Faculty of Computer and Information Sciences, Department of Computer Engineering, Trabzon, Türkiye

⁵Karadeniz Technical University Faculty of Engineering, Department of Electrical and Electronics Engineering, Trabzon, Türkiye

⁶University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Clinic of Radiology, İstanbul, Türkiye

⁷Sadika Sabancı State Hospital, Clinic of Radiology, Sakarya, Türkiye

Handling editor: Serap Gültekin

Corresponding author: Aykut Teymur

E-mail: aykuteymur@gmail.com

Received 02 December 2025; revision requested 02 January 2026; accepted 25 January 2026.



Epub: 12.02.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263222

PURPOSE

To differentiate benign and malignant breast masses by extracting radiomic features from low-energy and recombined contrast-enhanced mammography (CEM) images and to evaluate the diagnostic performance of multiple machine learning classifiers.

METHODS

In this retrospective, single-center study, 145 patients who underwent CEM between February 2019 and January 2022 were included. Radiomic features were extracted from manually segmented regions of interest on low-energy and recombined images using an open-source workflow (ITK-SNAP and PyRadiomics). The dataset was split at the patient level into a training set (75%) and an independent test set (25%); within the training set, feature selection and model optimization were performed using 10-fold cross-validation. Diagnostic performance [as measured by area under the curve (AUC), accuracy, sensitivity, specificity, positive predictive value, and negative predictive value] was reported on the held-out independent test set.

RESULTS

Ensemble learning demonstrated the best performance for both image types. The highest accuracy and AUC were 91.8% and 0.978 for recombined images and 89.7% and 0.968 for low-energy images, respectively. For recombined images, ensemble learning yielded the highest sensitivity (91.8%), whereas neural networks achieved the highest specificity (95.8%). For low-energy images, ensemble learning reached the highest sensitivity (98.0%), and decision trees achieved the highest specificity (91.7%).

CONCLUSION

Radiomics analysis of CEM images can effectively differentiate between benign and malignant breast masses, potentially enhancing diagnostic accuracy in breast imaging.

CLINICAL SIGNIFICANCE

A radiomics workflow based on recombined CEM images and open-source tools may complement conventional CEM interpretation, improve non-invasive lesion characterization, and support further research toward clinically validated decision-support applications.

KEYWORDS

Breast neoplasms, computer-aided diagnosis, contrast-enhanced mammography, machine learning, radiomics

Breast cancer is the most commonly diagnosed malignancy in women and a leading cause of cancer-related mortality.¹ Early detection, predicting treatment response, and estimating prognosis are crucial for improving survival rates.² Mammography remains the primary screening modality, reducing breast cancer mortality by approximately 30%.³ However, its sensitivity is limited in dense breasts.

Contrast-enhanced mammography (CEM) is a digital mammographic technique that provides functional and morphological information using iodinated contrast agents. The technique has been shown to have higher sensitivity than standard mammography and comparable performance to breast magnetic resonance imaging (MRI) while reducing false positives, and it is not affected by breast density.⁴ CEM has been increasingly used for lesion characterization, staging, and treatment monitoring.

In CEM, benign and malignant lesions are differentiated based on tumor shape, contour, contrast enhancement patterns, and kinetic characteristics. However, medical images contain quantitative data that are invisible to the human eye but can provide valuable diagnostic insights. Radiomics involves extracting and analyzing these high-dimensional features to characterize tissue properties. The radiomics workflow includes feature extraction through statistical, filtering, and morphological techniques, followed by feature selection to retain the most diagnostically relevant parameters. Machine learning algorithms then classify lesions as benign or malignant based on these features.

Main points

- Radiomics-based analysis of recombined contrast-enhanced mammography (CEM) images achieved high diagnostic performance in differentiating benign from malignant breast masses.
- An open-source workflow using ITK-SNAP and PyRadiomics provides a transparent and reproducible pipeline that can be implemented in radiology departments equipped with digital mammography systems.
- Shape- and texture-based radiomic features derived from CEM may serve as quantitative biomarkers to support lesion characterization and risk stratification research in breast imaging.
- Recombined CEM images outperformed low-energy images across multiple machine learning classifiers, highlighting the added value of contrast-enhanced information for lesion characterization.

Although radiomics has been widely studied in non-contrast mammography and MRI, its application in CEM is relatively new. Preliminary studies have demonstrated that radiomics analysis of CEM images can achieve classification accuracies ranging from 80% to 90% in tumor classification and holds promise for distinguishing subtypes, assessing invasiveness, and predicting tumor grade.⁵⁻⁹

This study extracts radiomic features from CEM images for benign–malignant mass differentiation and evaluates their diagnostic performance using machine learning algorithms.

Methods

Study population

A total of 145 patients with suspicious breast masses on CEM were retrospectively included, yielding 164 breast masses (73 benign and 91 malignant) for the final radiomics analysis. Patients who had contraindications to iodinated contrast agents or incomplete imaging data were excluded. Malignant lesions were one per patient (91 lesions in 91 patients); the higher number of lesions than patients was due to the benign group, in which 58 patients contributed 73 benign lesions (i.e., multiple lesions occurred only in the benign subgroup). Each lesion was segmented and analyzed as a separate lesion-level sample. The patient and lesion selection process is summarized in Figure 1. This retrospective study was approved by the Institutional Review Board of Karadeniz Technical University (approval number: 2022/121, date: June 2, 2022), and the requirement for informed consent was waived.

Imaging protocol

CEM was performed using a digital mammography unit (Senographe Essential, GE Healthcare, Buc, France). An intravenous contrast agent (1.5 mL/kg, 50–120 mL) was administered at 3 mL/s. Craniocaudal (CC) and mediolateral oblique (MLO) views of both breasts were acquired, starting approximately 2 minutes after contrast injection, and all views were completed within 6–7 minutes, generating low-energy and recombined images.

Radiomics analysis

Image assessment and segmentation

CEM images were evaluated using a dedicated mammography workstation by two radiologists with 20 and 3 years of experience

in breast imaging, respectively. Lesion size and histopathological type were recorded for each patient.

All images were stored in DICOM format and processed using ITK-SNAP 3.8 (University of Pennsylvania, Philadelphia, PA, USA; www.itksnap.org), an open-source image segmentation tool. The radiologist with 3 years of experience manually segmented the lesions, ensuring that the region of interest (ROI) strictly encompassed the lesion itself (Figure 2). This segmentation was applied to all low-energy and recombined CC and MLO images.

Feature extraction and selection

Radiomic feature extraction was conducted using PyRadiomics (AIM-Harvard, Boston, MA, USA), an open-source Python package for radiomic feature extraction from two-dimensional and three-dimensional images. No pre-processing was applied before extraction. A total of 102 radiomic features were computed, categorized into the following matrices:

- Shape-based features
- First-order statistical features
- Gray-Level Co-occurrence Matrix (GLCM)
- Gray-Level Run Length Matrix (GLRLM)
- Gray-Level Size Zone Matrix (GLSZM)
- Gray-Level Dependence Matrix (GLDM)
- Neighboring Gray-Tone Difference Matrix (NGTDM) (Supplementary Table 1)

To reduce dimensionality and improve model efficiency, minimum redundancy–maximum relevance, ReliefF, and ANOVA algorithms were implemented in MATLAB R2022b (MathWorks, Inc., Natick, MA, USA). Each algorithm generated a ranking based on feature importance scores. The top 10 most significant features were selected by each algorithm (Figures 3 and 4). A total of 22 features for recombined images and 25 features for low-energy images were selected for analysis (Supplementary Table 2).

Classification and model optimization

Supervised machine learning classifiers were developed to classify ROIs as benign or malignant. All analyses were performed using MATLAB R2022b. Prior to model development, the dataset was split at the patient level into a training set (75%) and an independent test set (25%) to prevent information leakage. When multiple benign lesions

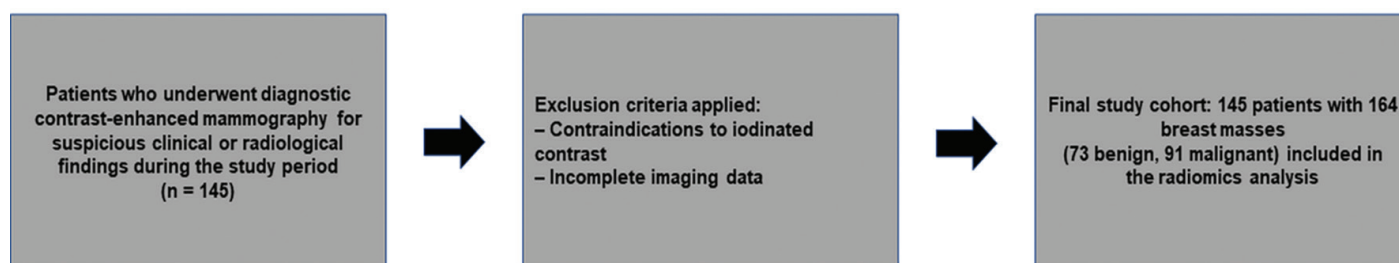


Figure 1. Flowchart of patient and lesion selection. Flow diagram showing the selection of patients and lesions for inclusion in the radiomics analysis. During the study period, 145 patients who underwent diagnostic contrast-enhanced mammography for suspicious clinical or radiological findings were screened. After application of the inclusion and exclusion criteria, 145 patients with 164 breast masses (73 benign and 91 malignant) were included in the final study cohort.

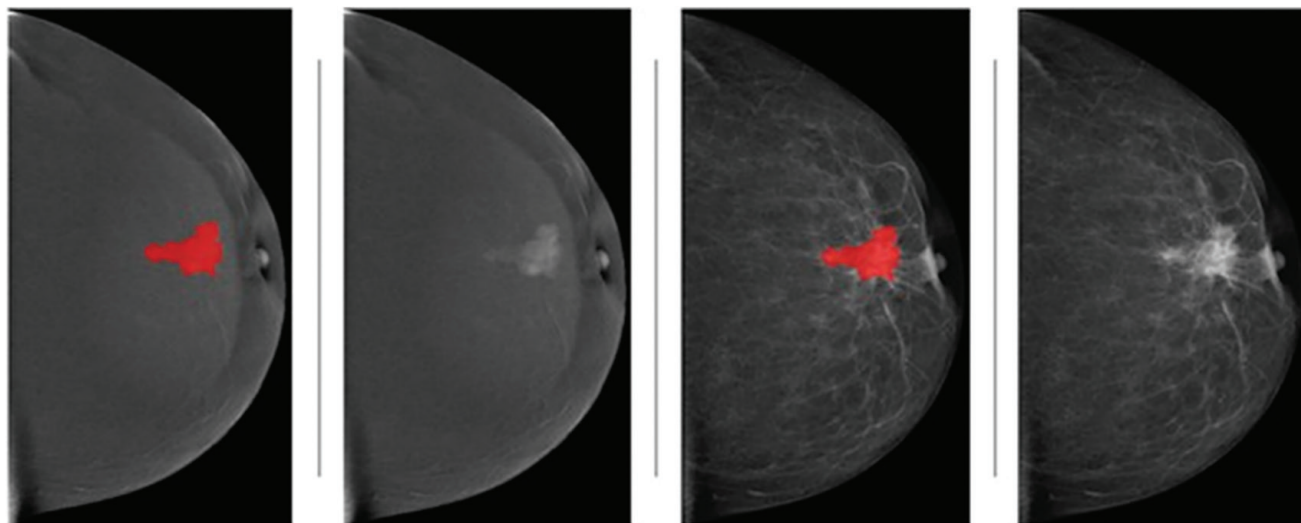


Figure 2. Representative malignant breast mass and manual segmentation on contrast-enhanced mammography. Contrast-enhanced mammography images of a biopsy-proven malignant breast mass. The region of interest is manually delineated on both low-energy and recombined images, precisely including only the enhancing tumor and excluding surrounding parenchyma, skin, and vessels. The same segmentation strategy was applied to all low-energy and recombined craniocaudal mediolateral and oblique images before radiomic feature extraction.

were present in the same patient, all lesions from that patient were kept within the same split. The independent test set was not used at any stage of feature selection, model tuning, or model selection.

The following classifiers were evaluated: ensemble learning, decision trees, naïve Bayes, support vector machines, and neural networks (Supplementary Table 3). Within the training set, 10-fold cross-validation was used for model optimization and selection. The final selected model was then trained on the full training set and evaluated on the held-out independent test set. Performance was assessed using area under the receiver operating characteristic curve (ROC) AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

All steps of the radiomics workflow, including image acquisition, lesion segmentation, feature extraction, feature selection, and model training, are summarized in Figure 5.

Statistical analysis

Descriptive statistics for continuous variables were expressed as mean $\pm$ standard deviation, and categorical variables as numbers and percentages. The machine learning workflow, including patient-level data partitioning, cross-validation, and leakage prevention, is described in the “Classification and Model Optimization” section. Diagnostic performance metrics (AUC, accuracy, sensitivity, specificity, PPV, and NPV) were reported for the held-out independent test set.

Results

Patient characteristics

A total of 164 ROIs were analyzed. Of these, 44.5% (73) were benign, and 55.5% (91) were malignant. The mean age was 48.1 ± 9.5 years for benign cases and 49.9 ± 10.5 years for malignant cases. Invasive ductal carcinoma was the most common malignant diagnosis (57%), and stable follow-up mass-

es were the most common benign findings (41%). The mean lesion size was 23.2 ± 20.6 mm in the benign group and 35.7 ± 21.5 mm in the malignant group (Tables 1 and 2).

Performance of radiomics analysis

All reported performance metrics represent results from the held-out independent test set (25% of patients). Model selection and optimization were performed within the training set using 10-fold cross-validation.

The classification performance of five different machine learning algorithms was evaluated. The confusion matrices and ROC curves for recombined images are illustrated in Supplementary Figure 1. Among all models, ensemble learning exhibited the best diagnostic performance, achieving an AUC value of 0.9783. The corresponding AUC values for the other classifiers were 0.9728 for neural networks, 0.9583 for support vector machines, 0.8793 for decision trees, and 0.8520 for naïve Bayes.

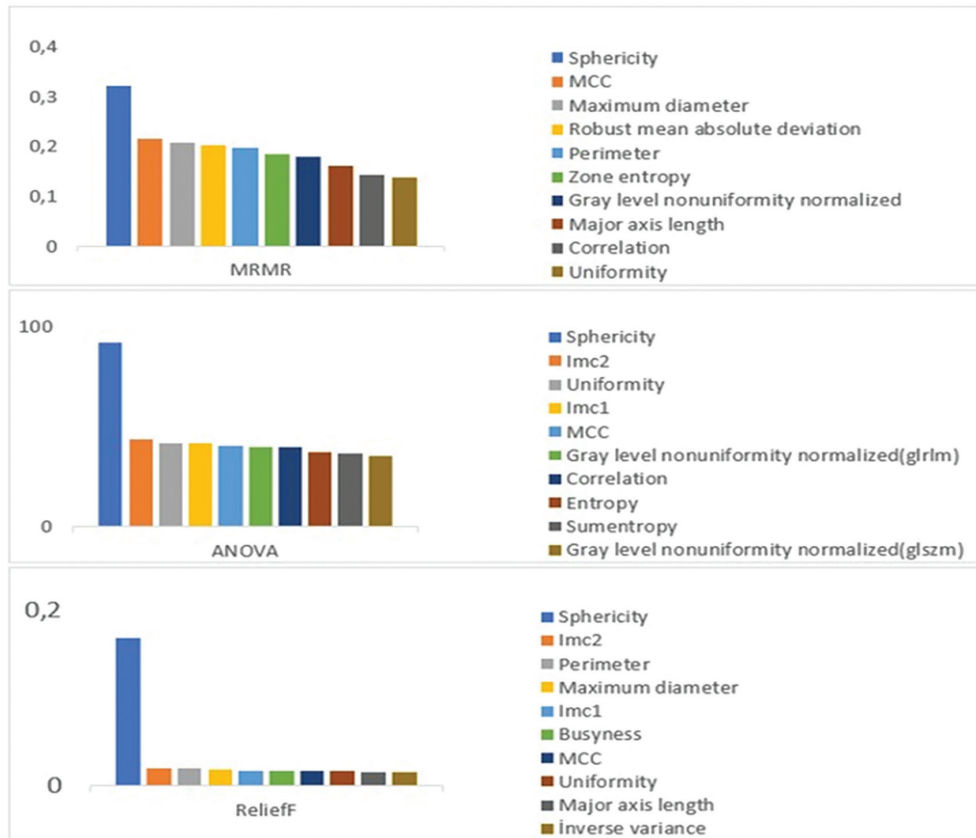


Figure 3. Top 10 most relevant radiomic features in recombined images. Bar chart showing the 10 most relevant radiomic features selected from recombined contrast-enhanced mammography images by the feature selection algorithms (minimum redundancy–maximum relevance, ReliefF, and ANOVA). Feature importance scores are normalized to allow visual comparison of the relative contribution of each feature. ANOVA, analysis of variance; MCC, maximal correlation coefficient.

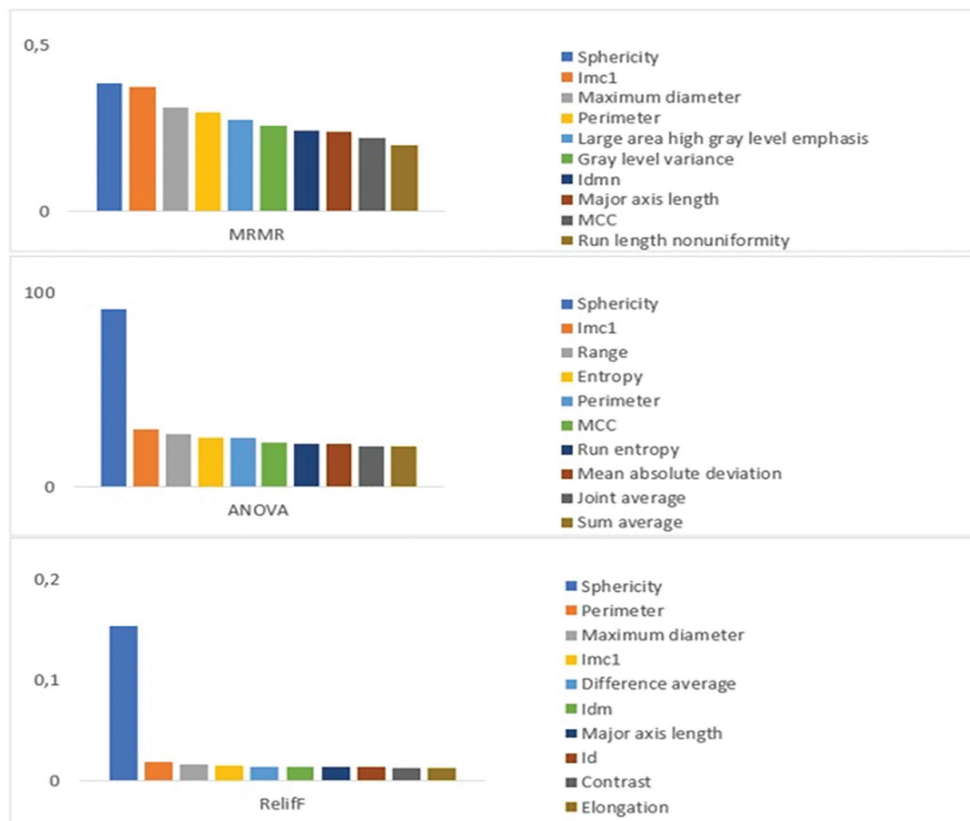


Figure 4. Top 10 most relevant radiomic features in low-energy images. Bar chart showing the 10 most relevant radiomic features selected from low-energy contrast-enhanced mammography images by the same feature selection algorithms (minimum redundancy–maximum relevance, ReliefF, and ANOVA). As in recombined images, texture- and shape-based features predominate among the highest-ranking variables. ANOVA, analysis of variance; MCC, maximal correlation coefficient.

The classification results for low-energy images are depicted in Supplementary Figure 2. As with recombined images, ensemble learning exhibited the highest classification performance, with an AUC value of 0.9677. The AUC values for the other classifiers were 0.9418 for neural networks, 0.9332 for support vector machines, 0.8695 for decision trees, and 0.8818 for naïve Bayes.

The accuracy and AUC values for each classifier in both recombined and low-energy images are summarized in Table 3. The ensemble learning classifier achieved the highest accuracy, with values of 91.75% for recombined images and 89.69% for low-en-

ergy images. The performance of the other classifiers is outlined below:

- For recombined images, the accuracy rates were as follows: neural networks (89.85%), support vector machines (89.69%), decision trees (84.74%), and naïve Bayes (84.54%).
- For low-energy images, the accuracy rates were as follows: neural networks (87.63%), support vector machines (85.57%), decision trees (86.60%), and naïve Bayes (82.47%).

Table 4 summarizes the sensitivity, specificity, PPV, and NPV of each classifier. In re-

combined images, ensemble learning and support vector machines exhibited the highest sensitivity (91.83% and 91.82%, respectively). The highest specificity and PPV were obtained with the neural network classifier (95.83% and 95.34%, respectively). Ensemble learning achieved the highest NPV at 91.66%.

For low-energy images, ensemble learning exhibited the highest sensitivity (97.95%) and NPV (95.50%). The highest specificity (91.66%) and PPV (90.90%) were achieved using the decision tree classifier.

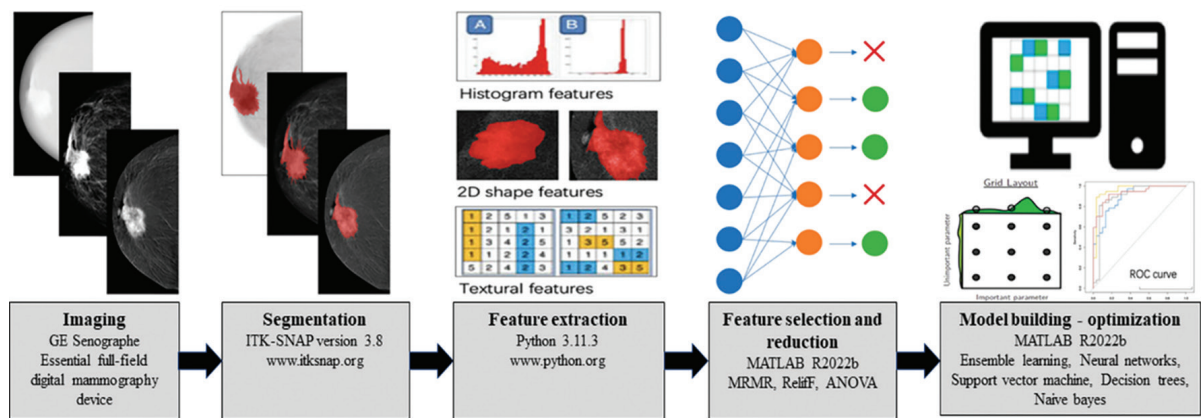


Figure 5. Radiomics workflow for contrast-enhanced mammography-based lesion classification. Schematic overview of the radiomics workflow applied in this study. Following diagnostic labeling, the dataset was split at the patient level into a training set (75%) and an independent test set (25%). Within the training set, feature selection and model optimization were performed using 10-fold cross-validation. Final performance metrics were calculated on the held-out test set to differentiate benign from malignant breast masses on contrast-enhanced mammography (CEM).

Table 1. Demographic characteristics of the study cohort (lesion-based analysis, n = 164)			
			n (%)
Definitive diagnosis (n = 164)	Benign		73 (44.5)
	Malignant		91 (55.5)
Age (n = 164)	Benign	< 40	19 (26.0)
		40–49	20 (27.4)
		50–59	22 (30.1)
		≥ 60	12 (16.4)
		Mean ± SD: 48.1 ± 9.5; median: 49; range: 30–65	
	Malignant	< 40	16 (17.6)
		40–49	28 (30.8)
		50–59	25 (27.5)
		≥ 60	22 (24.2)
		Mean ± SD: 49.9 ± 10.5; median: 50; range: 26–75	
Percentages are calculated within each subgroup (benign/malignant) for age categories; overall percentages are reported for the total sample where applicable. SD, standard deviation.			

A) Final diagnosis, n (%)		
	Diagnosis/category	n (%)
Benign (n = 73)	Fibroadenoma	2 (2.7)
	Papilloma	8 (11.0)
	Mastitis	11 (15.1)
	Fibrocystic disease	2 (2.7)
	Epithelial hyperplasia	2 (2.7)
	Stable on follow-up	30 (41.1)
	Other benign	18 (24.7)
Malignant (n = 91)	Invasive ductal carcinoma	52 (57.1)
	Invasive lobular carcinoma	8 (8.8)
	IDC + ILC	4 (4.4)
	IDC + ductal carcinoma <i>in situ</i>	14 (15.4)
	Other malignant	13 (14.3)
B) Maximum tumor diameter categories, n (%)		
	Maximum diameter	n (%)
Benign (n = 73)	< 10 mm	28 (38.4)
	11–20 mm	16 (21.9)
	21–30 mm	11 (15.1)
	31–40 mm	8 (11.0)
	41–50 mm	3 (4.1)
	> 51	7 (9.6)
	Mean ± SD, mm: 23 ± 20; median: 18; range: 4–90	
Malignant (n = 91)	<10 mm	3 (3.3)
	11–20 mm	17 (18.7)
	21–30 mm	31 (34.1)
	31–40 mm	15 (16.5)
	41–50 mm	9 (9.9)
	> 51	16 (17.6)
	Mean ± SD, mm: 36 ± 22; Median: 28; Range: 10–95	

Percentages are calculated within the benign (n = 73) and malignant (n = 91) subgroups. IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; SD, standard deviation.

	Recombined images		Low-energy images	
Model	Accuracy (%)	AUC	Accuracy (%)	AUC
Ensemble learning	91.75	0.9783	89.69	0.9677
Neural network	89.85	0.9728	87.63	0.9418
Support vector machine	89.69	0.9583	85.57	0.9332
Decision trees	84.74	0.8793	86.60	0.8695
Naïve Bayes	84.54	0.8520	82.47	0.8818

Performance metrics are reported for the held-out test set; 10-fold cross-validation was used within the training set for model optimization. AUC, area under the curve.

Table 4. Sensitivity, specificity, and positive and negative predictive values of five different models for recombined and low-energy images

	Model	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)
Recombined images	Ensemble learning	91.83	91.66	91.83	91.66
	Neural network	83.67	95.83	95.34	85.18
	Support vector machine	91.82	87.50	88.23	91.30
	Decision trees	85.71	83.33	84.00	85.10
	Naïve Bayes	89.79	79.16	81.48	88.37
Low-energy images	Ensemble learning	97.95	81.25	84.21	97.50
	Neural network	87.75	87.50	87.75	87.50
	Support vector machine	85.71	85.41	85.71	85.41
	Decision trees	81.63	91.66	90.90	83.01
	Naïve Bayes	87.75	77.08	79.62	86.04

Discussion

CEM enhances the diagnostic accuracy of digital mammography by visualizing tumor neovascularity through iodinated contrast administration. The CEM technique enables the concurrent evaluation of morphological and enhancement characteristics, achieving a sensitivity range of 96%–100%, comparable to that of MRI, and improving the specificity of digital mammography from 42% to 87.7%.^{10,11}

Radiomics is an emerging field that uses extracted quantitative imaging biomarkers to enhance diagnostic accuracy and improve patient management. By analyzing texture, shape, and statistical features, radiomics provides an objective and reproducible assessment of tumor characteristics. Although most radiomics studies have focused on MRI, CEM offers a unique advantage by combining mammographic and contrast-enhanced information, positioning it as a promising modality for radiomics analysis.^{2,5,12}

In the present study, 102 radiomic features, including first-order, shape-based, and texture-based metrics (GLCM, GLRLM, GLSZM, NGTDM, and GLDM), were evaluated. Compared with previous radiomics studies using CEM, our models achieved numerically higher AUCs and accuracies in differentiating benign from malignant lesions (Supplementary Table 4).^{5,9,13–17} The more comprehensive feature set used in this study may have contributed to these improved results.

Our findings indicate that recombined CEM images outperform low-energy images in distinguishing benign from malignant masses. Ensemble learning achieved the highest accuracy for both image types, whereas neural networks demonstrated the

best specificity for recombined images, and support vector machines performed best for low-energy images.

Clinical relevance and practical applicability

The current study aims to bridge the gap between advanced image-based computational analysis and clinical decision-making in breast imaging. By using widely accessible open-source tools such as ITK-SNAP and PyRadiomics and integrating standard segmentation and feature selection algorithms, the proposed methodology can be feasibly replicated in radiology departments equipped with digital mammography infrastructure. The selected radiomic features, particularly shape- and texture-based descriptors, may aid in distinguishing benign from malignant lesions and complement routine CEM assessment. However, this retrospective study did not evaluate integration with established clinical assessment systems (e.g., BI-RADS), perform comparisons with radiologist interpretation, or conduct decision-threshold and clinical-utility analyses, meaning the potential impact on clinical decision-making (including biopsy decisions) should be interpreted cautiously. With additional multi-reader robustness testing, external validation in larger multicenter cohorts, and prospective clinical-utility assessment, this approach could serve as a foundation for future decision-support tools.

This retrospective, single-center study has several limitations. First, the sample size was modest, with an imbalanced distribution of benign and malignant lesions. Second, lesions were manually segmented by a single reader; thus, interobserver agreement and segmentation robustness (e.g., intraclass correlation coefficient-based feature stability)

could not be assessed, which may affect reproducibility. Third, no external validation cohort was available, and the extremely high AUC values may partly reflect optimistic performance and potential overfitting in a limited dataset with high-dimensional radiomic features. Future research should incorporate larger, multicenter datasets, harmonization strategies, multi-reader or automated segmentation approaches, and external validation.

In conclusion, radiomics analysis of CEM—particularly recombined images—showed strong discrimination between benign and malignant breast masses within an open-source and reproducible workflow. These findings highlight the potential of integrating radiomics-derived quantitative biomarkers into artificial intelligence-based decision-support systems for breast lesion characterization. Future studies should evaluate model generalizability and robustness in larger cohorts by incorporating multi-reader segmentation robustness assessment and independent external multicenter validation and by benchmarking performance against standard clinical assessment frameworks.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Figures 1 and 2: <https://d2v96fxpocvxx.cloudfront.net/bda9171a-fae8-4995-8276-2138323f1e16/content-images/fdebcb55-e020-4311-b20b-1c52c401fc4a.pdf>

Supplementary Tables 1–4: <https://d2v96fxpocvxx.cloudfront.net/87200e80-fa4b-4309-8816-eff4c4e638c8/content-images/60a4326b-2e21-4bcd-8d2e-bb3c4fc4ee99.pdf>

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229-263. [\[Crossref\]](#)
2. Valdora F, Houssami N, Rossi F, Calabrese M, Tagliafico AS. Rapid review: radiomics and breast cancer. *Breast Cancer Res Treat.* 2018;169(2):217-229. [\[Crossref\]](#)
3. Liu H, Chen Y, Zhang Y, et al. A deep learning model integrating mammography and clinical factors facilitates the malignancy prediction of BI-RADS 4 microcalcifications in breast cancer screening. *Eur Radiol.* 2021;31(8):5902-5912. [\[Crossref\]](#)
4. James JJ, Tennant SL. Contrast-enhanced spectral mammography (CESM). *Clin Radiol.* 2018;73(8):715-723. [\[Crossref\]](#)
5. Losurdo L, Fanizzi A, Basile TMA, et al. Radiomics analysis on contrast-enhanced spectral mammography images for breast cancer diagnosis: a pilot study. *Entropy.* 2019;21(11):1110. [\[Crossref\]](#)
6. Marino MA, Pinker K, Leithner D, et al. Contrast-enhanced mammography and radiomics analysis for noninvasive breast cancer characterization: initial results. *Mol Imaging Biol.* 2020;22(3):780-787. [\[Crossref\]](#)
7. La Forgia D, Fanizzi A, Campobasso F, et al. Radiomic analysis in contrast-enhanced spectral mammography for predicting breast cancer histological outcome. *Diagnostics (Basel).* 2020;10(9):708. [\[Crossref\]](#)
8. Fusco R, Piccirillo A, Sansone M, et al. Radiomics and artificial intelligence analysis with textural metrics extracted by contrast-enhanced mammography in the breast lesions classification. *Diagnostics (Basel).* 2021;11(5):815. [\[Crossref\]](#)
9. Massafra R, Bove S, Lorusso V, et al. Radiomic feature reduction approach to predict breast cancer by contrast-enhanced spectral mammography images. *Diagnostics (Basel).* 2021;11(4):684. [\[Crossref\]](#)
10. Lobbes MB, Lalji U, Houwers J, et al. Contrast-enhanced spectral mammography in patients referred from the breast cancer screening programme. *Eur Radiol.* 2014;24(7):1668-1676. [\[Crossref\]](#)
11. Lalji UC, Houben IP, Prevos R, et al. Contrast-enhanced spectral mammography in recalls from the Dutch breast cancer screening program: validation of results in a large multireader, multicase study. *Eur Radiol.* 2016;26(12):4371-4379. [\[Crossref\]](#)
12. Pathak B, Barooah D. Texture analysis based on the gray-level co-occurrence matrix considering possible orientations. *Int J Adv Res Electr Electron Instrum Energy.* 2013;2:4206-4212. [\[Crossref\]](#)
13. Patel BK, Ranjbar S, Wu T, et al. Computer-aided diagnosis of contrast-enhanced spectral mammography: A feasibility study. *Eur J Radiol.* 2018;98:207-213. [\[Crossref\]](#)
14. Perek S, Kiryati N, Zimmerman-Moreno G, Sklair-Levy M, Konen E, Mayer A. Classification of contrast-enhanced spectral mammography (CESM) images. *Int J Comput Assist Radiol Surg.* 2019;14(2):249-257. [\[Crossref\]](#)
15. Fanizzi A, Losurdo L, Basile TMA, et al. Fully automated support system for diagnosis of breast cancer in contrast-enhanced spectral mammography images. *J Clin Med.* 2019;8(6):891. [\[Crossref\]](#)
16. Mateos MJ, Gastelum A, Márquez J, et al. Texture analysis of contrast-enhanced digital mammography (CEDM) images. *Lecture Notes in Computer Science.* 2016;9699:585-592. [\[Crossref\]](#)
17. Wang S, Mao N, Duan S, et al. Radiomic analysis of contrast-enhanced mammography with different image types: classification of breast lesions. *Front Oncol.* 2021;11:600546. [\[Crossref\]](#)



A scoping review of photon-counting detector computed tomography in cardiovascular imaging

Deniz Alis¹
 Mehmet Onur Önal²
 Müjgan Orman³
 Mustafa Ege Seker⁴
 Ahmet Akyol⁵
 Cem Alhan⁶
 Ercan Karaarslan¹

¹Acıbadem Mehmet Ali Aydınlar University Faculty of Medicine, Department of Radiology, İstanbul, Türkiye

²İstanbul University-Cerrahpaşa, Cerrahpaşa Faculty of Medicine, Department of Radiology, İstanbul, Türkiye

³Acıbadem Healthcare Group, Department of Radiology, İstanbul, Türkiye

⁴University of Wisconsin–Madison, Department of Radiology, Wisconsin, United States of America

⁵Acıbadem Mehmet Ali Aydınlar University Faculty of Medicine, Department of Cardiology, İstanbul, Türkiye

⁶Acıbadem Mehmet Ali Aydınlar University Faculty of Medicine, Department of Cardiovascular Surgery, İstanbul, Türkiye

ABSTRACT

Photon-counting detector computed tomography (PCD-CT) employs direct-conversion detectors that record the arrival and energy of individual photons, enabling high-resolution, multi-energy cardiovascular imaging. We searched MEDLINE, Embase, and Scopus from January 2021 through September 2025 and included 59 studies. Owing to heterogeneity in study designs, protocols, and endpoints, the findings were synthesized narratively across five domains (coronary, myocardial, structural/valvular, pulmonary–cardiopulmonary function, and aortic/visceral/peripheral arteries). In coronary imaging, a routine-practice cohort ($n = 7,833$) reported a per-patient specificity of 98% vs. 93%, lower invasive angiography of 9.9% vs. 13.1%, and a higher revascularization yield of 43.4% vs. 35.5% [PCD-CT vs. energy-integrating detector CT (EID-CT)]; ultra-high-resolution protocols achieved a vessel-level area under the curve (AUC) of up to 0.99. Low-dose CCTA was feasible at a CTDIvol of 1.72 mGy, and contrast-saving protocols supported diagnostic studies with a volume of 30 mL. Virtual non-contrast calcium scoring showed an intraclass correlation coefficient of 0.98 vs. true non-contrast. In myocardial tissue characterization, PCD-CT-derived extracellular volume differed from cardiovascular magnetic resonance by $\leq 2\%$ in selected cohorts, with a kappa of up to 0.956 for late-enhancement agreement; segment-level inflammation classification reached an AUC of 0.95. For structural/valvular assessment, a comparative cohort reported an effective dose of 8.8 ± 4.5 vs. 15.3 ± 5.8 mSv, with visual image quality scores of 4.8 vs. 3.3, respectively, for PCD-CT vs. EID-CT. Lung-perfusion iodine maps for chronic thromboembolic pulmonary hypertension achieved an accuracy of 0.85–0.88 at approximately one-fifth of the dose of single-photon emission CT. For aortic/peripheral applications, thoracoabdominal protocols reported a dose of 4.2 ± 1.4 vs. 7.2 ± 2.2 mGy, with a higher signal-to-noise ratio/contrast-to-noise ratio (PCD-CT vs. EID-CT); intrapopliteal imaging used 60 versus 140 mL of contrast, respectively, with improved vessel sharpness for PCD-CT vs. EID-CT; diagnostic performance for peripheral stenosis reached a sensitivity of 91% and a specificity of 95%, respectively, when compared with digital subtraction angiography. Overall, the evidence—predominantly single center—indicates that PCD-CT may enable dose- and contrast-efficient cardiovascular imaging with strong diagnostic metrics, and confirmation in multicenter outcome and cost-effectiveness studies remains a priority.

KEYWORDS

Photon counting, computed tomography, cardiovascular imaging, diagnostic accuracy, computed tomography angiography

Handling editor: Furkan Ufuk

Corresponding author: Deniz Alis

E-mail: drdenizalis@gmail.com

Received 07 August 2025; revision requested 08 September 2025; last revision received 17 October 2025; accepted 27 October 2025.



Epub: 20.11.2025

Publication date: 01.09.2026

DOI: 10.4274/dir.2025.253597

Cardiovascular computed tomography (CT) is a critical tool for the non-invasive assessment of coronary artery disease, myocardial tissue pathology, structural/valvular disease, pulmonary vascular disorders, and aortic/peripheral arterial pathology. Photon-counting detector CT (PCD-CT) represents a shift in detector design. By registering the arrival and energy of individual X-ray photons, PCD-CT can reduce electronic noise and support high-resolution (HR), multi-energy imaging more effectively than energy-integrating detector CT (EID-CT).

PCD-CT can achieve submillimeter spatial resolution at routine radiation doses, enabling coronary CT angiography (CCTA) that depicts 1–2-mm distal vessels and fine morphologic features (e.g., non-calcified plaques, napkin-ring signs, ostial lesions). In parallel, energy-based

You may cite this article as: Alis D, Önal MO, Orman M, et al. A scoping review of photon-counting detector computed tomography in cardiovascular imaging. *Diagn Interv Radiol.* 2026;32(5):626–642.

photon weighting can mitigate calcium and stent blooming and improve in-stent lumen depiction and calcium quantification compared with EID-CT.¹⁻⁴ Beyond morphology, the multi-energy output allows retrospective monoenergetic reconstructions, iodine maps, and (where available) K-edge material decomposition. These capabilities have supported single-acquisition protocols capturing coronary and myocardial phases, *in vivo* plaque imaging, and angiographic strategies that have reported a higher contrast-to-noise ratio (CNR) with iodine dose reductions of up to 40% in selected settings. Low-keV reconstructions have been reported to maintain diagnostic quality despite suboptimal bolus timing and to facilitate single-scan endoleak evaluation after aortic repair.⁵⁻¹⁰

These capabilities extend beyond coronary imaging: iodine-based maps and delayed-enhancement surrogates enable myocardial tissue characterization; HR sizing and peripheral access planning support structural/valvular workflows; iodine mapping supports pulmonary perfusion assessment alongside embolus detection; and low-keV reconstructions can improve vascular conspicuity in aortic and peripheral arteries, enabling contrast-saving protocols.⁵⁻¹⁰ Given heterogeneity in protocols, reconstructions, and endpoints across studies, a structured synthesis is warranted.

Accordingly, we present a state-of-the-art scoping review of peer-reviewed adult cardiovascular PCD-CT studies published through September 2025, synthesizing findings across five domains—coronary arteries, myocardial tissue, structural heart/valves, pulmonary–cardiopulmonary function, and aortic/visceral/peripheral arteries—and

Main points

- This scoping review synthesizes findings from 59 clinical studies evaluating photon-counting detector computed tomography (PCD-CT) in cardiovascular imaging.
- PCD-CT can provide higher spatial resolution and spectral detail than energy-integrating detector CT, enhancing the detection of small anatomical structures and subtle pathologies.
- In selected protocols and cohorts, PCD-CT can reduce radiation exposure and contrast-medium use, typically by up to 40%–60%, although effects vary by indication and technique.
- Its multi-energy capabilities can enable functional assessments—such as myocardial characterization and pulmonary perfusion mapping—often within a single scan.

complementing these with clearly labeled, illustrative examples from our single-center experience [$>1,000$ PCD-CT examinations on a (Siemens, Forchheim, Germany), not included in the current synthesis].

Methods

A scoping literature review was conducted to map and summarize clinical studies of PCD-CT in cardiovascular imaging. The search covered January 2021 (onset of clinical availability) through September 1, 2025. Searches were performed in PubMed/MEDLINE, Embase, and Scopus using a broadened keyword set combining PCD-CT terms and cardiovascular terms. The core logic included PCD-CT synonyms (e.g., “photon-counting detector computed tomography,” “photon-counting CT,” “PCD-CT,” “SPCCT,” and device names where reported) AND cardiovascular concepts (cardiac, cardiovascular, coronary, vascular, myocardial, aorta, valvular/structural, TAVI/TAVR, pulmonary).

The initial search yielded 828 records. After deduplication and title/abstract screening, 467 records remained for full-text as-

essment. Original clinical studies in adults that used PCD-CT for any cardiovascular indication, were published in peer-reviewed journals, and reported at least one imaging, diagnostic, or workflow outcome were included. Phantom-, animal-, or simulation-only studies; pediatric or congenital cohorts; neurovascular studies; abstract-only publications; and purely technical reports without a clinical cohort were excluded. A total of 59 studies met the criteria and were included in the review. The evidence-identification flow is summarized in Figure 1.

Two reviewers independently screened titles/abstracts and full texts, resolving disagreements by consensus. For each included study, the following was charted: study design and setting; patient characteristics; scanner/vendor; acquisition and reconstruction parameters [e.g., kilovolt peak (kVp), pitch, ultra-HR (UHR) protocols, virtual monoenergetic image (VMI) keV levels, kernels/iterative settings]; dose metrics as reported [CT dose index volume (CTDIvol), dose-length product (DLP), size-specific dose estimate (SSDE), and/or effective dose]; contrast volume; image-quality measures [objective

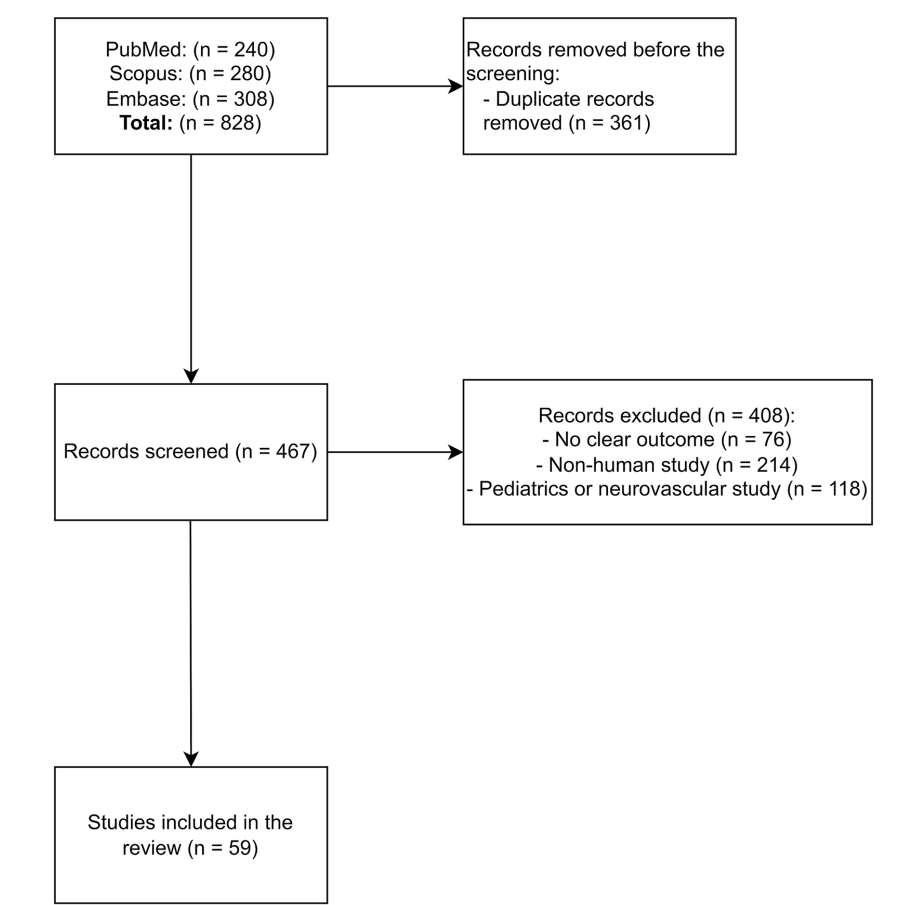


Figure 1. Literature search diagram. The initial search yielded 828 publications, which were reduced to 467 articles after the removal of duplicates and irrelevant studies. A total of 59 studies were ultimately included in the final review and synthesis, as they met all predefined inclusion criteria.

and subjective, including the signal-to-noise ratio (SNR)/CNR when provided]; and diagnostic performance vs. the stated reference standard [e.g., area under the curve (AUC), sensitivity, specificity].

Given heterogeneity in populations, protocols, endpoints, and reporting formats, no quantitative pooling or meta-analysis was performed. Findings are synthesized narratively, with numeric ranges reported by application domain (coronary, myocardium, structural/valvular, pulmonary, aortic/peripheral).

Results

Coronary computed tomography angiography

A total of 31 investigations evaluating CCTA with PCD-CT were identified (Table 1). Study designs included retrospective and prospective diagnostic accuracy cohorts, protocol-optimization studies for radiation/contrast reduction, and quantitative assessments using virtual monoenergetic imaging and iterative reconstruction. Several studies used UHR (0.2 mm) acquisitions.^{11,12}

In a large retrospective routine-practice cohort (n = 7,833), per-patient specificity with PCD-CT was 98% and 93% with EID-CT, positive predictive value was 83% and 63%, invasive coronary angiography (ICA) referral was 9.9% and 13.1%, and revascularization yield was 43.4% and 35.5%, respectively.¹³ In a high-risk cohort with heavy calcification or stents (n = 68), sensitivity and specificity were 96% and 84%, respectively.¹⁴ Two UHR studies reported a mean stenosis-measurement error of 6% and a vessel-level AUC of up to 0.99.^{11,12}

Table 1. Photon-counting CT studies related to coronary CT angiography

Author	Journal	Study design	Cohort	Imaging	Results	Conclusion
1. Greffier et al. ¹⁹ 2023	<i>Eur Radiol.</i>	Phantom + preliminary human	8 pts + phantom	Spectral PCD-CT HR (0.67 mm) and UHR (0.43 mm) VMIs 40–90 keV	Across all keV, PCD-CT f peak, f 50, d' > dual-layer DECT; radiologists rated lumen sharpness and IQ higher for PCD-CT	High- and ultra-high-resolution PCD-CT VMIs markedly enhance coronary-lumen detectability versus dual-energy CT
2. Hagar et al. ¹⁴ 2023	<i>Radiology.</i>	Prospective diagnostic accuracy	68 pts	Retrospectively ECG-gated UHR CCTA, 120/140 kV, 120 × 0.2 mm, 100 mL iopromid	AUC of 0.93 pp/0.94 pv/0.92 ps; sensitivity of 96%, specificity of 84%, accuracy of 88% (per patient)	UHR PCD-CT CCTA delivers high accuracy for CAD even in pts at high risk and with heavy calcification and stents
3. Kahmann et al. ²⁹ 2023	<i>Front Cardiovasc Med.</i>	Retrospective single-center radiomics analysis	66 pts	First-gen PCD-CT; manual PCAT segmentation (LCA & RCA); pyradiomics	Two texture features (gldm_HighGrayLevelEmphasis 23.95 vs. 22.99; glrlm_HighGrayLevelRunEmphasis 24.21 vs. 23.31; $P = 0.013$ – 0.24) differentiated hyper-cholesterolemic vs. normo-lipidemic patients	PCAT radiomics on PCD-CT discriminates hypercholesterolemia, suggesting a non-invasive biomarker
4. Mundt et al. ²⁸ 2023	<i>BMC Med Imaging.</i>	Retrospective radiomics association	55 pts	Unenhanced first-gen PCD-CT chest	Radiomics “glcm_ClusterProminence” differentiates Agatston 0 vs. ≥ 100 ; periaortic adipose texture and coronary calcium	Periaortic adipose radiomics on PCD-CT correlates with coronary calcium, hinting at inflammatory biomarker value
5. Pinos et al. ⁶⁵ 2023	<i>Eur J Radiol.</i>	Prospective intra-individual crossover	20 pts	PCD-CT polychromatic + VMI of 40–70 keV vs. EID-CT	Coronary CNR and subjective sharpness higher for all PCD-CT reconstructions (especially BMI > 30 kg/m ²)	PCD-CT significantly improves CCTA image quality over EID-CT, with greatest gain in pts with obesity
6. Rajiah et al. ¹⁷ 2023	<i>Invest Radiol.</i>	Prospective two-cohort evaluation	27 pts undergoing low-contrast + 26 pts undergoing routine-contrast	High-pitch (3.2) multi-energy PCD-CT CCTA; 30 vs. 60 mL iohexol; VMIs of 50 and 100 keV	CT numbers increased to 96% at 50 keV vs. 120 kV; CNR increased ($P < 0.0001$); CTDI of 2.5 vs. 3.1 mGy; CAD-RADS changed in 9 pts	High-pitch ME PCD-CT yields diagnostic CCTA at half the contrast dose, with VMI boosting CNR and reader confidence

Table 1. Continued

Author	Journal	Study design	Cohort	Imaging	Results	Conclusion
7. Rotkopf et al. ⁶⁶ 2023	<i>Int J Cardiovasc Imaging.</i>	Retrospective feasibility analysis (FLASH mode)	73 pts	Ultra-fast high-pitch PCD-CT coronary CTA	HRV strongly predicted image quality ($P < 0.001$); HR alone not independent	In FLASH PCD-CT CCTA, HRV—not absolute heart rate—is the key driver of motion artefacts and image degradation
8. Vattay et al. ²¹ 2023	<i>Eur Radiol.</i>	Cross-sectional VMI energy sweep	51 plaques/51 pts	Dual-source PCD-CT; VMIs of 40–180 keV (10 keV steps) vs. 120 kVp polychromatic	CP volume fell and LAP volume rose monotonically with keV; HU and CNR at 70 keV matched 120 kVp; all VMI CP volumes $\neq$ T3D ($P < 0.05$)	VMI energy strongly alters quantified plaque components—standardized thresholds are required for cross-study comparability
9. Vecsey-Nagy et al. ²³ 2023	<i>J Cardiovasc Comput Tomogr.</i>	Retrospective diagnostic agreement	197 pts	PCD-CT VNI CCTA at 120/140 kV vs. true non-contrast	18.3% false-zero reclassification; higher kVp decreased deviation ($\beta = -0.21$, $P = 0.020$); low-density CAC increased error	VNI CAC acceptable overall but misses low-density lesions; 140 kV improves agreement
10. Araki et al. ¹⁵ 2024	<i>Eur Heart J Imaging Methods Pract.</i>	Prospective observational	40 pts	Dual-source PCD-CT; 70 kVp, high-pitch helical; recon kernels Bv40 and Bv64	CTDIvol of 1.72 ± 0.38 mGy, DLP of 29.1 ± 6.8 mGy·cm, and an effective dose of 0.41 ± 0.09 mSv; >95% segments “good” IQ; sharp kernel altered stenosis grading ($P < 0.001$)	Ultra-low-dose (≈ 0.4 mSv) PCD-CT CCTA feasible; sharp kernels maintain IQ and refine stenosis evaluation
11. Cundari et al. ¹⁸ 2024	<i>Acad Radiol.</i>	Prospective three-arm CM-reduction study with VMI	100 pts (groups 1–3)	Dual-source PCD-CT CCTA; 45 keV VMI; standard CM vs. -20% and -40% volume	At -20% CM: attenuation is 890 HU, CNR is 26, IQ equals standard; at -40% CM: attenuation 676 is HU, CNR is 21, and all but 1 scan diagnostic	High-contrast VMI on PCD-CT permits up to 40% contrast-media reduction while retaining diagnostic CCTA quality
12. Kahmann et al. ³⁰ 2024	<i>Insights Imaging.</i>	Retrospective single-center radiomics study	61 pts and 306 plaques	First-gen PCD-CT, manual plaque segmentation, pyradiomics	Plaques with HR features/ relevant stenosis showed higher heterogeneity (multiple texture features) and higher EAT density (P significant)	Plaque-texture radiomics plus elevated EAT density on PCD-CT improve identification of patients at elevated cardiac risk
13. Haag et al. ²² 2024	<i>Radiology.</i>	Retrospective single-center comparison of PureCalcium VNC vs. true non-contrast CACS	170 pts (contrast-enhanced PCD-CT CCTA + TNC)	Dual-source PCD-CT CCTA with PureCalcium VNC reconstruction; TNC acquired separately	ICC of 0.98 (95% CI: 0.97–0.99); $\kappa = 0.88$; 74% correctly classified; median Agatston scores of 4.8 vs. 2.7 ($P = 0.99$); TNC added 19.7% $\pm$ 8.8% dose	PureCalcium VNC can replace TNC for Agatston scoring, cutting radiation while maintaining plaque-burden class accuracy

Table 1. Continued

Author	Journal	Study design	Cohort	Imaging	Results	Conclusion
14. McCollough et al. ⁶⁷ 2024	<i>J Comput Assist Tomogr.</i>	Prospective intra-individual comparison	21 pts (42–46 arteries)	Dual-source PCD-CT CAC exam; 3- and 1-mm axial (Qr36/Qr56 QIR4); dose of 2.1 ± 0.6 mGy	PCD-CT Agatston & volume increased to 5–10 mm ³ /approximately 8–12 points vs. EID ($P \leq 0.001$); dose 13% lower	PCD-CT yields slightly higher CAC scores while cutting the dose by 13% compared with EID-CT
15. Fink et al. ⁶⁸ 2024	<i>Radiol Cardiothorac Imaging.</i>	Prospective phantom + participant study	63 pts + phantom	Contrast CCTA with VNC at 55–80 keV; QIR of 1–4; safety-net recon at 55 keV/QIR2/110 HU	Safety-net recon detected 89% (50/56) previously missed subtle plaques; ICC VNC vs. TNC improved to 0.51–0.61	Safety-net VNC recon settings rescue detection of small, low-density plaques that standard VNC would miss
16. Kahmann et al. ³⁰ 2024	<i>Front Cardiovasc Med.</i>	Retrospective radiomics analysis (PCAT vs. CAD)	36 pts	Manual LAD and RCA PCAT segmentation on PCD-CT; pyradiomics	“original_glszm_GrayLevelNonUniformity” lower in CAD-LAD (155.2) vs. non-CAD-LAD (163.2) and CAD-RCA (189.1); strongest differentiator by random-forest	PCD-CT-based PCAT texture distinguishes CAD pts, even in non-stenotic RCA, indicating systemic adipose alterations
17. Mundt et al. ⁶⁹ 2024	<i>Diagnostics.</i>	Retrospective radiomics study	53 pts	Unenhanced PCD-CT; manual EAT segmentation; pyradiomics	Four texture features (GLRLM, GLSZM) differed between Agatston scores of 0 vs. 1–99/ ≥ 100 groups	PCD-CT-based EAT texture features associate with coronary calcium, offering a potential biomarker of perivascular inflammation
18. Sharma et al. ²⁴ 2024	<i>Eur Radiol.</i>	Cross-sectional VNC/VNI vs. true non-contrast	88 pts	PCD-CT CCTA with VNC and VNI recon	VNC median CAC of 8.5 vs. TNC of 27.8 (underestimate); VNI of 72.2 (over-estimate); mis-classification in 55% vs. 32%	VNI recon classifies calcium risk better than VNC but still needs optimization before replacing TNC scan
19. van der Bie et al. ⁷⁰ 2024	<i>Med Phys.</i>	Retrospective dose audit + phantom IQ experiment	143 pts + size-varying phantom	PCD-CT at 120 kV, 0.4 mm, kernels Bv40–56, QIR 3/4; mono-E at 40–55 keV tested	Large-patient phantom: PCD-CT detectability index (d') increased by 31% vs. EID-CT; small/medium: EID increased by 7%–17% unless low-keV PCD-CT narrows gap to 1%–6%	Thin-slice PCD-CT at 120 kV maintains vessel detectability for large pts without dose penalty; low-keV recon may offset IQ gap in smaller pts
20. Wang et al. ⁷¹ 2024	<i>Int J Cardiovasc Imaging.</i>	Phantom + pilot patient study	12 pts + CACS phantom	Flash and sequence modes; 90/120/140 kV and Sn100/140 kV; recon at 70 keV	ICC > 0.99 in phantom, >0.98 in pts; RMSE of 5.4–11.5; dose decreased by 16%–75% (flash); no CACS category change	PCD-CT provides accurate, repeatable calcium scoring with up to 75% dose reduction; flash Sn100/90/120 kV IQ20 advised
21. Vattay et al. ²⁰ 2024	<i>Eur J Radiol.</i>	Prospective protocol comparison	45 pts	Standard (0.4/0.6 mm, Bv40/Bv44, QIR of 0–4) vs. UHR (0.2/0.4 mm, Bv44/Bv56, QIR 0–4) PCD-CT	Best CNR 25.8 ± 4.1 at 0.6 mm Bv40 QIR4; worst was 8.3 ± 1.6 at 0.4 mm Bv44 QIR0 ($P < 0.001$); highest subjective IQ on Bv44 QIR3/4 (std) and Bv56 QIR4 0.2 mm (UHR)	Tailoring slice-thickness, kernel and QIR leverages PCD-CT to maximize CCTA quality—UHR is feasible without sacrificing distal vessel detail

Table 1. Continued						
Author	Journal	Study design	Cohort	Imaging	Results	Conclusion
22. Ayx et al. ²⁶ 2025	<i>Eur J Radiol.</i>	Single-center PCD-CT CT-FFR vs. ICA	28 pts	Dual-source PCD-CT CCTA; Δ CT-FFR (-1.8/+1.8 cm) threshold of ≥ 0.06	Δ CT-FFR: PPV of 66.7%, NPV of 100%, accuracy of 74%; could have avoided ICA in 39%	On-site Δ CT-FFR with PCD-CT reliably flags hemodynamically significant stenoses and may reduce unnecessary catheterization
23. Brendel et al. ²⁷ 2025	<i>Diagn Interv Imaging.</i>	Retrospective AI vs. expert	140 pts	Non-UHR PCD-CT; DL models CorEx and Spimed-AI	Patient-level: sensitivity of 97%, specificity of 82%, NPV of 99%, AUC of 0.90; vessel-level AUC of 0.92	Deep-learning on routine-res PCD-CT detects $\geq 50\%$ CAD with near-expert accuracy
24. Fahrni et al. ⁷² 2025	<i>Invest Radiol.</i>	Prospective paired CCTA study against invasive angiography	26 pts at very high risk (26 stenoses)	UHR spectral PCD-CT CCTA plus standard CT, both < 3 days before ICA	Mean error of 6% PCD-CT vs. 12% CT; sensitivity/specificity. 100%/90% vs. 75%/50%; 38% lesions reclassified by PCD-CT vs. 4% by CT	UHR SPCD-CT quantifies stenosis far more accurately than conventional CT, reclassifying CAD severity in > 1/3 cases
25. Fahrni et al. ²⁵ 2025	<i>Eur Radiol.</i>	Prospective paired PCD-CT vs. DECT	26 high-risk pts	PCD-CT in UHR mode; 40 vs. 70 keV VMIs	Overall IQ score of 5 vs. 4 ($P < 0.001$); stenosis bias of -1% vs. -6% DECT; blooming unchanged (+2% vs. +7% DECT)	UHR 40 keV PCD-CT VMIs outperform DECT at both energies, giving sharper, more consistent coronary stenosis measurements
26. Kaldas et al. ²⁵ 2025	<i>J Clin Med.</i>	Cross-sectional TNC vs. VNC	77 pts	PCD-CT recon: TNC at 3×3 and 3×1.5 mm, VNC at 3×3 and 3×1.5 mm	ICC of 0.97–0.99 across methods; κ of 0.94 for thin-slice TNC vs. 0.83–0.85 for VNC; slice/recon choices altered mean scores and mis-classification rate	VNC CACS shows excellent agreement but needs adjusted cut-offs; 1.5 slices minimize category errors
27. Kotronias et al. ¹¹ 2025	<i>JACC Cardiovasc Imaging.</i>	Prospective ICA-validated	100 pts/257 vessels	UHR (0.2 mm) and multi-energy standard-resolution PCD-CTA vs. 3D QCA	Median luminal diff. of 3% (UHR) vs. 6% (SR, $P < 0.001$); per-vessel AUC of 0.99 vs. 0.94 ($\Delta 0.05$, $P = 0.01$); κ of 0.90 vs. 0.63 for CAD-RADS	UHR PCD-CTA closely matches 3D QCA and outperforms standard PCD-CT, especially in severe calcification
28. Sakai et al. ¹³ 2025	<i>J Am Coll Cardiol.</i>	Retrospective device-comparative cohort	7,833 pts (3,876 PCD-CT vs. 3,957 EID)	Routine CCTA on NAEOTOM Alpha PCD-CT scanner	Specificity of 98% vs. 93%; PPV of 83.3% vs. 63%; accuracy of 97.2% vs. 92.8%; ICA referral of 9.9% vs. 13.1%; revasc. of 43.4% vs. 35.5%	PCD-CT CCTA increases specificity/PPV and reduces unnecessary catheterization while guiding revascularization
29. Sartoretti et al. ⁷³ 2025	<i>Int J Cardiovasc Imaging.</i>	Retrospective temporal-resolution study	70 pts	ECG-gated non-contrast PCD-CT; reconstructions of 66 vs. 125 ms	CAC, AVC, MAC smaller at 66 ms ($P < 0.001$); CAC category re-class in 4%; blur artifacts decrease at 66 ms	Higher temporal resolution (66 ms) on PCD-CT curbs motion blur and prevents calcium-score overestimation
30. Vecsey-Nagy et al. ³¹ 2025	<i>J Cardiovasc Comput Tomogr.</i>	Monte Carlo cost-effectiveness model	55 coronary lesions (diagnostic dataset); simulated 15,000 pts	UHR PCD-CT CCTA (UHR) vs. EID-CT	Projected decrease by 18.9% in downstream functional tests, 6% decrease in ICAs, 9.4% decrease in major complications; cost saving $\approx$ USD 795/patient, USD 11.9 M overall	UHR PCD-CT CCTA saves costs by reducing follow-up tests and complications in stable chest pain evaluation
31. Zhao et al. ¹⁶ 2025	<i>Quant Imaging Med Surg.</i>	Prospective low-dose vs. standard CAC-CT	105 pts	Low-dose chest PCD-CT; 3 and 1.5 mm; QIR 0–4	Sensitivity of 96%–100%, specificity of 100%; ICC of 0.983–0.993; dose reduction of 56% (1.0 vs. 2.3 mGy, $P < 0.001$); best agreement at LD 1.5 mm-QIR1	Low-dose 1.5-mm QIR1 PCD-CT chest scan delivers accurate CACS while halving radiation

PCD-CT, photon-counting computed tomography; UHR, ultra-high resolution; HR, high resolution; VMIs, virtual monoenergetic images; DECT, dual-energy computed tomography; CCTA, coronary computed tomography angiography; ECG, electrocardiogram; AUC, area under the curve; pts, patients; PCAT, pericoronary adipose tissue; LCA, left coronary artery; RCA, right coronary artery; VMI, virtual monoenergetic imaging; EID-CT, energy-integrating detector computed tomography; CNR, contrast-to-noise ratio; BMI, body mass index; CTDI, computed tomography dose index; mGy, milligray; CAD-RADS, coronary artery disease reporting and data system; ME, multi-energy; HRV, heart rate variability; kVp, kilovolt peak; CP, calcified plaque; LAP, low-attenuation plaque; HU, Hounsfield unit; T3D, three-dimensional threshold-based segmentation; VNI, virtual non-iodine; CAC, coronary artery calcium; TNC, true non-contrast; CTDIvol, computed tomography dose index volume; DLP, dose-length product; mSv, millisievert; IQ, image quality; CM, contrast media; EAT, epicardial adipose tissue; ICC, intraclass correlation coefficient; CI, confidence interval; CACS, coronary artery calcium scoring; RMSE, root mean square error; ICA, invasive coronary angiography; FFR, fractional flow reserve; Δ CT-FFR, change in computed tomography fractional flow reserve; PPV, positive predictive value; NPV, negative predictive value; DL, deep learning; AVC, aortic valve calcium; MAC, mitral annular calcium; ms, millisecond; LD, low dose; QIR, quantum iterative reconstruction; 3D QCA, three-dimensional quantitative coronary angiography; revasc., revascularization.

Low-dose protocols achieved diagnostic image quality at a CTDIvol of 1.72 mGy (DLP:29.1 mGy·cm; effective dose:0.41 mSv), with >95% of segments rated diagnostic.¹⁵ A prospective comparison for calcium scoring showed a 56% dose reduction (1.0 vs. 2.3 mGy) with 1.5-mm slices and low-strength iterative reconstruction.¹⁶ Contrast-saving protocols delivered diagnostic CCTA with 30 mL of iodinated agent using 50-keV reconstructions,¹⁷ and preserved image quality after 40% contrast reduction using 45-keV reconstructions.¹⁸ Kernel and iterative-reconstruction optimization maintained a coronary CNR with finer matrices.^{19,20} An energy-sweep analysis (40–180 keV) documented systematic shifts in quantified plaque components across keV levels.²¹

Virtual non-contrast (VNC) approaches for calcium scoring showed high agreement with true non-contrast (TNC) scans [intraclass correlation coefficient (ICC): 0.98; kappa (κ): 0.88] and contributed to dose reduction.²² Additional evaluations reported risk-category misclassification for low-density calcium,^{23,24} with improved agreement using thin-slice reconstructions and higher tube potentials.²⁵

Functional and data-driven analyses included on-site CT fractional flow reserve (FFR) with 100% negative-predictive value and a projected reduction of invasive angiography in 39% of cases;²⁶ deep-learning stenosis detection with a vessel-level AUC of 0.92;²⁷ and radiomics of pericoronary adipose tissue and plaque texture differentiating hypercholesterolemia and coronary disease status.^{28–30}

A Monte Carlo economic model estimated reductions in downstream functional testing (18.9%), invasive angiography (6.0%), and major complications (9.4%), with an approximate cost saving of USD 800 per patient in stable chest pain pathways.³¹

Figure 2 presents the CCTA of a patient with a stent in the left anterior descending artery obtained with PCD-CT in our center.

Myocardial tissue characterization

Four clinical studies evaluated PCD-CT against cardiovascular magnetic resonance imaging (MRI) for late iodine enhancement (LIE) and/or extracellular volume (ECV) mapping (Table 2). In this review, LIE denotes CT-based delayed-phase iodine-related myocardial hyperenhancement acquired 5–10 minutes after iodinated contrast and quantified on iodine maps/VMLs; it is analogous to—but distinct from—cardiovascular magnetic resonance (CMR) late gadolinium enhancement (LGE).

In a diagnostic accuracy study of 27 patients (459 myocardial segments), Tremamunno et al.³² used dual-source PCD-CT with electrocardiogram (ECG)-triggered sequential acquisition 5 minutes post-contrast (120 kVp; 144 × 0.4 mm collimation; iodine maps reconstructed with Qr40 and iterative reconstruction). For two readers, per-patient sensitivity was 100% and 91.7%, specificity was 73.3% and 80.0%, and accuracy was 85.2%. Per-segment sensitivity was 74.7% and 66.7%, specificity was 94.9% and 96.4%, and accuracy was 91%. Inter-reader agreement was $\kappa = 0.70$ at the patient level and $\kappa = 0.63$ at the segment level.³²

In a prospective series of 17 patients (24 CT/MRI pairs), Klambauer et al.³³ performed spectral dual-source PCD-CT with a 5-minute delayed LIE and atlas-based ECV mapping. Agreement with LGE-MRI was $\kappa = 0.832$ in the acute setting; agreement with combined LGE + edema was $\kappa = 0.944$; and at 3-month follow-up, $\kappa = 0.956$.³³

In 30 patients with systemic amyloidosis, Popp et al.³⁴ used first-generation PCD-CT with CCTA and a delayed phase. Global ECV was $42.93\% \pm 10.14\%$ (CMR), $42.51\% \pm 9.07\%$ [single energy (SE)], and $40.69\% \pm 9.24\%$ [dual energy (DE)]. Compared with CMR, SE showed a mean difference of 0.43% [95% confidence interval (CI): -1.83 to 2.68], whereas DE was -2.24% (95% CI: -4.42 to -0.06); DE vs. SE was -1.82% (95% CI: -2.70 to -0.94). Bland–Altman analysis: the mean bias for DE vs. CMR was -2.28% (limits of agreement: -11.16 to 6.59); for SE vs. CMR, it was -0.42% (-9.77 to 8.92); and for DE vs. SE, it was -1.82% (-5.46 to 1.83). Both CT approaches correlated strongly with CMR (r : 0.892 for DE; r : 0.882 for SE).³⁴

In a retrospective CT–MRI comparison of 32 patients with acute myocarditis, Gkizas et al.³⁵ reported a DLP of 96 ± 32 mGy·cm. The global ECV on PCD-CT was $29.4\% \pm 4.5\%$ and $30.0\% \pm 4.1\%$ on CMR ($P = 0.69$); correlation with LGE was $r = 0.82$, and the AUC for segment-level inflammation was 0.95 at a 26.9% threshold.³⁵

Structural heart and valvular assessment

Six clinical studies (Table 3) assessed PCD-CT across components of transcatheter aortic valve replacement/implantation (TAVR/TAVI) evaluation, including access planning, annulus sizing, valve-calcium quantification, and concomitant coronary assessment.

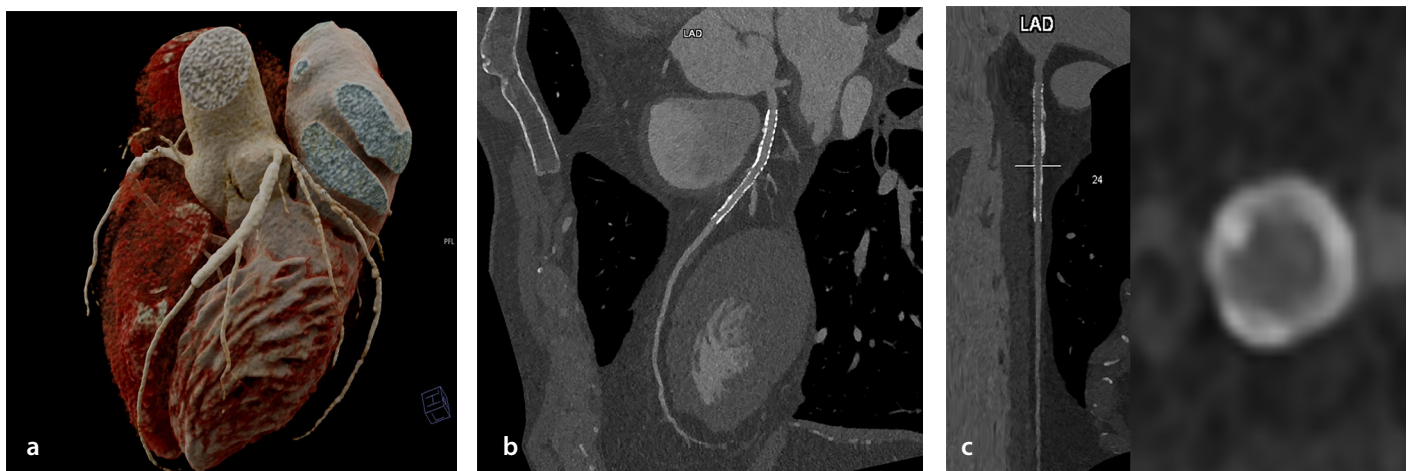


Figure 2. A 53-year-old male patient with a stent in the left anterior descending artery. Coronary computed tomography (CT) angiography was performed using a prospective ultra-high-resolution-mode electrocardiogram-triggered sequential acquisition, with 70 mL of iodinated contrast and a dose-length product of 604 mGy·cm. Three reconstructions are shown: cinematic volume-rendered image (a), curved planar reformation (b), and cross-sectional vessel analysis (c), all reconstructed at 0.2-mm slice thickness using a Bv72 kernel. Photon-counting CT enables high-resolution in-stent lumen visualization, clearly demonstrating stent patency.

Author	Journal	Study design	Cohort	Imaging	Results	Conclusion
1. Gkizas et al. ³⁵ 2025	<i>Diagn Interv Imaging.</i>	Retrospective CT–MRI comparison	32 pts	Late iodine enhancement PCD–CT (DL product: 96 ± 32 mGy-cm); ECV maps	Global ECV PCD–CT of $29.4\% \pm 4.5\%$ vs. MRI of $30.0\% \pm 4.1\%$ ($P = 0.69$); r : 0.82 with LGE; AUC of 0.95 at a 26.9% threshold	PCD–CT iodine-derived ECV quantifies acute myocarditis accurately at a low dose, matching MRI
2. Klambauer et al. ³³ 2025	<i>Invest Radiol.</i>	Prospective CT–MRI comparison	17 pts (24 CT/MRI pairs)	Spectral dual-source PCD–CT LE (5 min delay); atlas-map ECV	Acute κ : 0.832 (CT vs. LGE MRI), κ : 0.944 vs. LGE + edema; follow-up κ : 0.956	PCD–CT late-enhancement closely matches MRI for SCAD, aided by novel atlas ECV maps
3. Popp et al. ³⁴ 2025	<i>Invest Radiol.</i>	Cohort comparison with CMR reference	30 pts cardiac amyloidosis	First-gen PCD–CT CCTA + delayed phase; single-energy vs. dual-energy ECV methods	SE-PCD–CT ECV not different from CMR (mean diff: 0.43 , $P = 1.00$); DE-PCD–CT slightly lower (-2.24 , $P = 0.04$); both $r \approx 0.89$ vs. CMR	PCD–CT simultaneously quantifies myocardial ECV (in tight agreement with CMR) and extensive CAD burden in amyloidosis
4. Tremamunno et al. ³² 2025	<i>Eur Radiol.</i>	Retrospective analysis of a prospective cohort; diagnostic-accuracy study vs. LGE-MRI	27 pts and 459 myocardial segments	Dual-source PCD–CT, ECG-triggered sequential (5 min post-contrast); 120 kVp, 144×0.4 -mm collimation; iodine maps reconstructed with Qr40 kernel + iterative recon	Per-patient: sens of $100\%/91.7\%$, spec of $73.3\%/80.0\%$, acc of 85.2% • Per-segment: sens of $74.7\%/66.7\%$, spec of $94.9\%/96.4\%$, acc of 91% • Inter-reader κ : 0.70 (patient), 0.63 (segment)	PCD–CT iodine maps deliver high accuracy and substantial agreement for detecting myocardial late enhancement, suggesting a viable alternative when MRI is contraindicated

CT, computed tomography; MRI, magnetic resonance imaging; pts, patients; PCD–CT, photon-counting detector computed tomography; DL, dose-length; mGy-cm, milligray-centimeter; ECV, extracellular volume; LGE, late gadolinium enhancement; AUC, area under the curve; LE, late enhancement; κ , kappa coefficient; SCAD, spontaneous coronary artery dissection; CMR, cardiovascular magnetic resonance; SE-PCD–CT, single-energy photon-counting computed tomography; DE-PCD–CT, dual-energy photon-counting computed tomography; CCTA, coronary computed tomography angiography; r , correlation coefficient; acc, accuracy; spec, specificity; sens, sensitivity; kVp, kilovolt peak; mm, millimeter; Qr, quantum reconstruction.

Author	Journal	Study Design	Cohort	Imaging	Results	Conclusion
1. Brendel et al. ⁴⁰ 2024	<i>Diagn Interv Imaging.</i>	Retrospective, single-center, paired PC–CCTA $\pm$ FFR vs. ICA in TAVR work-up	260 pts	Dual-source PCD–CT CCTA, AI stenosis (CorEx) and FFR (Spimed-AI)	PC–CCTA $\geq 50\%$: sens of 96.0% , spec of 68.7% , AUC of 0.82 ; FFR ≤ 0.8 : sens of 96.8% , spec of 87.3% , AUC of 0.92 ; FFR could avoid ICA in 46.5% vs. 37.3%	AI-derived FFR on PCD–CT outperforms diameter stenosis, reducing invasive angiography during TAVR assessment
2. Dirrachs et al. ³⁶ 2024	<i>Acad Radiol.</i>	Retrospective comparative cohort	300 pts (202 PCD–CT, 100 DSCT)	Aorto-ilio-femoral contrast CT for TAVI; PCD–CT vs. dual-source EID–CT	SNR of 33 ± 10.5 vs. 47.3 ± 16.4 ; CNR of 47.3 ± 14.8 vs. 59.3 ± 21.9 (PCD–CT < DSCT, $P < 0.001$), yet visual quality of 4.8 vs. 3.3 ($P < 0.001$); suitability of 99% vs. 85% ; eRD of 8.8 ± 4.5 vs. 15.3 ± 5.8 mSv	Despite lower raw SNR/CNR, PCD–CT markedly improves subjective quality and halves radiation for TAVI planning
3. Feldle et al. ³⁹ 2024	<i>Sci Rep.</i>	Retrospective diagnostic-accuracy study VNI vs. TNC	123 pts (56 with AV calcification)	ECG-gated cardiac PCD–CT; 70 keV VNI vs. 70 keV TNC reference	Sens/Spec/Acc of $69\%/100\%/85\%$ (prospective gating); r : 0.983 – 0.986 with TNC ($P < 0.001$)	VNI PCD–CT accurately scores aortic-valve calcification, suggesting TNC can be omitted in AV calcification work-up
4. Hagar et al. ³⁸ 2024	<i>Int J Cardiovasc Imaging.</i>	Retrospective paired-scan comparison (pre-TAVR)	64 pts	Dual-source PCD–CT: UHR–CTA (120×0.2 mm) vs. high-pitch spiral CTA (144×0.4 mm)	Dose of 12.6 (UHR) vs. 4.1 mSv (HPS); annulus IQ median of 4 vs. 3 ($P < 0.001$); AAA strongly correlated (r^2 : 0.857); 91% identical valve sizing	Both PCD–CT modes give reliable aortic-annulus sizing; HPS halves the dose but inferior IQ may mis-size valves when image quality drops

Table 3. Continued

Author	Journal	Study Design	Cohort	Imaging	Results	Conclusion
5. Sharma et al. ⁴¹ 2024	<i>Int J Cardiovasc Imaging.</i>	Pre-TAVI HR vs. UHR vs. adj-UHR PCD-CT vs. QCA	60 pts	Dual-source PCD-CT: HR of 120 kV, UHR of 120 kV, adj-UHR of 90 kV/ IQ65	Per-pt AUC: HR of 0.57, UHR of 0.80, adj-UHR of 0.80; per-vessel AUC of 0.73, 0.69, 0.87 (UHR vs. adj-UHR $P = 0.04$)	Low-kV adjusted UHR PCD-CT boosts vessel-level CAD detection during TAVI planning without extra dose
6. Yang et al. ³⁷ 2024	<i>Eur J Radiol.</i>	Prospective ECG-gated high-pitch vs. reference	30 prospective + 30 matched controls	High-pitch PCD-CT ECG-gated (30% R-R); annulus sizing	Strong correlation with spiral CT ($r \geq 0.94$; $P \geq 0.09$); CTDI of 4.52 vs. 24.10 mGy ($P < 0.001$); systolic capture of 90% vs. 50%	ECG-gated high-pitch PCD-CT accurately sizes aortic annulus while slashing dose five-fold

PCD-CT, photon-counting computed tomography; PC-CCTA, photon-counting coronary computed tomography angiography; FFR, fractional flow reserve; ICA, invasive coronary angiography; TAVR, transcatheter aortic valve replacement; AI, artificial intelligence; CCTA, coronary computed tomography angiography; DSCT, dual-source computed tomography; CT, computed tomography; TAVI, transcatheter aortic valve implantation; EID-CT, energy-integrating detector computed tomography; SNR, signal-to-noise ratio; CNR, contrast-to-noise ratio; eRD, effective radiation dose; mSv, millisievert; VNI, virtual non-iodine; TNC, true non-contrast; AV, aortic valve; ECG, electrocardiogram; keV, kiloelectronvolt; UHR, ultra-high resolution; mm, millimeter; HPS, high-pitch spiral; IQ, image quality; AAA, aortic annulus area; r^2 , coefficient of determination; HR, high resolution; adj-UHR, adjusted ultra-high resolution; QCA, quantitative coronary angiography; kV, kilovolt; IQ65, image quality index 65; R-R, R-R interval (cardiac cycle); CTDI, computed tomography dose index.

In a retrospective comparative cohort of 300 patients (202 PCD-CT; 100 dual-source EID-CT), Dirrachs et al.³⁶ performed aorto-ilio-femoral contrast CT for TAVI. The SNR was 33 ± 10.5 vs. 47.3 ± 16.4 , and the CNR was 47.3 ± 14.8 vs. 59.3 ± 21.9 (both $P < 0.001$). The visual image-quality scores were 4.8 vs. 3.3 ($P < 0.001$), suitability for TAVI planning was 99% vs. 85%, and the effective radiation dose was 8.8 ± 4.5 vs. 15.3 ± 5.8 mSv.³⁶

In a prospective study with 30 patients (plus 30 matched controls), Yang et al.³⁷ implemented ECG-gated high-pitch PCD-CT at 30% R-R for annulus sizing. Correlation with spiral CT was strong ($r \geq 0.94$). Mean paired differences (bias) with 95% CIs between high-pitch PCD-CT and spiral CT were 0.16 mm (-0.10, 0.42) for mean diameter, 0.22 mm (-0.70, 1.13) for the perimeter, and 5.35 mm² (-22.02, 32.72) for the annular area; Bland-Altman plots showed minimal bias. Additionally, CTDI was 4.52 vs. 24.10 mGy ($P < 0.001$); systolic capture occurred in 90% vs. 50%.³⁷

In a paired-scan comparison of 64 pre-TAVR candidates, Hagar et al.³⁸ compared UHR PCD-CT CTA (120 × 0.2 mm) with high-pitch spiral CTA (144 × 0.4 mm). The effective dose was 12.6 (UHR) vs. 4.1 mSv (high pitch); annulus image-quality scores were median 4 vs. 3 ($P < 0.001$). Area-derived annulus measurements were highly correlated (r^2 : 0.857), and prosthesis size selection was identical in 91% of patients; the distribution of ± 1 size and $\geq \pm 2$ sizes was not reported in the paper.³⁸

In 123 patients (56 with aortic-valve calcification), Feldle et al.³⁹ evaluated ECG-gated cardiac PCD-CT with 70 keV virtual non-iodine (VNI) images against 70 keV TNC. Sensitivity, specificity, and accuracy were 69%, 100%, and 85%, respectively, under prospec-

tive gating. Correlation between VNI-derived and TNC Agatston scores was r : 0.983–0.986 ($P < 0.001$).³⁹

In a retrospective single-center cohort of 260 patients during TAVR workup, Brendel et al.⁴⁰ performed dual-source PCD-CT CCTA with artificial intelligence (AI) stenosis quantification (CorEx) and AI-derived FFR (Spimed-AI), both referenced to ICA. For $\geq 50\%$ stenosis, sensitivity was 96.0%, specificity was 68.7%, and the AUC was 0.82. For $\text{FFR} \leq 0.80$, sensitivity was 96.8%, specificity was 87.3%, and the AUC was 0.92. Decision analysis indicated that 46.5% vs. 37.3% of cases were classified as not requiring ICA with AI-based FFR vs. diameter-based stenosis.⁴⁰

During TAVI planning in 60 patients, Sharma et al.⁴¹ compared HR (120 kV), UHR (120 kV), and adjusted UHR (90 kV, IQ65) PCD-CT protocols against quantitative coronary angiography. Per-patient AUCs were 0.57 (HR), 0.80 (UHR), and 0.80 (adjusted UHR); per-vessel AUCs were 0.73, 0.69, and 0.87, respectively (UHR vs. adjusted UHR, $P = 0.04$).⁴¹

Figure 3 presents the PCD-CT angiography for a patient undergoing evaluation for TAVR.

Pulmonary and cardio-pulmonary functional imaging

Four clinical studies (Table 4; $n = 447$ patients) evaluated PCD-CT for pulmonary and cardiopulmonary functional assessment. Scharm et al.⁴² acquired contrast-enhanced inspiratory PCD-CT and expiratory PCD-CT

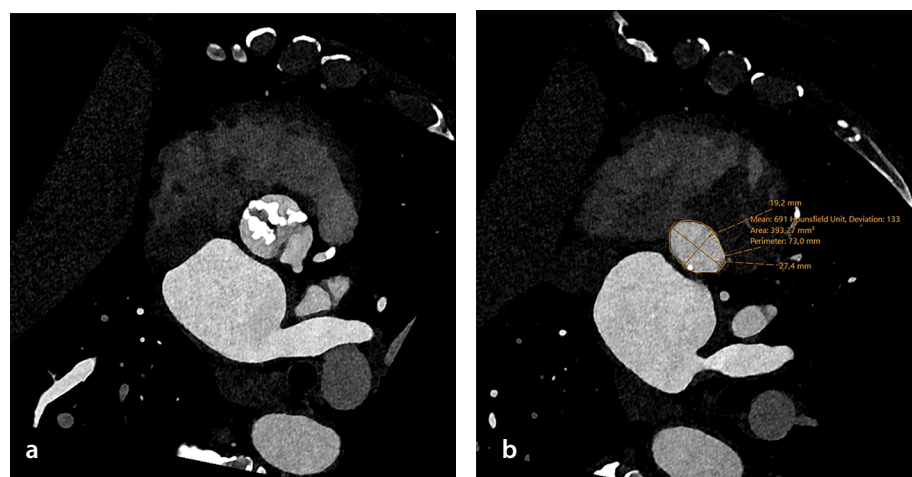


Figure 3. Photon-counting detector computed tomography angiography in a 78-year-old male patient undergoing evaluation for transcatheter aortic valve replacement. The scan was acquired using a retrospective low-pitch electrocardiogram-triggered protocol with 75 mL of iodinated contrast medium, a dose-length product of 1.928 mGy-cm, and a standard 0.4-mm slice thickness. Panel (a) shows an image reconstructed at 30% of the R-R interval, capturing the aortic valve in systole—the optimal phase for annular sizing. Panel (b) demonstrates quantitative measurements of the aortic annulus, including diameter, perimeter, and area, used to guide prosthesis selection.

5 minutes later in 166 patients, producing technically successful functional maps in 84.7% of patients; mean enhancement values were 325 HU for the pulmonary trunk, 260 HU for the left atrium, and 252 HU for the aorta, with per-phase dose indices of approximately CTDI: 3 mGy and DLP: 110 mGy-cm.

Kerber et al.⁴³ performed a retrospective comparison in 26 patients who received both PCD-CT iodine maps and single photon emission computed tomography (SPECT)/CT for suspected or known chronic thromboembolic pulmonary hypertension (CTEPH). Using multidisciplinary clinical diagnosis as the reference standard and per-patient CTEPH classification, two blinded PCD-CT readers achieved accuracies of 0.85 (95% CI: 0.66–0.94) and 0.88 (0.71–0.96), with sensitivities of 0.90 (0.60–0.98) and 0.90 (0.60–0.98) and specificities of 0.81 (0.57–0.93) and 0.88 (0.65–0.97). The SPECT/CT consensus had an accuracy of 0.73 (0.54–0.86), sensitivity of 0.80 (0.49–0.94), and specificity of 0.69 (0.44–0.86), and between-modality differences were not significant ($P > 0.688$). The lobar-level perfusion-defect extent on PCD-CT showed moderate correlations with right-heart catheter measures, and the dose was markedly lower with PCD-CT (1.19 ± 0.33 mSv) than with SPECT/CT (6.34 ± 1.68 mSv).⁴³

Saeed et al.⁴⁴ conducted a retrospective dose-reduction series of 105 patients undergoing high-pitch fast low-angle shot PCD-CT

pulmonary CTA with 35, 45, or 60 mL of contrast. Subjective image quality was 4.6 vs. 4.1 for 35 vs. 60 mL ($P < 0.001$), pulmonary-trunk attenuation was 320–347 HU, and all segmental arteries were assessable.⁴⁴

Yalon et al.⁴⁵ performed an evaluation of a three-arm comparative cohort ($n = 150$) of multi-energy high-pitch PCD-CT pulmonary CTA compared with high-pitch and routine dual-source DE-CT, reporting a CTDIvol of 8.1 vs. 9.6/16.2 mGy and a CNR ($P < 0.001$), a subjective artery-contrast score of 4.7/5 vs. 4.4/5 and 4.3/5, and fewer motion artifacts with PCD-CT.

Figure 4 illustrates the use of PCD-CT perfusion maps in a patient with a segmental pulmonary embolism.

Aortic, visceral, and peripheral arterial disease

Fourteen clinical studies (Table 5; $n = 851$ patients) evaluated PCD-CT across the thoracoabdominal aorta, renal-visceral branches, and lower-extremity run-off.

Euler et al.⁷ performed an intra-individual comparison ($n = 40$) at a matched dose using high-pitch PCD-CT with 40–55 keV VMIs and reported a CNR of 22 ± 7 at 40 keV vs. 17 ± 8 on EID-CT, with the greatest CNR gain in patients who were overweight; subjective noise increased at 40–45 keV, whereas overall image quality was similar.

Dillinger et al.⁴⁶ prospectively evaluated arterial-phase PCD-CT of the abdomen ($n = 20$) with VMI reconstructions from 40–190 keV: a cohort CTDIvol of 7.90 ± 3.92 mGy, a DLP of 330.6 ± 198.5 mGy-cm, and an effective dose of 4.92 ± 2.97 mSv (Radimetrics v3.4, ICRP-103 Monte Carlo; verification with k-factors of 0.015 for the abdomen and 0.014 for the chest). The CNR peaked at 60 keV and the SNR at 70 keV (no significant difference vs. 60 keV, $P = 0.294$), and subjective image quality was rated optimal at 70 keV. Acquisition and reconstruction settings were automatic 100–120 kVp, a pitch of 0.80, and 1-mm VMI in Qr40.⁴⁶

Hennes et al.⁴⁷ conducted an intra-individual comparison ($n = 57$) of ECG-triggered, high-pitch aortic CTA using PCD-CT [120 kVp; 144×0.4 mm; VMI 55 keV; kernel Bv36, quantum iterative reconstruction 3 (QIR-3)] and EID-CT with automatic tube voltage selection (ATVS) at 90/100 kVp [effective collimation: 192×0.6 mm; Bv36, advanced modeled iterative reconstruction 3 (ADMIRE-3)]. The CTDIvol was 3.95 ± 0.54 (PCD-CT) vs. 4.97 ± 0.57 mGy (EID-CT) ($P < 0.001$), and the SSDE was 4.88 ± 0.48 vs. 6.28 ± 0.50 mGy ($P < 0.001$); the DLP and effective dose were not reported. The CNR was 41.11 ± 8.68 vs. 27.05 ± 6.73 ($P < 0.001$), with higher overall image quality and luminal contrast on PCD-CT; vessel sharpness was similar, whereas blooming and beam hardening were less pronounced on EID-CT.⁴⁷

Table 4. Photon-counting CT studies related to pulmonary and cardio-pulmonary functional imaging						
Author	Journal	Study design	Cohort	Imaging	Results	Conclusion
1. Scharm et al. ⁴² 2023	<i>Radiology</i> .	Retrospective observational	166 pts (166 inspiratory + expiratory scans)	Contrast-enhanced inspiratory PCD-CT, expiratory PCD-CT 5 min later	84.7% success; mean HU of the pulmonary trunk 325, LA 260, Ao 252; DLP $\approx$ 110 mGy-cm/phase; CTDI $\approx$ 3 mGy; all functional metrics differed between 6 subgroups ($P < 0.05$)	Two-phase PCD-CT permits simultaneous morphology, ventilation and perfusion mapping at reference-level dose
2. Saeed et al. ⁴⁴ 2024	<i>Acad Radiol</i> .	Retrospective dose-reduction series	105 pts (35/45/60 mL CM)	High-pitch FLASH PCD-CT CTPA	Subjective quality of 4.6 vs. 4.1 (35 vs. 60 mL, $P < 0.001$); trunk HU of up to 320–347 (ns); all segmental arteries assessable	Diagnostic CTPA achievable with 35 mL contrast on PCD-CT without quality loss
3. Yalon et al. ⁴⁵ 2024	<i>J Comput Assist Tomogr</i> .	Three-arm comparative cohort (PCD-CT vs. high-pitch and routine DECT)	150 pts	High-pitch multi-energy PCD-CT pulmonary CTA	CTDIvol of 8.1 mGy vs. 9.6/16.2; highest CT number & CNR ($P < 0.001$); subjective artery contrast of 4.7 vs. 4.4/4.3; fewer motion artifacts	High-pitch PCD-CT pulmonary CTA delivers superior contrast, fewer artifacts, and lower dose than conventional dual-source protocols
4. Kerber et al. ⁴³ 2025	<i>Invest Radiol</i> .	Retrospective comparative	26 pts	Lung-perfusion PCD-CT iodine maps (mean dose = 1.19 ± 0.33 mSv)	Accuracy of 0.85/0.88 vs. 0.73 (SPECT); sensitivity of 0.90 vs. 0.80; specificity of 0.81/0.88 vs. 0.69; dose reduced to $\approx 5\times$ vs. SPECT	PCD-CT iodine maps detect CTEPH as accurately as SPECT at one-fifth the dose and correlate with hemodynamics
PCD-CT, photon-counting computed tomography; HU, Hounsfield unit; LA, left atrium; Ao, aorta; DLP, dose-length product; mGy-cm, milligray-centimeter; CTDI, computed tomography dose index; CM, contrast media; FLASH, fast low-angle shot; CTPA, computed tomography pulmonary angiography; ns, not significant; CTA, computed tomography angiography; CTDIvol, computed tomography dose index volume; CNR, contrast-to-noise ratio; DECT, dual-energy computed tomography; mSv, millisievert; SPECT, single photon emission computed tomography; CTEPH, chronic thromboembolic pulmonary hypertension.						

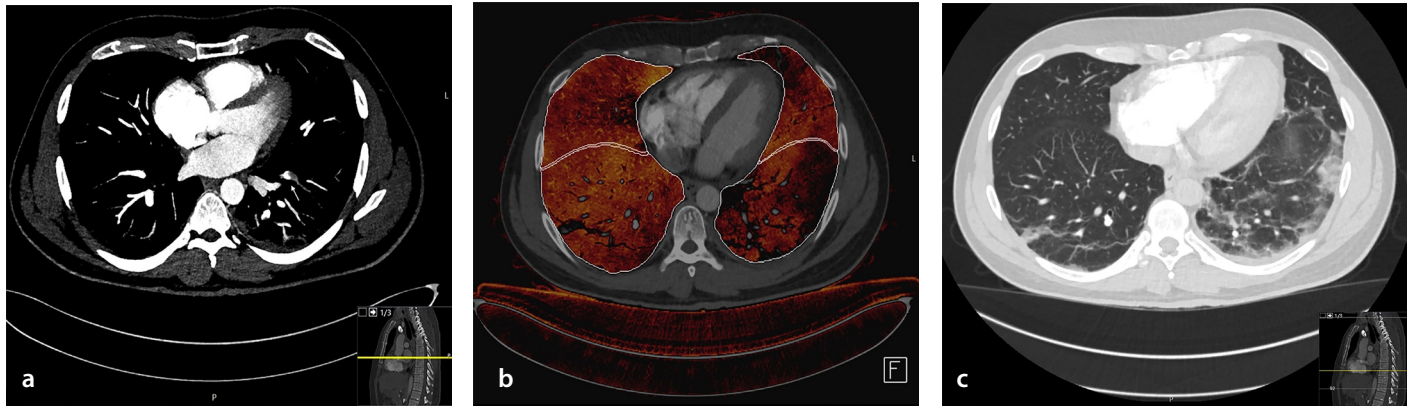


Figure 4. A 35-year-old male patient presenting with shortness of breath underwent pulmonary computed tomography (CT) angiography using photon-counting CT with 50 mL of iodinated contrast medium, a dose-length product of 256 mGy-cm, and a standard 0.4-mm slice thickness. The axial maximum intensity projection image (a) demonstrates an acute pulmonary embolism in the lateral segmental branch of the left lower lobe pulmonary artery (arrow). The corresponding axial image in the parenchymal window (b) shows increased attenuation in the left lower lobe lateral segment, compatible with pulmonary infarction. The iodine perfusion map (c), derived from pulmonary blood volume imaging, reveals a perfusion defect in the same segment. Iodine quantification was achieved by comparing attenuation on contrast-enhanced images with the respective virtual non-contrast reconstructions.

Table 5. Photon-counting CT studies related to aortic, visceral, and peripheral arterial diseases

Author	Journal	Study design	Cohort	Imaging	Results	Conclusion
1. Euler et al. ⁷ 2022	<i>Invest Radiol.</i>	Intra-individual (same-patient) comparison	40 pts	High-pitch PCD-CT at 120 kV; VMI of 40–55 keV; matched dose to EID-CT	CNR of 22 ± 7 (40 keV) vs. 17 ± 8 (EID); CNR gain greatest in pts who are overweight (+34%); subjective noise increased at 40–45 keV, overall quality similar	High-pitch PCD-CT aortic CTA with 40–45 keV VMI boosts CNR vs. matched-dose EID-CT, especially in patients who are overweight
2. Dillinger et al. ⁴⁶ 2023	<i>Diagnostics.</i>	Prospective single-arm observational	20 pts	Arterial-phase PCD-CT; VMIs of 40–110 keV	Peak CNR at 60 keV; peak SNR at 70 keV (ns vs. 60 keV $P = 0.294$); best subjective quality at 70 keV	60–70 keV VMIs optimize abdominal-vessel contrast and perceived image quality
3. Mundt et al. ²⁸ 2023	<i>BMC Med Imaging.</i>	Retrospective radiomics	55 pts	First-gen PCD-CT, non-contrast; descending-aorta adipose segmentation	Two GLCM features (ClusterProminence, ClusterTendency) distinguished Agatston score of ≥ 100 vs. 0; ClusterProminence most stable (10-fold CV)	Periaortic adipose texture on PCD-CT correlates with coronary calcium, hinting at inflammatory changes
4. Hennes et al. ⁴⁷ 2023	<i>Diagnostics.</i>	Intra-individual PCD-CT vs. EID-CT	57 pts	High-pitch PCD-CT of 120 kVp, 55 keV mono-E; EID-CT of 90/100 kVp	SSDE of 4.88 ± 0.48 vs. 6.28 ± 0.50 mGy ($P < 0.001$); CNR of 41.1 ± 8.7 vs. 27.1 ± 6.7 ($P < 0.001$); superior overall IQ and luminal contrast on PCD-CT; reduced blooming on EID	PCD-CT aortic CTA halves dose and boosts CNR versus EID-CT, though blooming still favors EID
5. Rippel et al. ⁴⁹ 2023	<i>Eur J Radiol.</i>	Retrospective matched-cohort comparison	40 pts with PCD-CT vs. 40 pts with EID	Run-off CTA on first-gen PCD-CT; VMI of 40–120 keV	SNR and CNR at 40–60 keV exceeded EID; subjective image quality higher at low keV and not different from EID	Low-keV VMI PCD-CT run-off CTA delivers higher vessel SNR/CNR than low-kVp EID without loss of perceived quality
6. Dane et al. ⁵¹ 2024	<i>J Comput Assist Tomogr.</i>	Retrospective intra-individual comparison	50 pts	Portal-venous PCD-CT with 20 mL contrast reduction; 70 keV mono-E recon	No sig. difference vs. EID in hepatic HU, portal-vein HU, noise, SNR or CNR (all $P > 0.0016$); image quality and metastasis confidence similar; κ : 0.86 (PCD-CT) vs. 0.78 (EID) for metastasis detection	Portal-venous PCD-CT maintains image quality and lesion detection with 20 mL less contrast than weight-based EID-CT
7. Dane et al. ⁵² 2024	<i>J Comput Assist Tomogr.</i>	Retrospective within-patient PCD-CT vs. EI-DECT VNC	74 pts	Portal-venous PCD-CT VNC compared with prior EI-DECT VNC	Qualitative IQ, noise, small-structure delineation all better on PCD-CT ($P < 1 \times 10^{-4}$); lower noise ($P = 0.006$), higher CNR ($P < 0.0001$); dose of 9.2 vs. 9.4 mGy ($P = 0.06$)	PCD-CT-derived VNC images outperform EI-DECT VNC for abdominal imaging without increasing radiation dose

Table 5. Continued

Author	Journal	Study design	Cohort	Imaging	Results	Conclusion
8. Graafen et al. ⁵⁰ 2024	<i>Eur Radiol Exp.</i>	Phantom + prospective patient evaluation	20 pts + in-vitro tubes	Lower-leg CTA reconstructed at 0.4 mm; kernels Qr36–76; QIR of 2–4	Phantom: noise increased with kernel sharpness (16→77 HU); sharpness plateau at Qr60; higher QIR cut noise (51→25 HU) without blurring. <i>In vivo</i> : Qr60 + highest QIR gave best overall quality	Sharp kernel Qr60 combined with highest QIR optimizes lower-leg PCD-CT angiography quality
9. Rippel et al. ⁴⁸ 2024	<i>J Endovasc Ther.</i>	Prospective matched-cohort comparison	50 pts	ECG-gated high-pitch thoracoabdominal PCD-CT; VMI of 40–120 keV	Dose of 4.2 ± 1.4 vs. 7.2 ± 2.2 mGy ($P < 0.001$); SNR increased at 40 and 70 keV; CNR increased at 40–45 keV; low-keV salvages low-contrast scans	High-pitch PCD-CTA halves radiation and boosts SNR/CNR; low-keV VMI rescues sub-optimal contrast studies
10. Yalon et al. ⁵³ 2024	<i>AJR.</i>	Prospective intra-individual comparison	32 pts	Same-day infrapopliteal PCD-CTA (60 mL contrast) vs. EID-CTA (140 mL)	Fibular perforators: 6.4 ± 3.2 vs. 4.2 ± 2.4 ($P < 0.001$); sharpness of 3.2 ± 0.5 vs. 1.8 ± 0.5 ($P < 0.001$); contrast dose decreased by 60%	PCD-CTA improves infrapopliteal vessel visualization and sharpness while halving contrast dose
11. Ayx et al. ⁵⁴ 2025	<i>Eur J Radiol.</i>	Retrospective PCD-CT UHR vs. EID abdominal CTA	25 + 25 pts	PCD-CT UHR (abdominal arterial phase)	CTDIvol of 4.7 vs. 7.3 mGy; eff dose of 3.4 vs. 6.5 mSv; higher SNR/CNR (renal $P = 0.043$); subjective IQ increased ($P < 0.0001$)	UHR PCD-CT abdominal CTA halves dose and boosts image quality versus energy-integrating CT
12. Ghibes et al. ⁵⁵ 2025	<i>Eur J Radiol.</i>	Retrospective PCD-CTA vs. DSA gold standard	109 pts and 933 segments	Lower-extremity PCD-CT CTA ± pure-lumen recon	PCD-CT: sens of 91%, spec of 95%, acc of 93%; κ : 0.79–0.83	PCD-CT CTA shows high accuracy for peripheral-artery stenosis, matching DSA
13. Ota et al. ⁵⁶ 2025	<i>Radiol Med.</i>	Retrospective biomarker study with VNCA maps	200 pts	Abdominal PCD-CT angiography + VNCA subtraction	PCV AUC of 0.94 vs. ACV 0.90; PCV cut-off at 14.8%–73% sens/99% spec for high CVD risk	PCD-CT-derived%-calcification of abdominal aorta is a strong imaging biomarker for systemic CVD risk
14. Sala et al. ⁵⁷ 2025	<i>Bioengineering.</i>	Retrospective PCD-CT vs. histology	14 pts	Pre-op PCD-CT in ATAA; wall thickness measured radiologically and pathologically	pc-CT min/max of 1.05/1.69 mm vs. histology of 1.66/2.82 mm; Bland–Altman shows no systematic bias (min of -0.61 mm; max of -1.1 mm)	Preliminary evidence that PCD-CT wall-thickness measurements agree with histology, supporting risk-stratification beyond diameter

PCD-CT, photon-counting detector computed tomography; VMI, virtual monoenergetic imaging; EID-CT, energy-integrating detector computed tomography; CNR, contrast-to-noise ratio; pts, patients; SNR, signal-to-noise ratio; ns, not significant; GLCM, gray level co-occurrence matrix; CV, coefficient of variation; kVp, kilovolt peak; mono-E, monoenergetic; SSDE, size-specific dose estimate; IQ, image quality; CTA, computed tomography angiography; HU, Hounsfield unit; VNC, virtual non-contrast; EID-CT, energy-integrating dual-energy computed tomography; QIR, quantum iterative reconstruction; Qr, quantum reconstruction; mGy, milligray; TNC, true non-contrast; ECG, electrocardiogram; PCD-CTA, photon-counting detector computed tomography angiography; AJR, American Journal of Roentgenology; UHR, ultra-high resolution; EID-CT, energy-integrating computed tomography; eff dose, effective dose; mSv, millisievert; acc, accuracy; spec, specificity; sens, sensitivity; DSA, digital subtraction angiography; PCD-CTA, photon-counting computed tomography angiography; VNCA, virtual non-contrast calcium maps; PCV, percent calcified volume; ACV, absolute calcified volume; CVD, cardiovascular disease; ATAA, ascending thoracic aortic aneurysm.

Rippel et al.⁴⁸ (prospective matched-cohort, ECG-gated high-pitch thoracoabdominal CTA; $n = 50$) reported an exam-level CT-DIvol of 4.0 [interquartile range (IQR): 3.1–4.9] vs. 6.5 mGy (5.5–9.7) and a DLP of 288 (207–402) versus 466 mGy·cm (365–681) (both $P < 0.001$). On PCD-CT, the SNR was higher at 40 and 70 keV VMIs, and the CNR was higher at 40–45 keV (each $P < 0.001$) than with EID-CT, and low-keV VMIs salvaged low-contrast studies (diagnostic acceptability 50% → 75% at 40 keV). Acquisition and reconstruction: PCD-CT at 120 kVp, 144 × 0.4-mm collimation, 3.2 pitch, 0.25-s rotation time, Bv36 + QIR-3; EID-CT with ATVS at 100/120/140 kVp,

123 × 0.6-mm collimation, 3.2 pitch, 0.28-s rotation time, I26s + ADMIRE-3. The effective dose was not reported.⁴⁸

In a retrospective matched run-off CTA cohort (40 PCD-CT vs. 40 EID-CT), the exam-level CTDIvol and DLP were 3.9 (IQR: 3.0–7.6) vs. 3.5 mGy (2.4–5.7) ($P = 0.024$) and 499 (353–1060) versus 456 mGy·cm (268–753) ($P = 0.029$), respectively. The SNR on PCD-CT exceeded EID-CT for 40–70 keV VMIs, whereas the CNR exceeded EID-CT at 40–45 keV (vs. 80 kVp EID) and 40–50 keV (vs. 100 kVp EID). Subjective image quality was optimal at 40–60 keV and not significantly different from EID-CT overall. Acquisition and reconstruc-

tion: PCD-CT at 120 kVp (QuantumPlus; 144 × 0.4-mm collimation, 0.8 pitch, Qr36 + QIR-3, 1-mm slices, 512 × 512 matrix) with VMI 40–120 keV; EID-CT with ATVS at 80/100 kVp (128 × 0.6 mm, 0.5 pitch, I26s + ADMIRE-3), with identical slice thickness and matrix.⁴⁹

An *in vitro/in vivo* study ($n = 20$) of lower-leg PCD-CT reconstructed at 0.4-mm isotropic resolution found that a sharp quantitative kernel (Qr60) combined with the highest QIR level (QIR-4) best reduced noise without degrading edge definition, yielding the highest qualitative scores. *In vivo* CTDIvol at the lower-leg level was 2.51 mGy (IQR: 2.50–2.57); the DLP, SSDE, and effective dose

were not reported. Acquisition/reconstruction: 120 kVp, CARE Dose4D (image-quality index 145), VMI 55 keV, 512 × 512 matrix, field of view of 205 × 205 mm, kernels Qr44/Qr60/Qr72 with QIR-2/-3/-4; inter-reader reliability was substantial overall (Krippendorff's α : 0.70–0.71) and excellent for noise (α : 0.84–0.86).⁵⁰

In a portal-venous intra-individual comparison ($n = 50$), PCD-CT used 20 mL less intravenous contrast than weight-based EID-CT (90.9 ± 23.0 vs. 111.0 ± 24.0 mL; $P < 0.001$) and, at 70-keV VMIs, showed no significant differences in hepatic or portal-vein attenuation, noise, SNR, or CNR (all $P > 0.0016$), with similar qualitative scores and metastasis-detection confidence [odds ratios: 0.58 (95% CI: 0.33–1.01), 1.25 (0.61–2.56), and 1.17 (0.54–2.52), respectively]. Exam-level dose metrics were a CTDIvol of 9.4 ± 4.0 vs. 11.1 ± 7.4 mGy ($P = 0.005$) and a DLP of 458.7 ± 219.9 vs. 534.6 ± 391.7 mGy·cm ($P = 0.01$); the effective dose was not reported. Inter-reader agreement for metastasis identification was $\kappa = 0.86$ (95% CI: 0.70–1.00) for PCD-CT and 0.78 (0.59–0.98) for EID-CT. Acquisition/reconstruction (fairness): PCD-CT at 120 kVp, 144 × 0.4 mm, CARE Dose4D/Care kV (IQ 145), 70-keV VMI, Br44; EID-CT at 120 kVp, Br44; 4-mm axial and 3-mm coronal/sagittal reconstructions.⁵¹

In a within-patient comparison ($n = 74$) of portal-venous VNC images, PCD-CT vs. EID-CT showed an exam-level CTDIvol of 9.2 ± 3.5 vs. 9.4 ± 9.0 mGy ($P = 0.06$) and a DLP of 417.9 ± 162.8 vs. 523.4 ± 290.9 mGy·cm ($P = 0.026$) (32-cm phantom for both). Qualitatively, PCD-CT VNC had higher overall image quality, lower perceived noise, better small-structure delineation, improved noise texture, and fewer artifacts (all $P < 0.00001$). Quantitatively, PCD-CT VNC had lower attenuation (all $P < 0.05$), lower noise ($P = 0.006$), and a higher CNR ($P < 0.0001$ –0.04); the SNR was lower for enhancing structures (reflecting greater iodine removal) but higher in fat. Acquisition/reconstruction (fairness): PCD-CT 120 kV (QuantumPlus), 144 × 0.4 mm, 0.8 pitch, 0.5-s rotation, CARE Dose4D/CARE kV; EI-DECT of 80–90/Sn150 kV, 0.6 pitch, 0.5-s rotation; VNC recon at 4 mm (Br44) for both.⁵²

For infrapopliteal evaluation, a same-day intra-individual study ($n = 32$) used 60.0 ± 11.0 mL contrast on PCD-CT vs. 139.6 ± 11.8 mL on EID-CT and reported an exam-level CTDIvol of 6.6 ± 2.2 vs. 4.6 ± 3.0 mGy (DLP and effective dose not reported). Acquisition/reconstruction parameters were as follows: PCD-CT UHR mode 120 kV, 120 × 0.2-mm collimation, 0.5 pitch, 0.25-s rotation,

1,024 matrix, Br68, IR-3; EID-CT SE with CARE kV (variable kV), 192 × 0.6-mm collimation, 0.4 pitch, 0.5-s rotation, 512 matrix, Bv44, IR-2. PCD-CT yielded more visualized fibular perforators (R1: 6.4 ± 3.2 vs. 4.2 ± 2.4 , $P < 0.001$; R2: 8.8 ± 3.4 vs. 7.6 ± 3.3 , $P = 0.04$) and greater arterial sharpness (both readers 3.2 vs. 1.7 – 1.8 , $P < 0.001$), with fewer total occlusions for one reader (0.5 ± 1.3 vs. 0.9 ± 1.7 , $P = 0.04$) and similar subjective noise.⁵³

In abdominal arterial-phase imaging, a retrospective comparison (25 + 25 patients) of UHR PCD-CT vs. EID-CT reported a median CTDIvol of 4.7 (IQR: 3.9–5.1) vs. 7.3 mGy (4.6–12.6) and a DLP of 229 (187–262) vs. 295 mGy·cm (233–595); the effective dose was 3.4 (2.8–3.9) vs. 4.4 mSv (3.5–8.9), calculated as $DLP \times 0.015 \text{ mSv} \cdot \text{mGy}^{-1} \cdot \text{cm}^{-1}$. PCD-CT showed higher SNR/CNR (significant for renal arteries, $P = 0.0432$) and higher subjective image quality ($P < 0.0001$). Acquisition/reconstruction details: PCD-CT UHR 120 kV, 0.25-s rotation, 0.8 pitch, 0.2-mm collimation; 0.6 /0.6 mm axial reconstruction, Bv40 kernel; EID-CT 80–140 kV with automatic dose modulation, 0.28-s rotation, 0.6 pitch, 0.6-mm collimation; I30f/Bv38 reconstruction; non-ECG gated.⁵⁴

Diagnostic performance against digital subtraction angiography (DSA) was assessed per segment in 109 patients (933 arterial segments): sensitivity was 91% (95% CI 87–94), specificity was 95% (92–96), and accuracy was 93% ($\approx 95\%$ CI: 91–95) overall; territory-level accuracies were 98% (iliac), 92% (femoro-popliteal), and 93% (calf). Inter-reader agreement was good (weighted κ : 0.791; κ : 0.829 for pure-lumen reconstruction). Agreement with DSA grading was κ : 0.905 (CTA) and κ : 0.825 (pure lumen); 95% CIs for κ were not reported.⁵⁵

Exploratory biomarker research included a radiomics study ($n = 55$) in which two gray-level co-occurrence matrix features from periaortic adipose tissue distinguished Agatston ≥ 100 vs. 0; ClusterProminence showed the most stable performance under 10-fold cross-validation.²⁹ In 200 patients, Ota et al.⁵⁶ derived %-calcification on VNCA maps of the abdominal aorta with an AUC of 0.94 (vs. an ACV AUC of 0.90); a 14.8% cut-off yielded 73% sensitivity and 99% specificity for high cardiovascular-risk classification.

A histology-matched study of ascending thoracic aortic aneurysms ($n = 14$) compared per-patient minimum and maximum aortic-wall thickness on PCD-CT with *ex vivo* histology. ECG-gated UHR-CTA of the aortic root (120 × 0.2-mm collimation; 66-ms temporal resolution) was followed by non-gated tho-

rax–abdomen–pelvis CTA (144 × 0.4-mm collimation). The effective-dose model, CTDIvol, DLP, and SSDE were not reported. The results (unit of analysis = patient-level paired measures) showed a PCD-CT mean minimum/maximum wall thickness of 1.05/1.69 mm versus histology of 1.66/2.82 mm. Bland–Altman (PCD-CT–histology) analysis revealed a mean bias of -0.61 (minimum) and -1.10 mm (maximum); the authors stated no systematic bias, and numerical limits of agreement were not tabulated. Inter-/intra-observer ICCs were not performed or reported.⁵⁷

Figure 5 presents aortic PCD-CT and peripheral angiography in a patient with an iliac stent and multiple stenotic segments in the peripheral arterial system.

Discussion

Evidence summary

Fifty-nine clinical studies published between January 2021 and September 1, 2025, evaluated PCD-CT across the cardiovascular spectrum: 31 coronary, 14 aortic–visceral–peripheral, 4 pulmonary/cardiopulmonary functional, 4 myocardial tissue characterization, and 6 structural-heart/valvular-planning investigations. Across domains, the included studies frequently reported at least one advantage of PCD-CT over EID-CT, including higher spatial resolution at a routine or reduced dose, opportunities for radiation and contrast savings, and robust diagnostic performance in real-world cohorts. Reported examples include vessel-level AUCs of up to 0.99 with UHR modes; reductions in CTDIvol of approximately 40%–60% in matched coronary comparisons; effective doses as low as 0.41 mSv for selected coronary protocols; and 40%–60% contrast-volume reductions in thoracoabdominal CTA, infrapopliteal run-off, and pulmonary embolism protocols without loss of diagnostic confidence.

In routine coronary practice, a large comparative cohort of 7,833 examinations reported an increase in per-patient specificity from 93% to 98% and a reduction in invasive angiography referrals from 13.1% to 9.9%. In the peripheral circulation, diagnostic performance approached that of DSA, with sensitivity around 91% and specificity around 95%. Quantitative capabilities extend beyond morphology: VNC calcium scores have shown ICCs of 0.97–0.99 compared with TNC; iodine-derived ECV estimates differ from cardiac MRI by less than 2% in selected studies; and decision-analytic modeling suggests potential cost savings by reducing downstream testing.

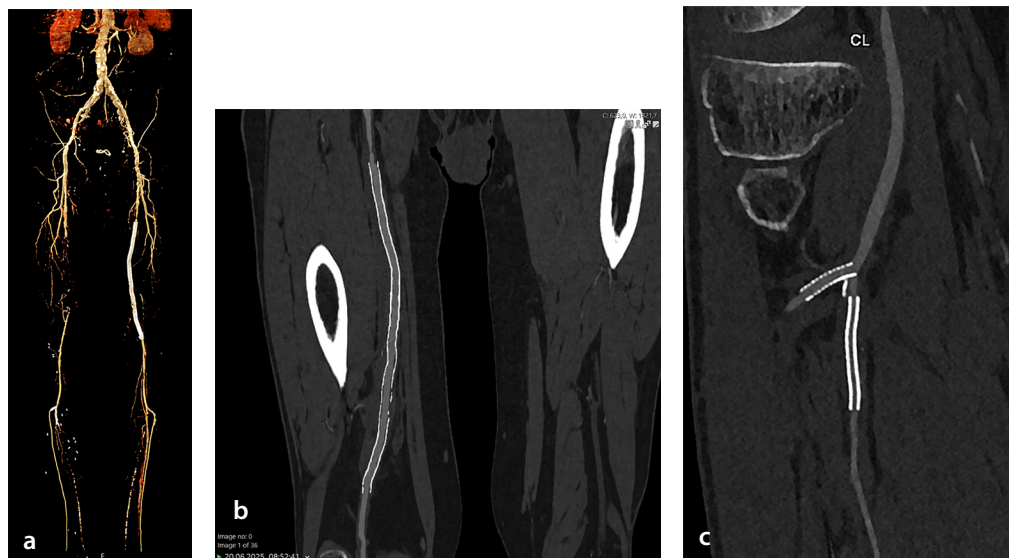


Figure 5. Photon-counting detector computed tomography (CT) angiography in a 65-year-old male patient with a left femoral–popliteal artery stent undergoing vascular evaluation. The scan was performed using a retrospective ECG-triggered protocol with 75 mL of iodinated contrast medium, a dose-length product of 945 mGy-cm, and a standard 0.4-mm slice thickness. The coronal cinematic-rendered maximum intensity projection image (a) and the curved planar reformation (b) depict the metallic tubular stent extending through the left femoral and popliteal arteries. Panel (c) shows the metallic tubular stents in the peroneal artery and anterior tibial arteries. Photon-counting detector CT provides excellent visualization of stent lumen patency and adjacent vascular segments, with high spatial resolution and reduced blooming artifacts, enabling the detailed assessment of in-stent and peri-stent regions.

Context within prior syntheses

Since 2023, several narrative or semi-structured reviews have highlighted the clinical promise of PCD-CT but have generally not aggregated core metrics and often predate workflow and economic data emerging in late 2024–2025. Flohr et al.⁵⁸ presented a seminal cardiac-focused overview in 2023, illustrating early findings such as an approximate 50% calcium-scoring dose reduction and an approximate 11% decrease in blooming-related stenosis overestimation while calling for multicenter outcome and economic evaluations.

Sharma et al.⁵⁹ offered a clinician-oriented digest combining phantom, animal, and human data, reporting ranges of 29%–41% in noise reduction, 20%–36% in CNR improvements, and 100%/87% in sensitivity/specificity for in-stent restenosis but without formal synthesis and with limited attention to non-coronary applications.

Hagen et al.⁶⁰ broadened the scope to oncology, cardiovascular, and pediatric imaging with a qualitative three-pillar framework. Hagar et al.⁶¹ introduced a more structured approach but included only around 20 cardiac studies up to August 2024. In the vascular domain, Wildberger and Alkadhi⁶² reviewed feasibility-level studies and emphasized prototype-related bias and the need for prospective surveillance in endovascular aortic repair follow-up. Van der Bie et al.⁶³ provided

a focused systematic review on stent imaging.

A separate review by Van der Bie et al.⁶³ addressed clinical aspects of PCD-CT utilization not only in cardiovascular imaging but also in abdominal, thoracic, musculoskeletal, neuro, and pediatric imaging. They specifically investigated stent imaging, coronary stenosis measurements, coronary calcium quantification, plaque component quantification, ECV quantification, TAVI planning, and calcium scoring in the cardiovascular imaging section, following a largely narrative methodology with limited comparative analysis.⁶⁴

Clinical impact and emerging signals

In coronary imaging, UHR and low-keV reconstructions sharpen lumen–plaque interfaces, mitigate blooming in calcified and stented segments, and enable contrast-sparing protocols while maintaining diagnostic performance. Real-world cohorts suggest improved specificity, fewer unnecessary invasive angiographies, and operational efficiencies when functional adjuncts can be derived from the same dataset. In myocardial tissue characterization, delayed iodine maps and ECV estimates demonstrate high concordance with MRI at segment and patient levels, supporting single-session coronary-plus-tissue assessment when MRI is contraindicated or impractical.

For structural-heart/TAVR planning, PCD-CT supports accurate annular measurements and peripheral-access assessment at reduced exposure; VMI reconstructions can facilitate calcium quantification without additional native scans. Within the pulmonary circulation, multi-energy datasets enable iodine-based perfusion mapping alongside embolus detection, often with lower radiation and reduced contrast loads. In aortic and peripheral vascular applications, low-keV VMIs improve vascular conspicuity—particularly in small-caliber infrapopliteal vessels—while supporting dose- and contrast-efficient protocols; early research also indicates potential quantitative biomarkers.

Challenges and evidence gaps

The evidence base remains dominated by single-center experiences with heterogeneous acquisition/reconstruction (tube potentials, matrix/slice thickness, kernels/iterative strengths, and VMI energies) and variable reference standards (ICA, CMR, and DSA). Quantitative thresholds—for example, plaque-component cut-points and calcium-score categories in VNC/VNI workflows—require harmonization. Health economic and workflow data are encouraging but largely model based; prospective utilization and cost-impact studies across health systems are needed. Finally, multicenter trials linking PCD-CT-guided decisions to hard clinical endpoints remain limited.

Future directions

Priorities include the following: (i) multicenter prospective studies with standardized acquisition and reporting checklists; (ii) consensus on recommended VMI energies and quantitative thresholds by indication; (iii) reproducibility studies for ECV/perfusion metrics and for VNC calcium scoring across vendors and sites; (iv) prospective evaluations of AI-enabled adjuncts (e.g., CT-FFR, radiomics) anchored to outcomes; and (v) robust cost-utility and budget-impact analyses in diverse clinical contexts.

Limitations of the evidence and of this review

Most included studies are observational, single-center studies, with heterogeneity that precludes formal pooling. Accordingly, we report study-level results and observed ranges rather than pooled effects. External validity across institutions and vendors and longer-term outcomes require further study.

In conclusion, across cardiovascular applications, PCD-CT has been reported to demonstrate higher spatial resolution, improved tissue/contrast characterization, and greater opportunities for radiation and contrast reduction. These technical gains, together with early signals of diagnostic and workflow efficiency, support an expanding clinical role for PCD-CT, contingent upon confirmation in multicenter outcome and economic evaluations.

Footnotes

Acknowledgements

We thank our radiation technicians Ahmet Sarimehmet and Celal Uzungüney for their help.

Conflict of interest disclosure

Deniz Alis is the CEO and co-founder of Hevi AI Health Tech. However, none of Hevi AI's solutions are mentioned in this paper. Other authors have nothing to disclose.

References

1. Si-Mohamed SA, Boccalini S, Lacombe H, et al. Coronary CT angiography with photon-counting CT: first-in-human results. *Radiology*. 2022;303(2):303-313. [\[Crossref\]](#)
2. Mergen V, Eberhard M, Manka R, Euler A, Alkadhi H. First in-human quantitative plaque characterization with ultra-high resolution coronary photon-counting CT angiography. *Front Cardiovasc Med*. 2022;9:981012. [\[Crossref\]](#)
3. Pourmorteza A, Symons R, Henning A, Ulzheimer S, Bluemke DA. Dose efficiency of quarter-millimeter photon-counting computed tomography: first-in-human results. *Invest Radiol*. 2018;53(6):365-372. [\[Crossref\]](#)
4. Symons R, De Bruecker Y, Roosen J, et al. Quarter-millimeter spectral coronary stent imaging with photon-counting CT: initial experience. *J Cardiovasc Comput Tomogr*. 2018;12(6):509-515. [\[Crossref\]](#)
5. Symons R, Cork TE, Lakshmanan MN, et al. Dual-contrast agent photon-counting computed tomography of the heart: initial experience. *Int J Cardiovasc Imaging*. 2017;33(8):1253-1261. [\[Crossref\]](#)
6. Si-Mohamed SA, Sigovan M, Hsu JC, et al. In vivo molecular K-edge imaging of atherosclerotic plaque using photon-counting CT. *Radiology*. 2021;300(1):98-107. [\[Crossref\]](#)
7. Euler A, Higashigaito K, Mergen V, et al. High-pitch photon-counting detector computed tomography angiography of the aorta: intraindividual comparison to energy-integrating detector computed tomography at equal radiation dose. *Invest Radiol*. 2022;57(2):115-121. [\[Crossref\]](#)
8. Higashigaito K, Mergen V, Eberhard M, et al. CT Angiography of the aorta using photon-counting detector CT with reduced contrast media volume. *Radiol Cardiothorac Imaging*. 2023;5(1):e220140. [\[Crossref\]](#)
9. Yalynska T, Polacin M, Frauenfelder T, Martini K. Impact of photon counting detector CT derived virtual monoenergetic images on the diagnosis of pulmonary embolism. *Diagnostics (Basel)*. 2022;12(11):2715. [\[Crossref\]](#)
10. Dangelmaier J, Bar-Ness D, Daerr H, et al. Experimental feasibility of spectral photon-counting computed tomography with two contrast agents for the detection of endoleaks following endovascular aortic repair. *Eur Radiol*. 2018;28(8):3318-3325. [\[Crossref\]](#)
11. Kotronias RA, de Maria GL, Xie C, et al. Benchmarking Photon-counting computed tomography angiography against invasive assessment of coronary stenosis: implications for severely calcified coronaries. *JACC Cardiovasc Imaging*. 2025;18(5):572-585. [\[Crossref\]](#)
12. Fahrni G, Boccalini S, Lacombe H, et al. Ultra-high-resolution 40 keV virtual monoenergetic imaging using spectral photon-counting CT in high-risk patients for coronary stenoses. *Eur Radiol*. 2025;35(6):3042-3053. [\[Crossref\]](#)
13. Sakai K, Shin D, Singh M, et al. Diagnostic performance and clinical impact of photon-counting detector computed tomography in coronary artery disease. *J Am Coll Cardiol*. 2025;85(4):339-348. [\[Crossref\]](#)
14. Hagar MT, Soschynski M, Saffar R, et al. Accuracy of ultrahigh-resolution photon-counting CT for detecting coronary artery disease in a high-risk population. *Radiology*. 2023;307(5):e223305. [\[Crossref\]](#)
15. Araki S, Nakamura S, Takafuji M, Ichikawa Y, Sakuma H, Kitagawa K. Ultra-low-dose coronary computed tomography angiography using photon-counting detector computed tomography. *Eur Heart J Imaging Methods Pract*. 2024;2(3):qyae125. [\[Crossref\]](#)
16. Zhao YE, Hu Q, Zhu J, et al. Evaluation of low-dose chest scans for coronary artery calcium scoring using photon-counting computed tomography with different slice thicknesses and iterative reconstruction levels. *Quant Imaging Med Surg*. 2025;15(4):3565-3574. [\[Crossref\]](#)
17. Rajiah PS, Dunning CAS, Rajendran K, et al. High-pitch multienergy coronary CT angiography in dual-source photon-counting detector CT scanner at low iodinated contrast dose. *Invest Radiol*. 2023;58(9):681-690. [\[Crossref\]](#)
18. Cundari G, Deilmann P, Mergen V, et al. Saving contrast media in coronary CT angiography with photon-counting detector CT. *Acad Radiol*. 2024;31(1):212-220. [\[Crossref\]](#)
19. Greffier J, Si-Mohamed SA, Lacombe H, et al. Virtual monochromatic images for coronary artery imaging with a spectral photon-counting CT in comparison to dual-layer CT systems: a phantom and a preliminary human study. *Eur Radiol*. 2023;33(8):5476-5488. [\[Crossref\]](#)
20. Vattay B, Boussoussou M, Vecsey-Nagy M, et al. Qualitative and quantitative image quality of coronary CT angiography using photon-counting computed tomography: standard and ultra-high resolution protocols. *Eur J Radiol*. 2024;175:111426. [\[Crossref\]](#)
21. Vattay B, Szilveszter B, Boussoussou M, et al. Impact of virtual monoenergetic levels on coronary plaque volume components using photon-counting computed tomography. *Eur Radiol*. 2023;33(12):8528-8539. [\[Crossref\]](#)
22. Haag NP, Michael AE, Lennartz S, et al. Coronary Artery calcium scoring using virtual versus true non-contrast images from photon-counting coronary CT Angiography. *Radiology*. 2024;310(3):e230545. [\[Crossref\]](#)
23. Vecsey-Nagy M, Varga-Szemes A, Emrich T, et al. Calcium scoring on coronary computed angiography tomography with photon-counting detector technology: predictors of performance. *J Cardiovasc Comput Tomogr*. 2023;17(5):328-335. [\[Crossref\]](#)
24. Sharma SP, van der Bie J, van Straten M, et al. Coronary calcium scoring on virtual non-contrast and virtual non-iodine reconstructions compared to true non-contrast images using photon-counting computed tomography. *Eur Radiol*. 2024;34(6):3699-3707. [\[Crossref\]](#)
25. Kaldas M, Weber J, Parikh R, et al. Coronary calcium scoring using true and virtual non-contrast reconstructions on photon-counting CT with differing slice increment: impact on calcium severity classifications. *J Clin Med*. 2025;14(9):2875. [\[Crossref\]](#)

26. Ayx I, Lichti L, Buettner S, et al. Feasibility of on-site CT-FFR analysis on cardiac photon-counting CT in evaluation of hemodynamically significant stenosis in comparison to invasive catheter angiography. *Eur J Radiol.* 2025;183:111927. [\[Crossref\]](#)
27. Brendel JM, Walterspiel J, Hagen F, et al. Coronary artery disease detection using deep learning and ultrahigh-resolution photon-counting coronary CT angiography. *Diagn Interv Imaging.* 2025;106(2):68-75. [\[Crossref\]](#)
28. Mundt P, Tharmaseelan H, Hertel A, et al. Periaortic adipose radiomics texture features associated with increased coronary calcium score-first results on a photon-counting-CT. *BMC Med Imaging.* 2023;23(1):97. [\[Crossref\]](#)
29. Kahmann J, Tharmaseelan H, Riffel P, et al. Pericoronary radiomics texture features associated with hypercholesterolemia on a photon-counting-CT. *Front Cardiovasc Med.* 2023;10:1223035. [\[Crossref\]](#)
30. Kahmann J, Nörenberg D, Papavassiliu T, Schoenberg SO, Froelich MF, Ayx I. Interrelation of pericoronary adipose tissue texture and coronary artery disease of the left coronary artery in cardiac photon-counting computed tomography. *Front Cardiovasc Med.* 2024;11:1499219. [\[Crossref\]](#)
31. Vecsey-Nagy M, Emrich T, Tremamunno G, et al. Cost-effectiveness of ultrahigh-resolution photon-counting detector coronary CT angiography for the evaluation of stable chest pain. *J Cardiovasc Comput Tomogr.* 2025;19(1):106-112. [\[Crossref\]](#)
32. Tremamunno G, Varga-Szemes A, Kravchenko D, et al. Accuracy of photon-counting detector CT-based iodine maps for myocardial late enhancement detection. *Eur Radiol.* 2025. [\[Crossref\]](#)
33. Klambauer K, Klotz E, Moser LJ, et al. Myocardial late enhancement with photon-counting detector CT in spontaneous coronary artery dissection: prospective comparison with cardiac MRI. *Invest Radiol.* 2026;61(1):32-40. [\[Crossref\]](#)
34. Popp S, Beitzke D, Strassl A, et al. Evaluation of extracellular volume and coronary artery disease in cardiac amyloidosis using photon-counting CT. *Invest Radiol.* 2025;60(10):698-707. [\[Crossref\]](#)
35. Gkizas C, Longere B, Sliwicka O, et al. Photon-counting CT-derived extracellular volume in acute myocarditis: comparison with cardiac MRI. *Diagn Interv Imaging.* 2025;106(7-8):255-263. [\[Crossref\]](#)
36. Dirrachs T, Schröder J, Frick M, et al. Photon-counting versus dual-source CT for transcatheter aortic valve implantation planning. *Acad Radiol.* 2024;31(12):4780-4789. [\[Crossref\]](#)
37. Yang Y, Richter R, Halfmann MC, et al. Prospective ECG-gated High-pitch photon-counting CT angiography: evaluation of measurement accuracy for aortic annulus sizing in TAVR planning. *Eur J Radiol.* 2024;178:111604. [\[Crossref\]](#)
38. Hagar MT, Kluemper T, Hein M, et al. Photon-counting CT-angiography in pre-TAVR aortic annulus assessment: effects of retrospective vs. prospective ECG-synchronization on prosthesis valve selection. *Int J Cardiovasc Imaging.* 2024;40(4):811-820. [\[Crossref\]](#)
39. Feldle P, Scheuber M, Grunz JP, et al. Virtual non-iodine photon-counting CT-angiography for aortic valve calcification scoring. *Sci Rep.* 2024;14(1):4724. [\[Crossref\]](#)
40. Brendel JM, Walterspiel J, Hagen F, et al. Coronary artery disease evaluation during transcatheter aortic valve replacement work-up using photon-counting CT and artificial intelligence. *Diagn Interv Imaging.* 2024;105(7-8):273-280. [\[Crossref\]](#)
41. Sharma SP, Verhemel S, Hirsch A, et al. Diagnostic performance of high and ultrahigh-resolution photon counting CT for detection of coronary artery disease in patients evaluated for transcatheter aortic valve implantation. *Int J Cardiovasc Imaging.* 2024. [\[Crossref\]](#)
42. Scharm SC, Schaefer-Prokop C, Winther HB, et al. Regional pulmonary morphology and function: photon-counting CT assessment. *Radiology.* 2023;308(1):e230318. [\[Crossref\]](#)
43. Kerber B, Hüllner M, Maurer A, et al. Photon-counting detector CT iodine maps versus SPECT/CT: advancing lung perfusion imaging in chronic thromboembolic pulmonary hypertension. *Invest Radiol.* 2025;60(9):569-576. [\[Crossref\]](#)
44. Saeed S, Niehoff JH, Boriesosdick J, et al. Minimizing contrast media dose in CT pulmonary angiography with clinical photon counting using high pitch technique. *Acad Radiol.* 2024;31(2):686-692. [\[Crossref\]](#)
45. Yalon M, Hoodeshenas S, Chan A, et al. Improved pulmonary artery evaluation using high-pitch photon-counting CT compared to high-pitch conventional or routine-pitch conventional dual-energy CT. *J Comput Assist Tomogr.* 2024;48(6):897-905. [\[Crossref\]](#)
46. Dillinger D, Overhoff D, Booz C, et al. Impact of CT photon-counting virtual monoenergetic imaging on visualization of abdominal arterial vessels. *Diagnostics (Basel).* 2023;13(5):938. [\[Crossref\]](#)
47. Hennes JL, Huflage H, Grunz JP, et al. An intra-individual comparison of low-keV photon-counting CT versus energy-integrating-detector CT angiography of the aorta. *Diagnostics (Basel).* 2023;13(24):3645. [\[Crossref\]](#)
48. Rippel K, Luitjens J, Habeeballah O, et al. Evaluation of ECG-gated, high-pitch thoracoabdominal angiographies with dual-source photon-counting detector computed tomography. *J Endovasc Ther.* 2024;15266028241230943. [\[Crossref\]](#)
49. Rippel K, Decker JA, Wudy R, et al. Evaluation of run-off computed tomography angiography on a first-generation photon-counting detector CT scanner - Comparison with low-kVp energy-integrating CT. *Eur J Radiol.* 2023;158:110645. [\[Crossref\]](#)
50. Graafen D, Bart W, Halfmann MC, et al. *In vitro* and *in vivo* optimized reconstruction for low-keV virtual monoenergetic photon-counting detector CT angiography of lower legs. *Eur Radiol Exp.* 2024;8(1):89. [\[Crossref\]](#)
51. Dane B, Mabud T, Melamud K, et al. Reduced intravenous contrast dose portal venous phase photon-counting computed tomography compared with conventional energy-integrating detector portal venous phase computed tomography. *J Comput Assist Tomogr.* 2024;48(5):675-682. [\[Crossref\]](#)
52. Dane B, Ruff A, O'Donnell T, et al. Photon-counting computed tomography versus energy-integrating dual-energy computed tomography: virtual noncontrast image quality comparison. *J Comput Assist Tomogr.* 2024;48(2):251-256. [\[Crossref\]](#)
53. Yalon M, Inoue A, Thorne JE, et al. Infrapopliteal segments on lower extremity CTA: prospective intraindividual comparison of energy-integrating detector CT and photon-counting detector CT. *AJR Am J Roentgenol.* 2024;222(3):e2329778. [\[Crossref\]](#)
54. Ayx I, Schwenke K, Rotkopf L, et al. Feasibility of ultra-high-resolution abdominal CT angiography in PCCT in outperforming conventional EICT. *Eur J Radiol.* 2025;186:112050. [\[Crossref\]](#)
55. Ghibes P, Hagen F, Weissinger M, et al. Diagnostic performance of photon-counting CT angiography in peripheral artery disease compared to DSA as gold standard. *Eur J Radiol.* 2025;182:111834. [\[Crossref\]](#)
56. Ota T, Nakamoto A, Hori M, et al. Quantification of abdominal aortic calcification using photon-counting CT angiography: an imaging biomarker for high-risk cardiovascular patients. *Radiol Med.* 2025;130(6):817-829. [\[Crossref\]](#)
57. Sala A, de Vincentiis C, Grimaldi F, et al. Can the novel photon-counting CT scan accurately predict aortic wall thickness? Preliminary results. *Bioengineering (Basel).* 2025;12(3):306. [\[Crossref\]](#)
58. Flohr T, Schmidt B, Ulzheimer S, Alkadhi H. Cardiac imaging with photon counting CT. *Br J Radiol.* 2023;96(1152):20230407. [\[Crossref\]](#)
59. Sharma SP, Lemmens MDK, Smulders MW, Budde RPJ, Hirsch A, Muhl C. Photon-counting detector computed tomography in cardiac imaging. *Neth Heart J.* 2024;32(11):405-416. [\[Crossref\]](#)
60. Hagen F, Soschynski M, Weis M, et al. Photon-counting computed tomography - clinical application in oncological, cardiovascular, and pediatric radiology. *Rofo.* 2024;196(1):25-35. [\[Crossref\]](#)

61. Hagar MT, Schlett CL, Oechsner T, et al. Photon-counting detector CT: advances and clinical applications in cardiovascular imaging. *Rofo*. 2025;197(5):509-517. [\[Crossref\]](#)
62. Wildberger JE, Alkadhi H. New horizons in vascular imaging with photon-counting detector CT. *Invest Radiol*. 2023;58(7):499-504. [\[Crossref\]](#)
63. van der Bie J, Sharma SP, Hirsch A, et al. Coronary stent imaging with photon-counting detector CT. *Radiol Cardiothorac Imaging*. 2025;7(3):e240338. [\[Crossref\]](#)
64. van der Bie J, van der Laan T, van Straten M, et al. Photon-counting CT: an updated review of clinical results. *Eur J Radiol*. 2025;190:112189. [\[Crossref\]](#)
65. Pinos D, Griffith J 3rd, Emrich T, et al. Intra-individual comparison of image quality of the coronary arteries between photon-counting detector and energy-integrating detector CT systems. *Eur J Radiol*. 2023;166:111008. [\[Crossref\]](#)
66. Rotkopf LT, Froelich MF, Riffel P, et al. Influence of heart rate and heart rate variability on the feasibility of ultra-fast, high-pitch coronary photon-counting computed tomography angiography. *Int J Cardiovasc Imaging*. 2023;39(5):1065-1073. [\[Crossref\]](#)
67. McCollough CH, Winfree TN, Melka EF, Rajendran K, Carter RE, Leng S. Photon-counting detector computed tomography versus energy-integrating detector computed tomography for coronary artery calcium quantitation. *J Comput Assist Tomogr*. 2024;48(2):212-216. [\[Crossref\]](#)
68. Fink N, Emrich T, Schoepf UJ, et al. Improved detection of small and low-density plaques in virtual non-contrast imaging-based calcium scoring at photon-counting detector CT. *Radiol Cardiothorac Imaging*. 2024;6(4):e230328. [\[Crossref\]](#)
69. Mundt P, Hertel A, Tharmaseelan H, et al. Analysis of epicardial adipose tissue texture in relation to coronary artery calcification in PCCT: The EAT Signature! *Diagnostics (Basel)*. 2024;14(3):277. [\[Crossref\]](#)
70. van der Bie J, Bos D, Dijkshoorn ML, Booij R, Budde RPJ, van Straten M. Thin slice photon-counting CT coronary angiography compared to conventional CT: objective image quality and clinical radiation dose assessment. *Med Phys*. 2024;51(4):2924-2932. [\[Crossref\]](#)
71. Wang M, Zhang X, Li J, et al. Quantification accuracy in photon-counting detector CT for coronary artery calcium score: a pilot study. *Int J Cardiovasc Imaging*. 2024;40(10):2181-2191. [\[Crossref\]](#)
72. Fahrni G, Boccalini S, Mahmoudi A, et al. Quantification of coronary artery stenosis in very-high-risk patients using ultra-high resolution spectral photon-counting CT. *Invest Radiol*. 2025;60(2):114-122. [\[Crossref\]](#)
73. Sartoretti T, Mergen V, Dzaferi A, et al. Effect of temporal resolution on calcium scoring: insights from photon-counting detector CT. *Int J Cardiovasc Imaging*. 2025;41(3):615-625. [\[Crossref\]](#)



Radiologic patterns of interstitial lung disease: a pictorial review based on the 2025 European Respiratory Society/American Thoracic Society Classification

 Furkan Ufuk

University of Chicago Medicine, Department of
Radiology, Chicago, United States of America

ABSTRACT

Interstitial lung diseases (ILDs) are a heterogeneous group of diffuse parenchymal disorders with overlapping clinical and physiological features, making imaging central to diagnosis and management. Thin-slice computed tomography (CT) enables recognition of characteristic patterns and disease monitoring, serving as a cornerstone of multidisciplinary discussion. The 2025 European Respiratory Society/American Thoracic Society classification introduces refinements to the 2013 framework, grouping ILDs into three categories: interstitial patterns, alveolar filling disorders, and rare or unclassified disorders. This pictorial essay illustrates the updated classification with representative CT examples.

KEYWORDS

Interstitial lung diseases, computed tomography, bronchiolocentric interstitial pneumonia, alveolar macrophage pneumonia, fibrosis, usual interstitial pneumonia

Interstitial lung diseases (ILDs) are a heterogeneous group of parenchymal disorders characterized by varying degrees of inflammation and fibrosis.¹⁻³ Imaging, particularly thin-section computed tomography (CT), is central to diagnosis, management, and monitoring.³⁻⁶ The 2025 European Respiratory Society/American Thoracic Society (ERS/ATS) consensus update on ILD classification refined the framework (Figure 1), reflecting advances in pathobiology, new terminology, and recognition of clinically relevant radiologic-pathologic patterns.⁷ This article aims to provide radiologists with a visual reference for the updated ILD classification, illustrated with representative CT images.

Interstitial patterns

Interstitial patterns are key for ILD classification and are labeled as fibrotic or non-fibrotic. Fibrotic disease shows distortion, traction bronchiectasis, and honeycombing, indicating ongoing, often irreversible remodeling.⁵⁻⁷ Non-fibrotic patterns reflect inflammation or acute injury without fibrosis. Accurate CT differentiation is vital for diagnosis, prognosis, and treatment.⁷

Usual interstitial pneumonia (UIP) is the most clinically significant form of fibrotic interstitial pneumonia and serves as the defining histopathological correlate of idiopathic pulmonary fibrosis.^{1,4,7} The condition reflects chronic, progressive parenchymal remodeling with temporal and spatially heterogeneous fibrosis. Histology shows subpleural-predominant fibrosis, architectural distortion, fibroblastic foci, and honeycomb cysts.⁸ CT enables non-invasive recognition, often avoiding biopsy. Hallmark findings include basal and subpleural reticulation, traction bronchiectasis, and honeycombing (Figure 2).^{1,4,7} Honeycombing is the most specific sign, with ancillary features such as volume loss and distortion.

Handling editor: Mehmet Ruhi Onur

Corresponding author: Furkan Ufuk

E-mail: furkan.ufuk@hotmail.com

Received 15 November 2025; revision requested 20
December 2025; last revision requested 31 January 2026;
accepted 10 February 2026.



Epub: 26.02.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263751

You may cite this article as: Ufuk F. Radiologic patterns of interstitial lung disease: a pictorial review based on the 2025 European Respiratory Society/American Thoracic Society Classification. *Diagn Interv Radiol.* 2026;32(5):643-654.

UIP carries a poor prognosis with progressive functional decline despite therapy.^{1,5-7}

Non-specific interstitial pneumonia (NSIP) is the most common pattern in connective tissue disease-associated ILD (CTD-ILD). However, it may also occur idiopathically or with drug reactions and other systemic disorders.⁴⁻⁷ Histology shows uniform interstitial inflammation and fibrosis, without the temporal heterogeneity of UIP.¹⁻³ CT usually shows symmetric ground-glass opacities (GGOs) and fine reticulation, with lower lobe and subpleural predominance, and possible subpleural sparing, which helps distinguish NSIP from UIP (Supplementary Figure 1). Traction bronchiectasis, volume loss, and architectural distortion are frequent in the fibrotic subtype, whereas honeycombing is rare (Figure 3).^{4-7,9} NSIP carries a better prognosis than UIP and often responds to immunosuppressive therapy.

Bronchiolocentric interstitial pneumonia (BIP) is a newly recognized interstitial pattern in the 2025 ERS/ATS classification, defined by fibrosis and inflammation centered around the small airways.⁷ It is relevant because it may occur in diverse contexts, including hypersensitivity pneumonitis, CTD-ILD, aspiration, and drug-induced lung injury, often mimicking other fibrotic processes. On CT, BIP typically manifests as airway-centered fibrosis with associated tree-in-bud opacities and centrilobular nodules, sometimes accompanied by peribronchiolar reticulation (Figure 4).^{7,10}

Diffuse alveolar damage (DAD), formerly termed acute interstitial pneumonia, is the histopathologic correlate of acute respiratory distress syndrome (ARDS) in idiopathic or secondary settings.^{7,11} It is characterized by

acute alveolar injury with hyaline membrane formation and interstitial edema, leading to rapid clinical deterioration.¹¹ On CT, DAD typically presents with diffuse GGOs and patchy consolidations, often with dependent predominance, reflecting the exudative phase of ARDS (Figure 5). In survivors, imaging may evolve toward reticulation and traction bronchiectasis, indicating fibrotic remodeling.^{5-7,10,11}

Alveolar filling disorders

Alveolar filling disorders are characterized by the accumulation of specific cells or materials within the alveolar spaces, distinguishing them from fibrotic interstitial pneumonias, which primarily involve the lung interstitium.⁷

Organizing pneumonia (OP) is characterized by intra-alveolar fibroblastic plugs (Masson bodies) that preserve underlying lung architecture. The condition may be idiopathic (cryptogenic OP) or occur secondarily. Secondary causes include infection, CTD, drug reactions, radiation, or aspiration.^{7,11} On CT, OP typically presents with patchy peripheral or peribronchovascular consolidations. These are often migratory in distribution. A characteristic but less common feature is the reversed halo sign, which is a central GGO surrounded by a ring of consolidation. Arcade-like opacities, representing curvilinear perilobular consolidations, may also be seen. These reflect the organization of tissue along the interstitium (Figure 6).^{7,11,12} OP frequently responds to corticosteroid therapy and carries a more favorable prognosis than fibrotic ILDs.¹¹

Respiratory bronchiolitis-interstitial lung disease (RB-ILD) is a smoking-related disorder characterized by the accumulation of pigmented macrophages within respiratory bronchioles and adjacent alveoli.¹³ It occurs almost exclusively in current or former heavy smokers. It represents the diffuse, symptomatic form of the histologic lesion known as RB, which is otherwise common and often subclinical in smokers. On CT, RB-ILD typically demonstrates ill-defined centrilobular nodules and patchy GGOs, usually with an upper lobe predominance (Figure 7). Associated findings may include bronchial wall

thickening and expiratory air trapping. The condition often improves with smoking cessation and generally carries a favorable prognosis.^{7,13}

Alveolar macrophage pneumonia (AMP), formerly termed desquamative interstitial pneumonia, is an alveolar filling disorder marked by macrophage accumulation within alveolar spaces.^{7,13} Although strongly linked to cigarette smoking, AMP can also, albeit rarely, occur in association with CTDs such as rheumatoid arthritis or systemic sclerosis, certain drug exposures (e.g., nitrofurantoin, amiodarone), or environmental factors (such as occupational dust). On CT, AMP shows diffuse, basally predominant GGOs with mild reticulation (Figure 8). Notably, honeycombing is rare, although small cystic changes may appear in advanced disease.¹³

Eosinophilic pneumonias comprise acute and chronic forms, both characterized by eosinophilic infiltration of alveoli and the interstitium, often associated with asthma, drug reactions, infection, or autoimmune disease.^{7,13} Acute eosinophilic pneumonia typically presents with rapidly progressive respiratory failure and diffuse pulmonary involvement. CT findings include bilateral GGOs and patchy consolidations.^{11,13} Chronic eosinophilic pneumonia follows a more indolent course and classically shows peripheral, upper-lobe-predominant consolidations and GGOs, sometimes described as the “photographic negative” of pulmonary edema (Figure 9). Pleural effusion is uncommon, and EP typically responds well to corticosteroid therapy.

Pulmonary alveolar proteinosis (PAP) is characterized by the intra-alveolar accumulation of surfactant material, which may be autoimmune, secondary to hematologic or environmental exposures, or congenital.¹⁴ On CT, PAP exhibits the classic “crazy-paving” pattern, defined by diffuse or patchy GGOs superimposed with smooth interlobular septal thickening and intralobular lines. Distribution is often bilateral and symmetric, with perihilar or lower-lung predominance (Figure 10).^{7,14} Imaging findings help to avoid invasive testing and guide management, including whole-lung lavage or targeted therapy in autoimmune cases.

Main points

- The 2025 European Respiratory Society/American Thoracic Society update organizes interstitial lung diseases (ILDs) into three computed tomography (CT)-based pattern groups.
- Recognizing key CT patterns is essential for diagnosing and managing ILD.
- This pictorial review provides practical CT examples for everyday interpretation.

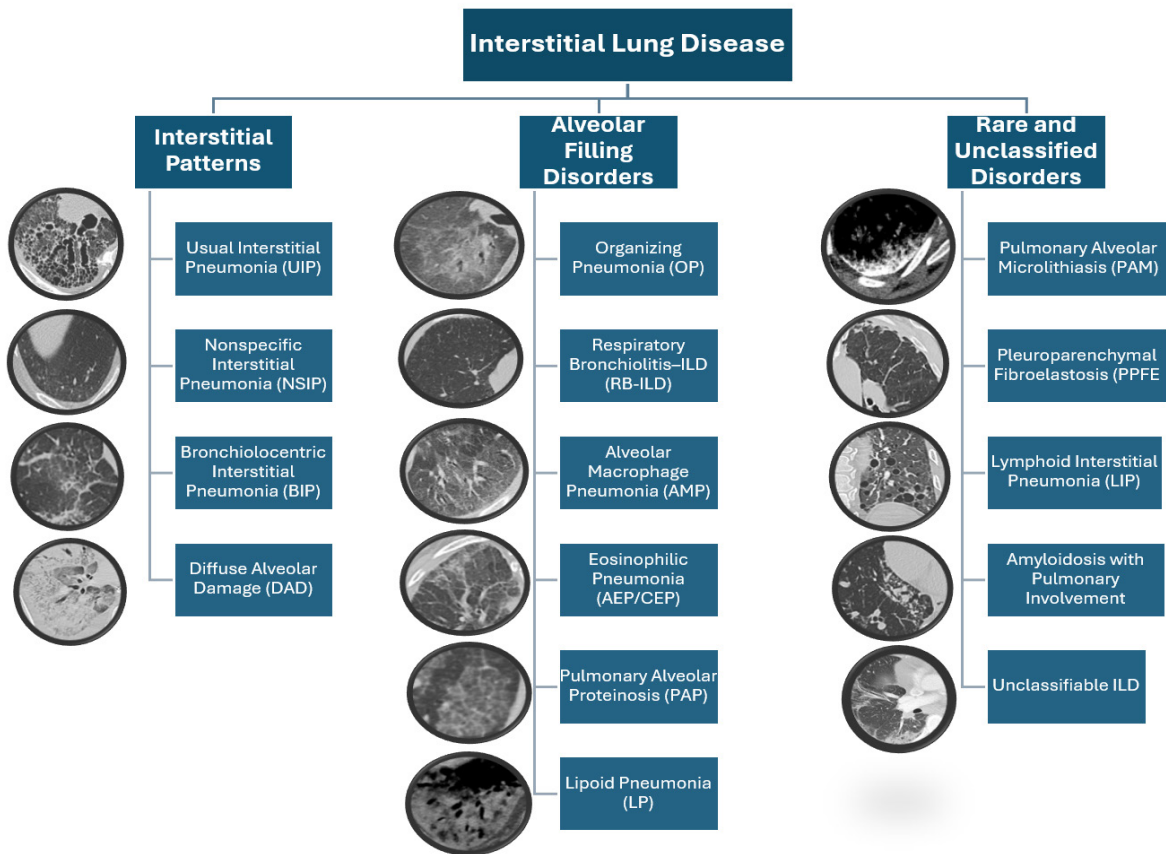


Figure 1. Radiological patterns of interstitial, alveolar filling, and uncommon or unclassified pulmonary disorders. Representative computed tomography images illustrate the predominant parenchymal patterns. ILD, interstitial lung disease.

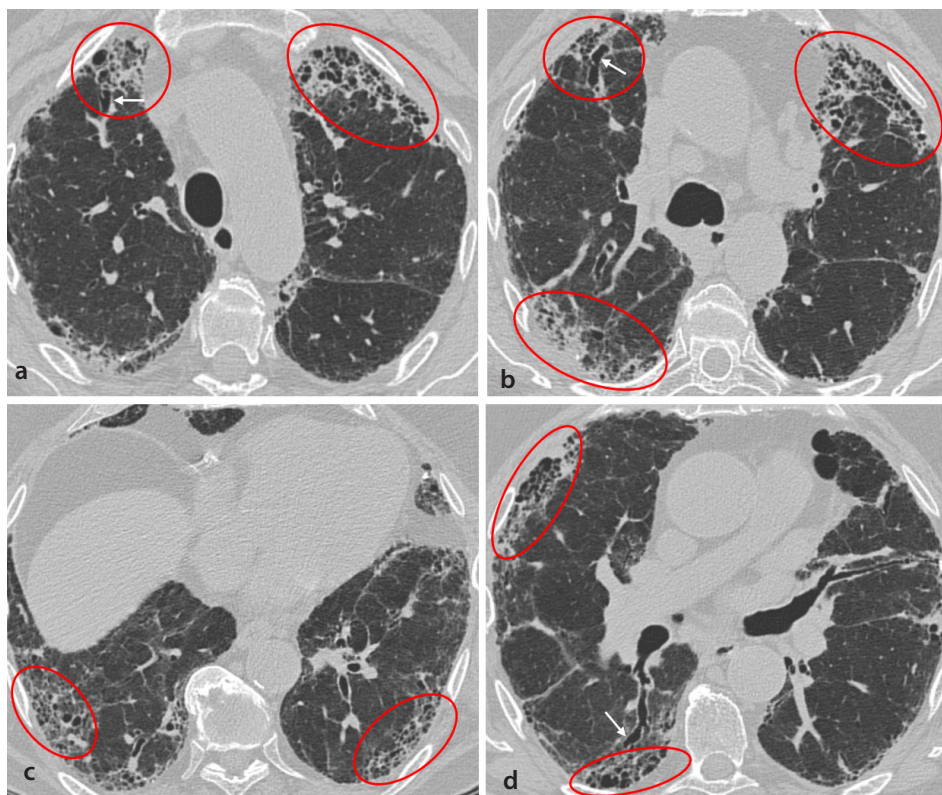


Figure 2. Usual interstitial pneumonia pattern on chest computed tomography (CT). Axial CT images (a–d) demonstrate characteristic UIP features, including basal and subpleural reticulation, honeycombing (red circles), and traction bronchiectasis (arrows). Fibrotic changes are heterogeneously distributed with basal and subpleural predominance and relative sparing of the central regions in a patient with idiopathic pulmonary fibrosis. UIP, usual interstitial pneumonia.

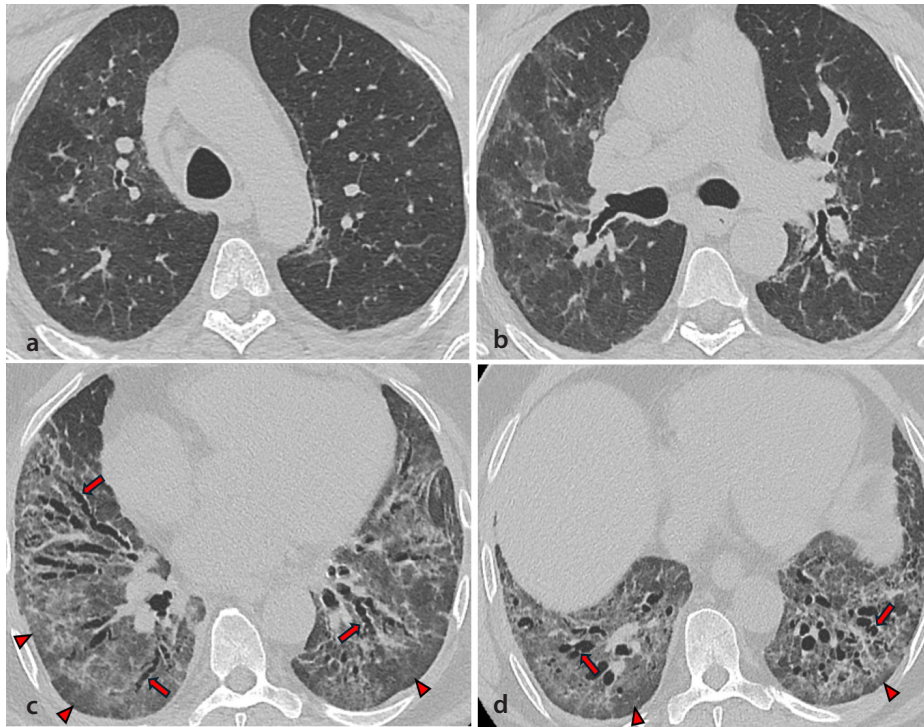


Figure 3. Non-specific interstitial pneumonia pattern on chest computed tomography (CT). Axial CT images (a–d) demonstrate bilateral ground-glass opacities and fine reticulation with lower-lobe predominance, accompanied by traction bronchiectasis (arrows) and areas of subpleural sparing (arrowheads). Findings are shown in a patient with systemic sclerosis–associated interstitial lung disease.

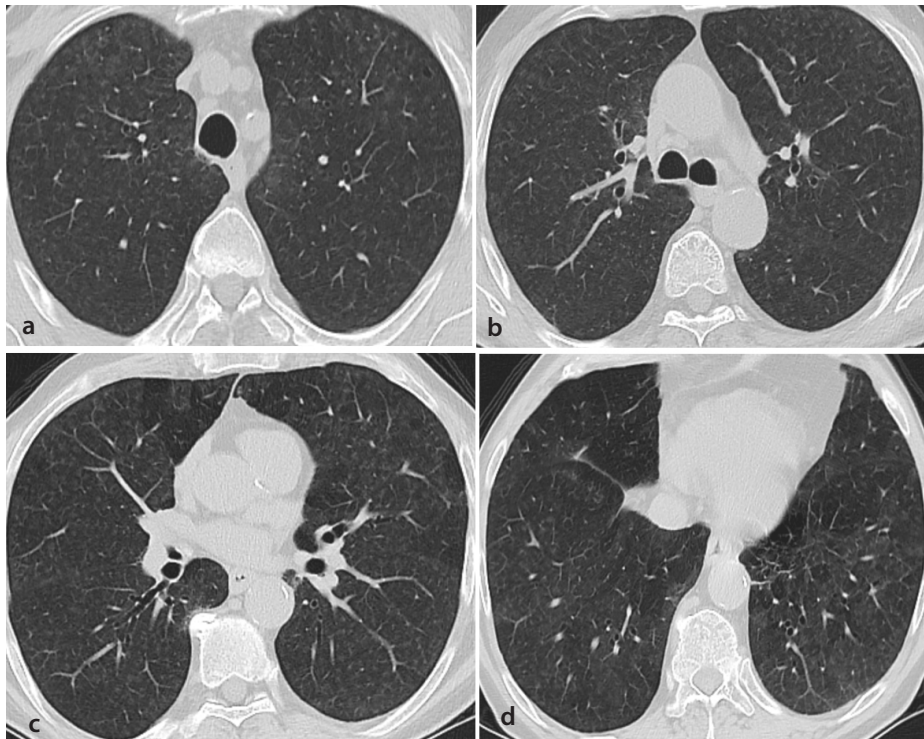


Figure 4. Bronchiolocentric interstitial pneumonia pattern on chest computed tomography (CT). Axial CT images (a–d) demonstrate diffusely distributed centrilobular nodules and ground-glass opacities, consistent with an airway-centered interstitial process. Basilar-predominant mosaic attenuation with associated lucent areas suggests air trapping. Based on integration of the imaging findings, histopathologic correlation, and exposure history, this pattern was interpreted as bronchiolocentric interstitial pneumonia, consistent with non-fibrotic hypersensitivity pneumonitis in the appropriate clinical context.

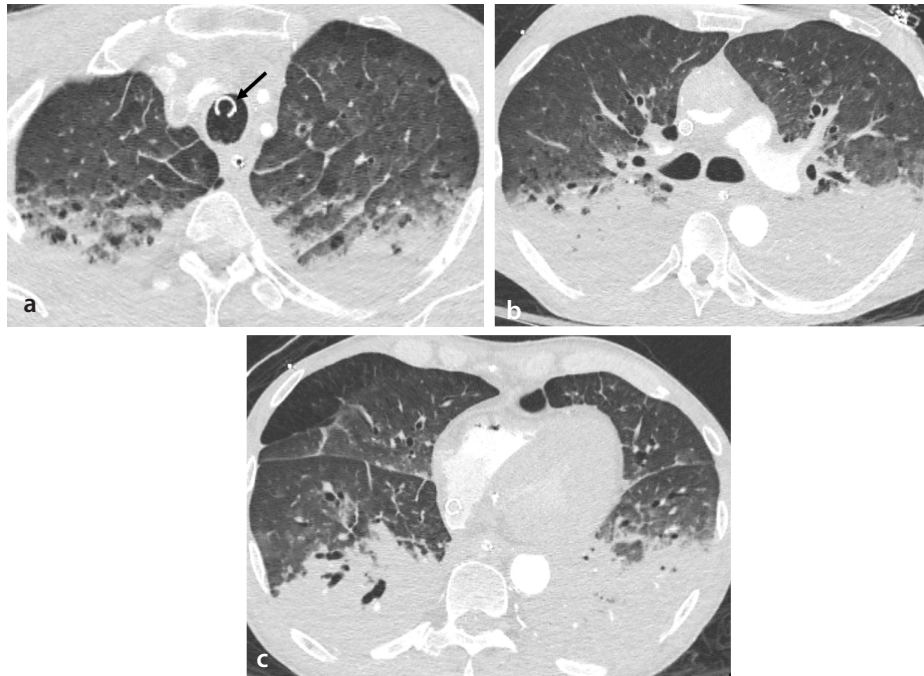


Figure 5. Diffuse alveolar damage pattern on chest computed tomography (CT). Axial CT images (a–c) demonstrate dependent gradient of consolidations, consistent with the exudative phase of diffuse alveolar damage. The patient experienced respiratory deterioration and met clinical criteria for acute respiratory distress syndrome. The endotracheal tube is noted (arrow), along with interstitial thickening, suggestive of underlying pulmonary edema.

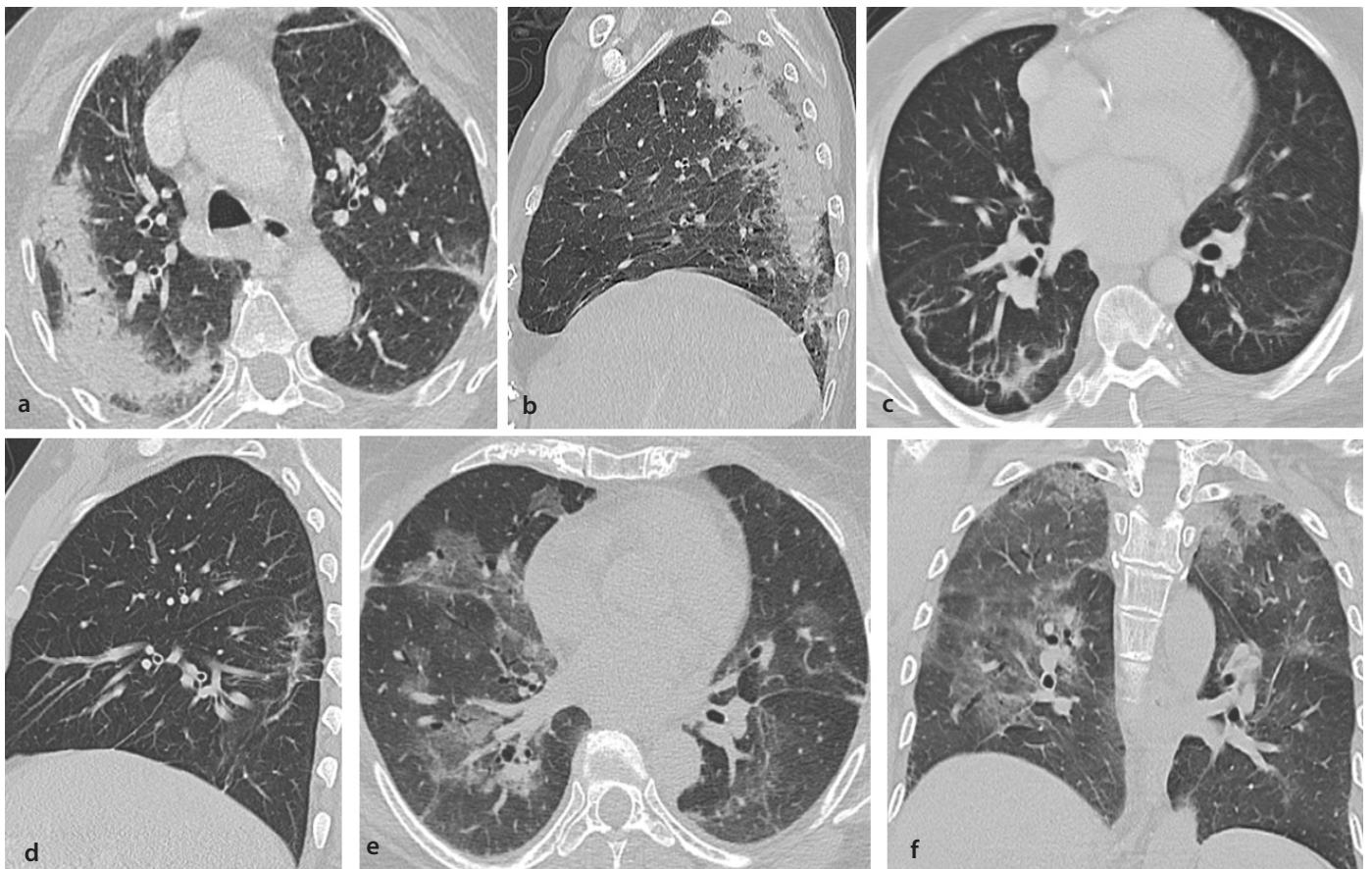


Figure 6. Organizing pneumonia patterns on chest computed tomography (CT). Axial (a) and sagittal (b) CT images demonstrate patchy peripheral consolidations, more prominent on the right side and involving both upper and lower lobes. Axial (c) and sagittal (d) images show peripheral, arcade-like opacities with areas of subpleural sparing. Axial (e) and coronal (f) images reveal patchy peripheral and peribronchial ground-glass opacities.

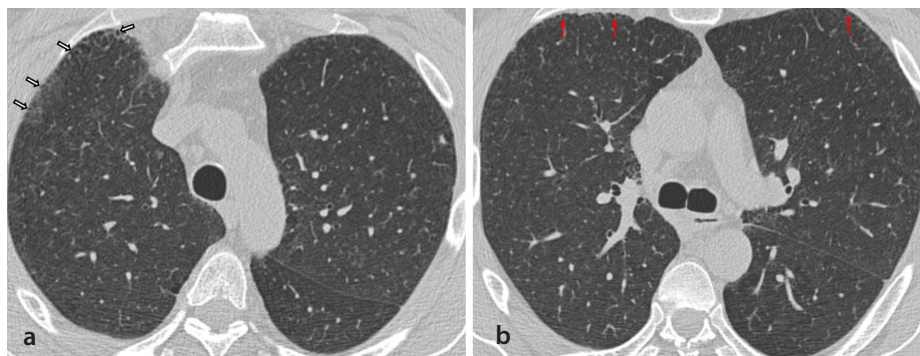


Figure 7. Respiratory bronchiolitis-interstitial lung disease (RB-ILD) pattern on chest computed tomography (CT). Axial CT images (a, b) demonstrate ill-defined centrilobular nodules and patchy, mild ground-glass opacities with upper-lobe predominance, representing the key imaging features of RB-ILD. Associated paraseptal emphysema is present (red arrows), consistent with a smoking-related ILD. Minimal subpleural reticulation and bronchiolectasis (white arrows) are incidental but are not defining features of RB-ILD.

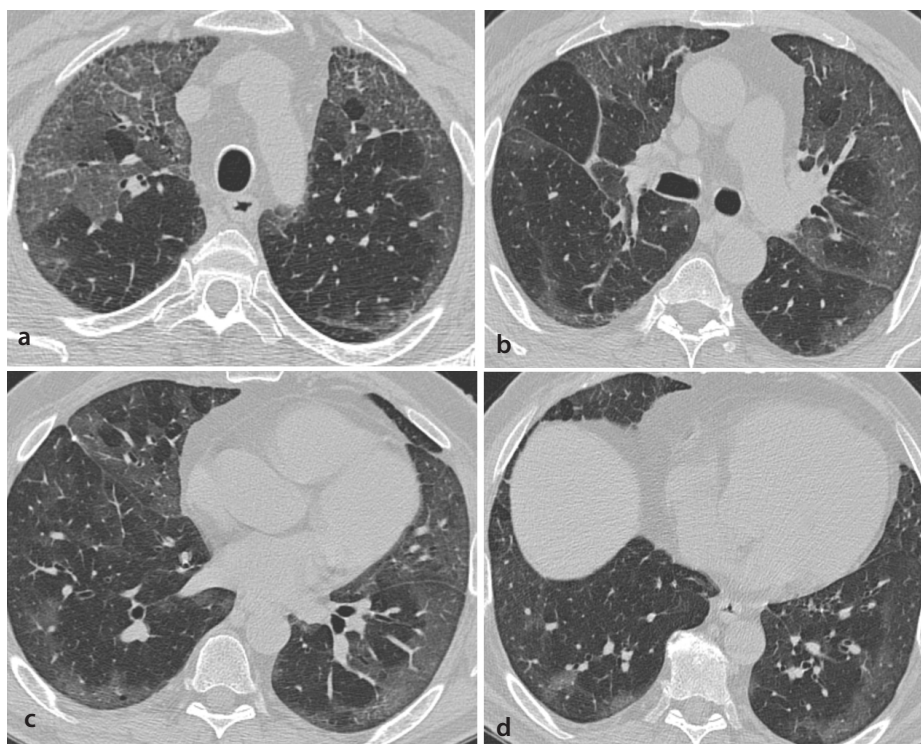


Figure 8. Alveolar macrophage pneumonia pattern, formerly termed desquamative interstitial pneumonia, on chest computed tomography (CT). Axial CT images (a–d) demonstrate diffuse, basilar-predominant ground-glass opacities in a patient with a 16 pack-year smoking history. The diagnosis was histopathologically confirmed.

Lipoid pneumonia (LP) is a rare disorder caused by the accumulation of lipid material within the alveoli, which may be exogenous (e.g., aspiration of mineral oils or fatty substances) or endogenous (associated with airway obstruction, lipid storage disorders, or tissue breakdown).⁷ On CT, LP typically manifests as consolidations or nodules with fat attenuation, often demonstrating negative Hounsfield unit values. These lesions may be focal or multifocal and are sometimes accompanied by GGOs or a crazy-paving pattern when inflammation coexists (Figure 11).¹⁵ Identification of fat density within pul-

monary opacities is highly suggestive of LP and helps distinguish it from infectious or neoplastic processes.

Rare and unclassified disorders

Rare and unclassified ILDs represent a heterogeneous group outside the major interstitial and alveolar filling patterns.⁷ Some cases show overlapping or atypical findings and are labeled unclassifiable ILD.

Pulmonary alveolar microlithiasis (PAM) is a rare genetic disorder caused by mutations in the *SLC34A2* gene, leading to intra-alve-

olar accumulation of calcium phosphate microliths and potentially resulting in restrictive lung physiology. The disease often progresses insidiously, with imaging findings preceding symptoms by years.^{7,16} On CT, PAM is characterized by diffuse, sand-like calcific micronodules that produce a dense “ground-glass” appearance and marked subpleural calcification (Figure 12). The black pleura sign, indicating relative subpleural lucency due to adjacent dense calcification, can be seen.¹⁶

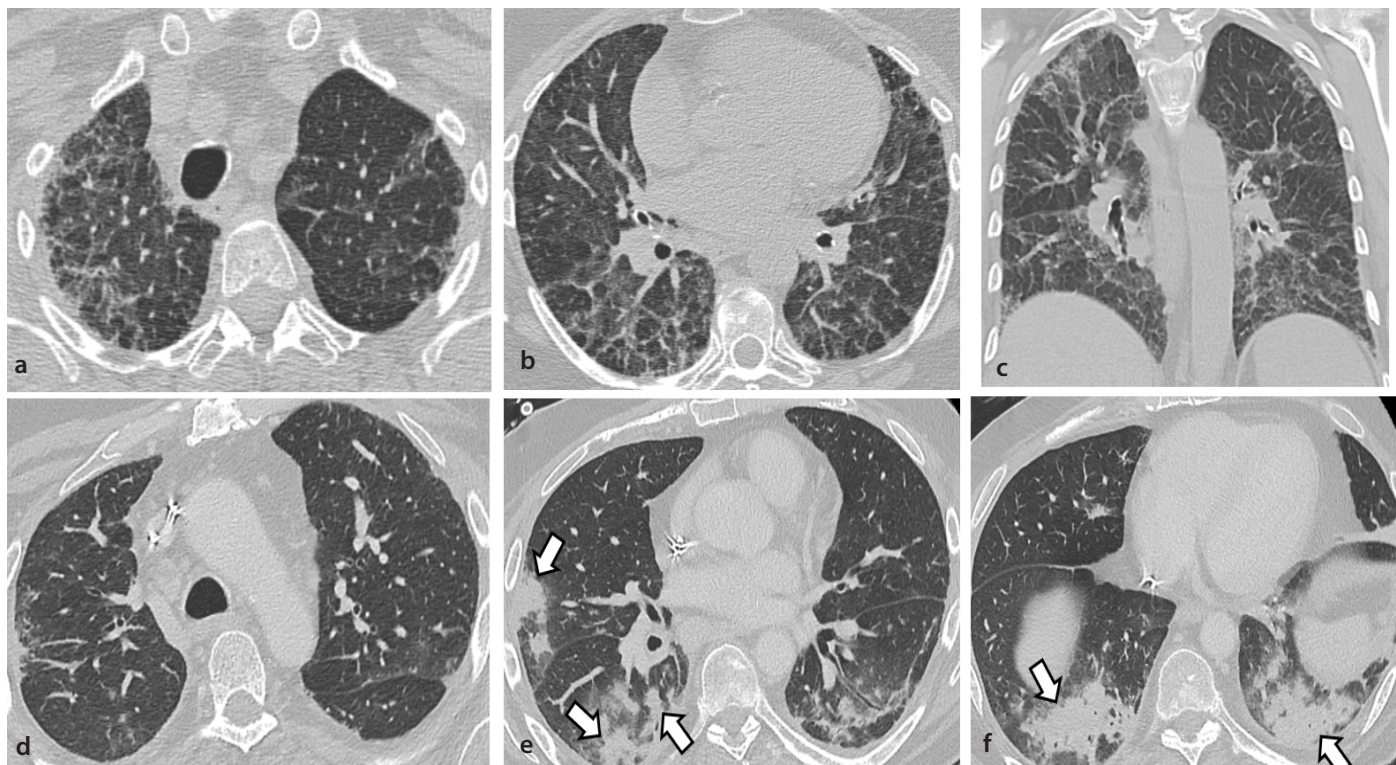


Figure 9. Eosinophilic pneumonia patterns on chest computed tomography (CT). Axial (a, b) and coronal (c) CT images demonstrate chronic eosinophilic pneumonia, characterized by peripheral, upper-lobe-predominant ground-glass opacities, architectural distortion, and reticulations in a patient with a subacute-to-chronic clinical course. Axial CT images (d–f) demonstrate acute eosinophilic pneumonia with bilateral patchy consolidations (arrows) and ground-glass opacities in the setting of acute respiratory symptoms.

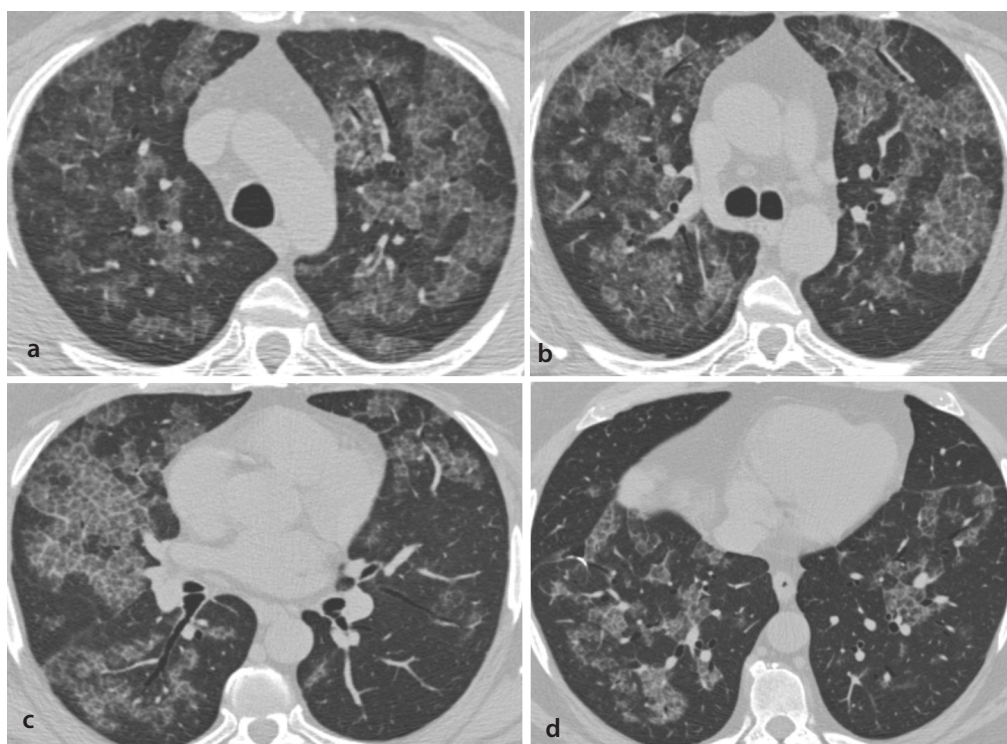


Figure 10. Pulmonary alveolar proteinosis pattern on chest computed tomography (CT). Axial CT images (a–d) demonstrate the characteristic “crazy-paving” pattern, consisting of diffuse or patchy ground-glass opacities superimposed with smooth interlobular septal thickening and intralobular lines. The distribution is bilateral and symmetric.

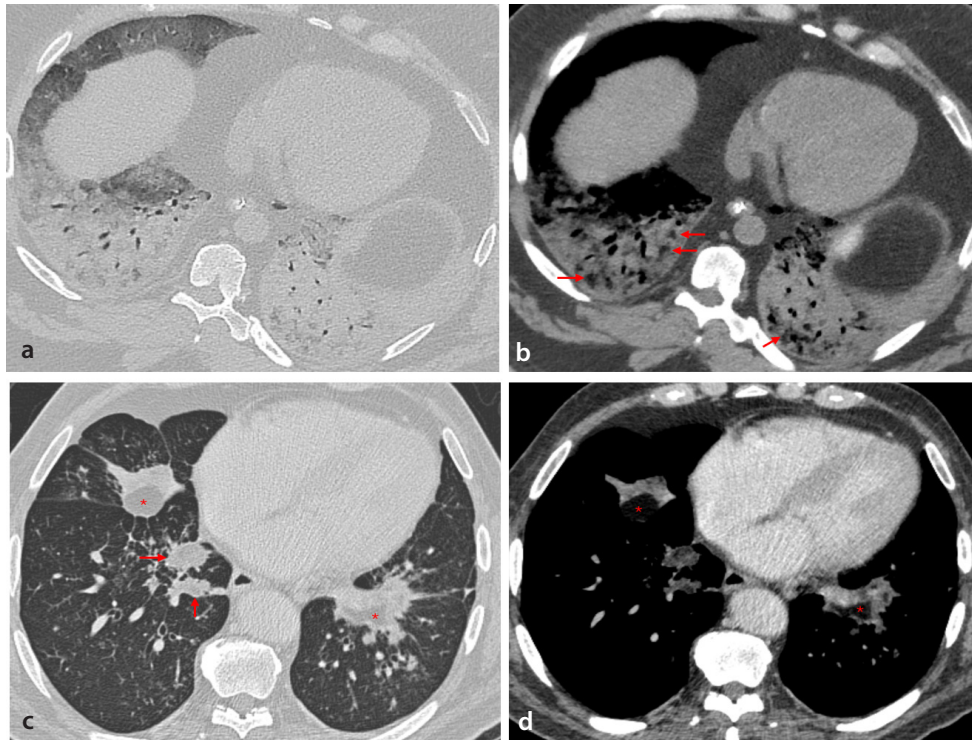


Figure 11. Lipoid pneumonia patterns on chest computed tomography (CT). Axial CT images (a, b) show bilateral lower-lobe consolidations with areas of negative attenuation (arrows) in a patient with acute respiratory failure, consistent with endogenous lipoid pneumonia. Axial CT images (c, d) demonstrate mixed-attenuation masses and nodules containing fat-density areas (asterisks) in a patient with chronic constipation and a history of oral oil-based laxative use, consistent with exogenous lipoid pneumonia due to aspiration.

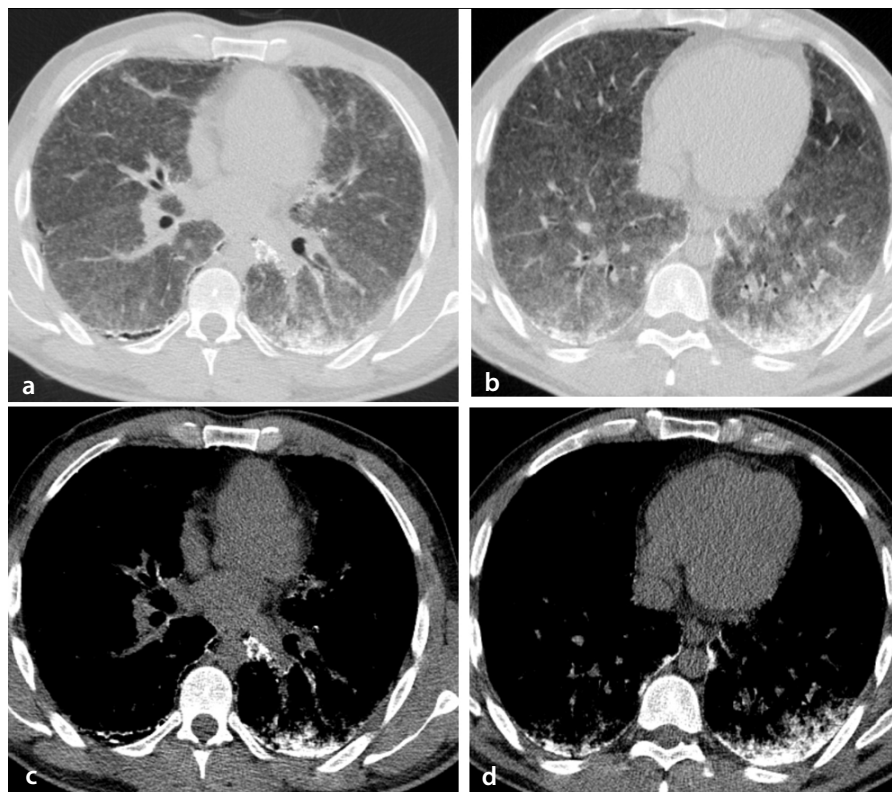


Figure 12. Pulmonary alveolar microlithiasis pattern on chest computed tomography (CT). Axial CT images (a, b) in lung windows show diffuse, sand-like calcific micronodules creating a dense "ground-glass" appearance with subpleural predominance. Corresponding mediastinal window images (c, d) highlight extensive subpleural and parenchymal calcifications.

Pleuroparenchymal fibroelastosis is characterized by dense fibroelastotic thickening of the pleura and adjacent subpleural lung parenchyma, predominantly in the upper lobes. It may arise idiopathically or secondary to transplantation, chemotherapy, autoimmune disease, or recurrent infections.⁵⁻⁷ CT typically shows upper-lobe pleural and subpleural fibrosis with significant volume loss, resulting in the characteristic platythorax (flattened chest) appearance. Additional findings can include traction bronchiectasis, interlobular septal thickening, and honeycombing (Figure 13).^{5,6} Pleuroparenchymal fibroelastosis follows a progressive course with poor prognosis, frequently leading to respiratory failure.

Lymphoid interstitial pneumonia (LIP) is characterized by diffuse lymphoid infiltration of the interstitium, most often associat-

ed with autoimmune diseases, immunodeficiency, or viral infections, although idiopathic cases may occur.⁷ On CT, LIP typically manifests as a combination of thin-walled cysts, GGOs, and interlobular septal thickening, often with lower-lobe predominance (Figure 14). Small nodules and peribronchovascular thickening may also be present.^{17,18} Although some patients respond to immunosuppressive therapy, LIP carries a risk of progression to lymphoma.

Pulmonary amyloidosis results from extracellular deposition of amyloid fibrils within the lung parenchyma, airways, or vasculature and may manifest in either systemic or localized forms.^{7,19} On CT, amyloidosis can appear as calcified or non-calcified parenchymal nodules, tracheobronchial wall thickening, or interstitial infiltration. In cases of interstitial infiltration, imaging findings

include patchy GGOs, smooth or nodular interlobular septal thickening, perilymphatic micronodules, and occasional architectural distortion (Figure 15).^{19,20}

Unclassifiable ILD refers to cases in which a confident diagnosis cannot be established despite comprehensive clinical, radiologic, and pathologic evaluation.⁷ This may result from overlapping features, atypical or indeterminate patterns, or limited biopsy samples. On CT, unclassifiable ILD often shows a mixture of findings, such as fibrotic changes coexisting with OP-like opacities, making assignment to a single category challenging (Supplementary Figure 2).^{5-7,11}

The 2025 ERS/ATS classification offers an updated framework for ILD, emphasizing CT's central role in distinguishing key radiologic patterns with clinical and prognostic relevance (Table 1).

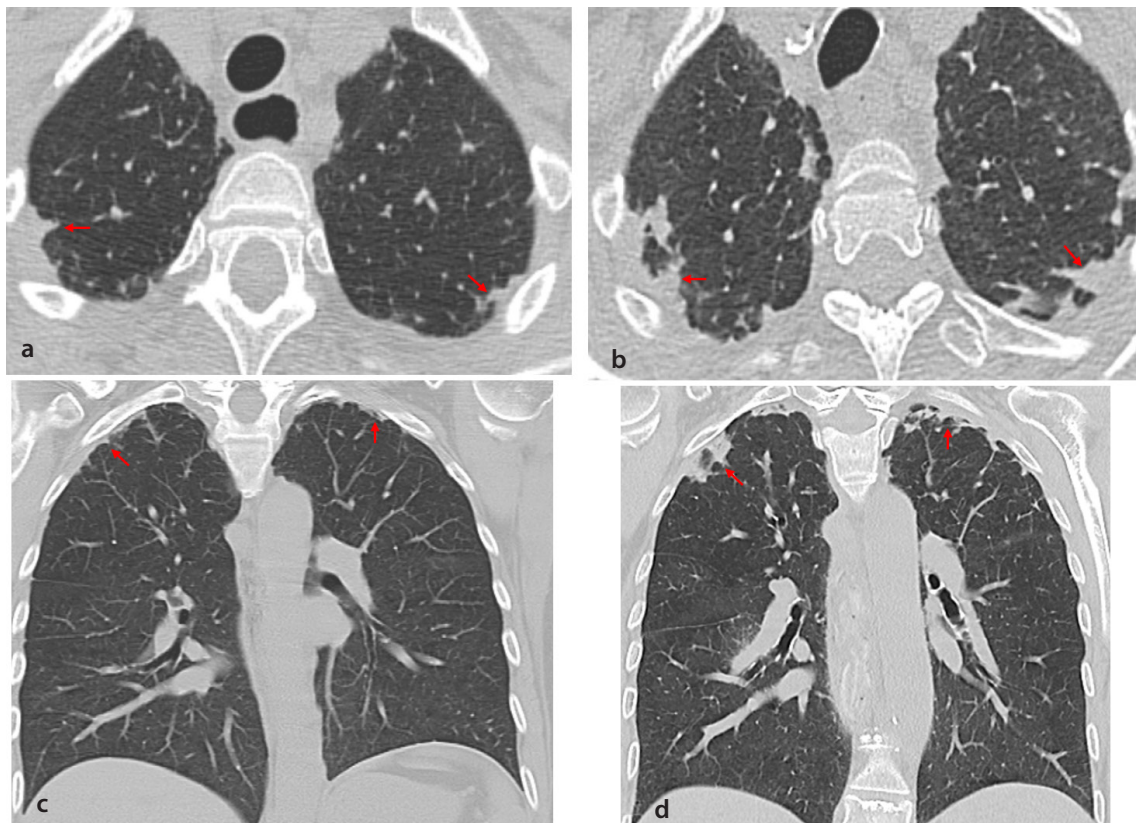


Figure 13. Pleuroparenchymal fibroelastosis pattern on high-resolution chest computed tomography (CT). Initial axial (a) and coronal (c) CT images demonstrate dense pleural and subpleural fibrosis with upper-lobe predominance. Follow-up CT images obtained one year later (b, d) show marked progression of fibrosis with associated volume loss and architectural distortion (arrows).

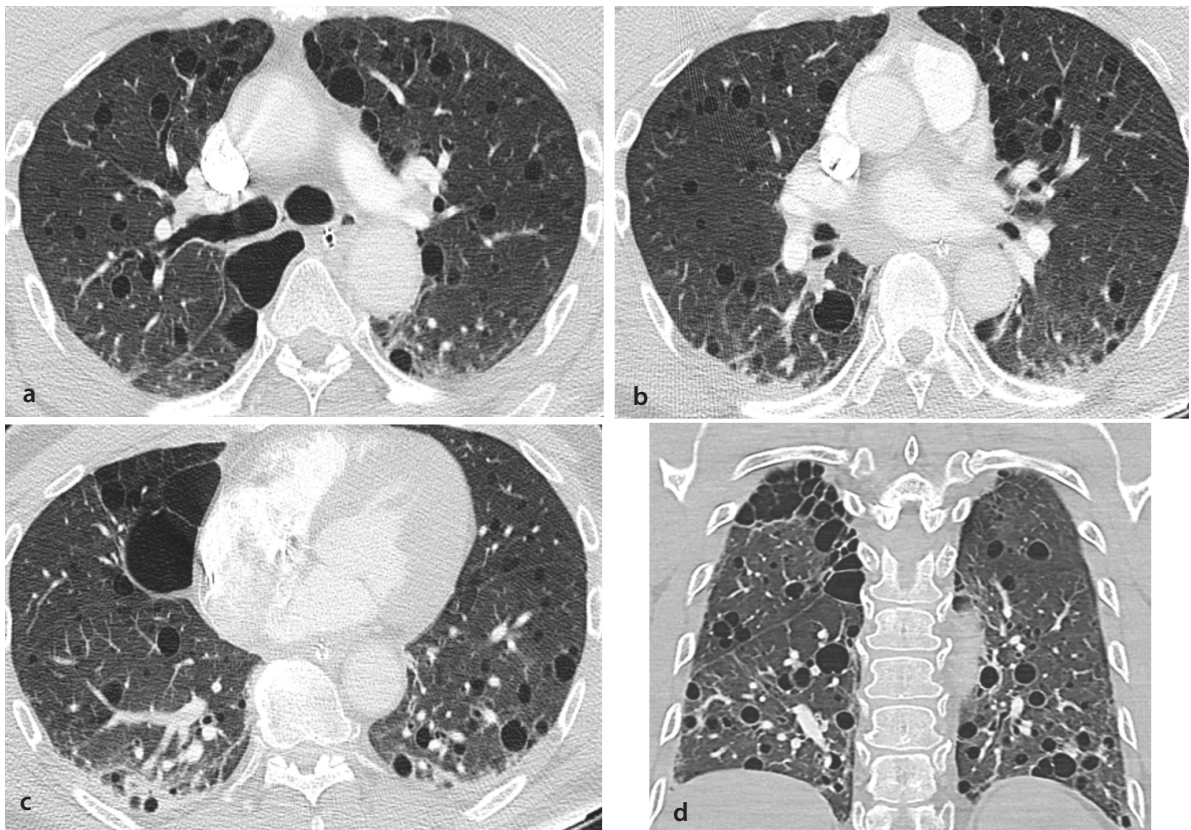


Figure 14. Lymphoid interstitial pneumonia (LIP) pattern on chest computed tomography (CT) in a patient with underlying Sjögren's disease. Axial (a–c) and coronal (d) CT images show numerous thin-walled cysts of variable size, lower-lobe–predominant ground-glass opacities, and interlobular septal thickening, characteristic of LIP.



Figure 15. Pulmonary amyloidosis pattern on chest computed tomography (CT). Axial CT images (a–c) demonstrate nodular interlobular septal thickening, scattered perilymphatic micronodules, and patchy ground-glass opacities. The imaging findings were suggestive of pulmonary amyloidosis in a patient with known systemic disease, and were supported by histopathologic correlation.

Table 1. Radiological subtypes of interstitial, alveolar filling, and rare/unclassified disorders		
Subtype/pattern	Definition/key features	Imaging characteristics [Radiograph/chest computed tomography (CT)]
Interstitial patterns		
Usual interstitial pneumonia	Chronic fibrosing interstitial lung disease (ILD); histopathologic and radiologic hallmark of idiopathic pulmonary fibrosis.	Radiograph: reticular opacities, basal volume loss. CT: basal and subpleural fibrosis, honeycombing, traction bronchiectasis.
Non-specific interstitial pneumonia	Temporally homogeneous interstitial pneumonia with varying degrees of inflammation and fibrosis; includes non-fibrotic and fibrotic subtypes; commonly associated with connective tissue disease (CTD) but may be idiopathic or secondary.	Radiograph: bilateral hazy opacities. CT: diffuse ground-glass opacities (GGOs) with or without fine reticulation, traction bronchiectasis in fibrotic subtype, lower-lobe predominance, possible subpleural sparing.
Bronchiolocentric interstitial pneumonia	Airway-centered interstitial pneumonia characterized by bronchiolocentric inflammation with or without fibrosis; may be seen in hypersensitivity pneumonitis, CTD-associated ILD, aspiration, or drug-related lung disease.	CT: peribronchiolar interstitial thickening or reticulation, centrilobular nodules, tree-in-bud opacities; may be associated with air trapping on expiratory imaging.
Diffuse alveolar damage	Radiologic–pathologic pattern of acute lung injury, seen in acute respiratory distress syndrome (ARDS) and acute exacerbations of ILDs; ARDS represents a clinical syndrome.	Radiograph: diffuse bilateral opacities. CT: diffuse GGOs and patchy consolidation, often with dependent predominance.
Alveolar filling disorders		
Organizing pneumonia (OP)	Intra-alveolar fibroblastic tissue plugs with preservation of lung architecture; idiopathic (cryptogenic OP) or secondary to infection, CTD, drugs, radiation, or aspiration.	Radiograph: patchy, often peripheral opacities. CT: peripheral or peribronchovascular consolidations, perilobular or arcade-like opacities, reversed halo sign.
Respiratory bronchiolitis-ILD	Smoking-related ILD characterized by accumulation of pigmented macrophages within respiratory bronchioles and adjacent alveoli.	CT: ill-defined centrilobular ground-glass nodules and patchy GGOs, often with upper-lobe predominance.
Alveolar macrophage pneumonia (desquamative interstitial pneumonia)	Smoking-related alveolar filling disorder with diffuse intra-alveolar macrophage accumulation; less commonly associated with CTD or drug exposure.	Radiograph: diffuse hazy opacities. CT: diffuse GGOs with basal predominance, mild reticulation, and occasional small cysts.
Eosinophilic pneumonia [acute eosinophilic pneumonia (AEP)/chronic eosinophilic pneumonia (CEP)]	Acute and chronic eosinophilic lung diseases characterized by eosinophilic infiltration of alveoli and interstitium; AEP presents with acute respiratory symptoms, whereas CEP follows a subacute to chronic course.	CT: AEP–diffuse bilateral GGOs and/or consolidation. CEP–peripheral and often upper-lobe–predominant consolidations and GGOs (“photographic negative” of pulmonary edema).
Pulmonary alveolar proteinosis	Intra-alveolar accumulation of surfactant material; most commonly autoimmune in origin.	Radiograph: bilateral perihilar opacities. CT: crazy-paving pattern (GGOs with superimposed septal thickening).
Lipoid pneumonia	Exogenous or endogenous accumulation of lipid material within the alveoli.	CT: consolidations or nodules with fat attenuation (negative Hounsfield units), sometimes with associated GGOs.
Rare and unclassified disorders		
Pulmonary alveolar microlithiasis	Rare genetic disorder with widespread intra-alveolar calcium phosphate microliths.	Radiograph: diffuse sand-like micronodules. CT: diffuse calcified micronodules with subpleural involvement; black pleura sign.
Pleuroparenchymal fibroelastosis	Rare fibrosing ILD characterized by upper-lobe–predominant pleural and subpleural fibroelastosis.	Radiograph: apical pleural thickening. CT: dense subpleural upper-lobe fibrosis, volume loss, platythorax.
Lymphoid interstitial pneumonia	Lymphoproliferative ILD; frequently associated with autoimmune disease or immunodeficiency.	CT: GGOs, interlobular septal thickening, scattered thin-walled cysts.
Pulmonary amyloidosis	Extracellular deposition of amyloid protein within lung parenchyma, airways, or vasculature.	CT: pulmonary nodules with or without calcification, tracheobronchial wall thickening, and interstitial infiltration.
Unclassifiable ILD	ILD with indeterminate or overlapping features that do not fit established categories despite multidisciplinary evaluation.	CT: mixed or overlapping patterns, often combining fibrotic changes with OP-like features.

Footnotes

Conflict of interest disclosure

Furkan Ufuk, MD, serves as Section Editor for Diagnostic and Interventional Radiology. He had no involvement in the peer review of this article and had no access to information regarding its peer review.

Supplementary Figures 1 and 2:
<https://d2v96fxpocvxx.cloudfront.net/beb8919b-f013-4ea1-b1c8-40332e840fe1/content-images/6d601dda-86a1-4f50-8964-b276c887a632.pdf>

References

1. Raghu G, Remy-Jardin M, Richeldi L, et al. Idiopathic pulmonary fibrosis (an update) and progressive pulmonary fibrosis in adults: an official ATS/ERS/JRS/ALAT clinical practice guideline. *Am J Respir Crit Care Med*. 2022;205(9):e18-e47. [\[Crossref\]](#)
2. Rajan SK, Cottin V, Dhar R, et al. Progressive pulmonary fibrosis: an expert group consensus statement. *Eur Respir J*. 2023;61(3):2103187. [\[Crossref\]](#)
3. Marinescu DC, Hague CJ, Muller NL, et al. Integration and application of radiologic patterns from clinical practice guidelines on idiopathic pulmonary fibrosis and fibrotic hypersensitivity pneumonitis. *Chest*. 2023;164(6):1466-1475. [\[Crossref\]](#)
4. Chung JH, Cox CW, Montner SM, et al. CT features of the usual interstitial pneumonia pattern: differentiating connective tissue disease-associated interstitial lung disease from idiopathic pulmonary fibrosis. *AJR Am J Roentgenol*. 2018;210(2):307-313. [\[Crossref\]](#)
5. Lee KS, Han J, Wada N, et al. Imaging of pulmonary fibrosis: an update, from the AJR special series on imaging of fibrosis. *AJR Am J Roentgenol*. 2024;222(2):e2329119. [\[Crossref\]](#)
6. Chelala L, Brixey AG, Hobbs SB, et al. Current state of fibrotic interstitial lung disease imaging. *Radiology*. 2025;316(1):e242531. [\[Crossref\]](#)
7. Ryerson CJ, Adegunsoye A, Piciucchi S, et al. Update of the international multidisciplinary classification of the interstitial pneumonias: an ERS/ATS statement. *Eur Respir J*. 2025;66(6):2500158. [\[Crossref\]](#)
8. Mukhopadhyay S. Usual interstitial pneumonia (UIP): a clinically significant pathologic diagnosis. *Mod Pathol*. 2022;35(5):580-588. [\[Crossref\]](#)
9. Shih AR, Nitiwarangkul C, Little BP, et al. Practical application and validation of the 2018 ATS/ERS/JRS/ALAT and Fleischner Society guidelines for the diagnosis of idiopathic pulmonary fibrosis. *Respir Res*. 2021;22(1):124. [\[Crossref\]](#)
10. Fernández Pérez ER, Travis WD, Lynch DA. Rethinking the CT and histopathological nomenclature of hypersensitivity pneumonitis: unveiling bronchiolocentric patterns of interstitial pneumonia. *Am J Respir Crit Care Med*. 2025;211(10):1738-1740. [\[Crossref\]](#)
11. Marquis KM, Hammer MM, Steinbrecher K, et al. CT approach to lung injury. *Radiographics*. 2023;43(7):e220176. [\[Crossref\]](#)
12. Ufuk F, Ocak I, Chelala L, Landeras L. The reverse halo sign in emergency chest CT: a diagnostic guide across diverse pulmonary conditions. *Emerg Radiol*. 2025;32(3):435-445. [\[Crossref\]](#)
13. Escalon JG, Girvin F. Smoking-related interstitial lung disease and emphysema. *Clin Chest Med*. 2024;45(2):461-473. [\[Crossref\]](#)
14. Trapnell BC, Nakata K, Bonella F, et al. Pulmonary alveolar proteinosis. *Nat Rev Dis Primers*. 2019;5(1):16. [\[Crossref\]](#)
15. Bučić D, Koršić M, Smojver-Ježek S, Hrabak-Paar M. Exogenous lipid pneumonia after long-term nasal petroleum jelly application. *Radiol Cardiothorac Imaging*. 2025;7(1):e240289. [\[Crossref\]](#)
16. Ufuk F. Pulmonary alveolar microlithiasis. *Radiology*. 2021;298(3):567. [\[Crossref\]](#)
17. Kuykendall K, Raptis C, Bhalla S. Interstitial lung disease in connective tissue disease. *Radiol Clin North Am*. 2025;63(4):595-607. [\[Crossref\]](#)
18. Ufuk F. Lymphoid interstitial pneumonia. *N Engl J Med*. 2025;392(8):806. [\[Crossref\]](#)
19. Li F, Li J, Li Y, et al. Imaging and clinical features of nodular pulmonary amyloidosis. *J Thorac Imaging*. 2025;40(5):e0830. [\[Crossref\]](#)
20. Eggleston RH, Hartman TE, Walkoff LA, Yi ES, Ryu JH, Baqir M. Clinical, radiologic, and pathologic features and outcomes of pulmonary transthyretin amyloidosis. *Respir Med*. 2022;194:106761. [\[Crossref\]](#)



The relationship between preoperative pulmonary hypertension risk in liver transplant and vascular and biliary complications in the postoperative period

Burak Yağdıran¹
 Can Ugochkwu Uchehara¹
 Adem Şafak²
 Nedim Çekmen³
 Mehmet Haberal²

¹Başkent University Faculty of Medicine, Department of Radiology, Ankara, Türkiye

²Başkent University Faculty of Medicine, Department of General Surgery, Ankara, Türkiye

³Başkent University Faculty of Medicine, Department of Anesthesiology and Intensive Care Unit, Ankara, Türkiye

PURPOSE

Liver transplantation (LT) is the gold standard definitive treatment for end-stage liver disease (ESLD). In patients with ESLD, preoperative pulmonary hypertension (PHT) secondary to the primary liver pathology is a common complication that causes high mortality and morbidity. This study aimed to determine whether thrombosis, hemorrhage, stenosis, occlusion, and biliary system complications may be indirectly affected by this vascular complication.

METHODS

This study retrospectively screened patients undergoing LT at our hospital between 2017 and 2024. A total of 265 patients were included. Preoperative PHT was assessed using transthoracic echocardiography-derived systolic pulmonary arterial pressure measurements, whereas post-transplant computed tomography scans were reviewed for vascular and biliary complications.

RESULTS

The portal vein complications in the PHT group (17.5%) were higher than those in the non-PHT group (9.4%). In the multivariate logistic regression analysis examining the relationship between PHT level and complications, a significant association with portal vein complications was found. No difference was found between the two groups regarding hepatic vein complications ($P = 0.496$). The incidence of biliary complications in patients with PHT (41.3%) was similar to that in those without PHT (41.1%) ($P = 0.980$). There was no statistically significant relationship between PHT levels (normal, mild, moderate/severe) and the distribution of hepatic artery complications ($P = 0.194$).

CONCLUSION

Moderate-to-severe PHT levels were found to be associated with a significantly increased risk of portal vein complications. We believe caution is warranted, as the presence of preoperative PHT can significantly increase the risk of post-LT complications.

CLINICAL SIGNIFICANCE

Changes in mediator balance in PHT lead to systemic vascular remodeling and, consequently, systemic vasoconstriction. We believe that this condition may negatively affect graft vascularity in patients undergoing LT, thereby increasing vascular and biliary complications. This study found that PHT is associated with the risk of portal vein complications.

KEYWORDS

Preoperative pulmonary hypertension, liver transplantation, complications, biliary system, vascular

Handling editor: Sonay Aydın

Corresponding author: Burak Yağdıran

E-mail: burakyagdiran@gmail.com

Received 21 October 2025; revision requested 27 November 2025; accepted 14 December 2025.



Epub: 29.12.2025

Publication date: 01.09.2026

DOI: 10.4274/dir.2025.253707

Liver transplantation (LT) is the gold standard definitive treatment for end-stage liver disease (ESLD).¹ In patients with ESLD, pulmonary vascular disorders secondary to the primary liver pathology are common complications that cause high mortality and morbidity. These patients have a 25% prevalence of pulmonary vascular complications. The most common of which is portopulmonary hypertension (PoPHT).² Pulmonary hypertension (PHT)

findings are deemed PoPHT if they develop due to portal venous hypertension. The prevalence of PoPHT ranges from 2.5% to 10.5% in the literature, an extremely high rate.^{3,4} The pathogenesis of PoPHT is explained by increased portal venous pressure, leading to hyperdynamic hepatic arterial flow, increased venous blood volume, vasoconstriction in the pulmonary arteries, and neuro-humoral activation. Although the condition shares similar histopathological features with other types of PHT, the severity of PoPHT is not related to the severity or etiology of liver disease.⁵⁻⁷

The presence of PHT, whether it develops due to PoPHT or any other etiology, significantly worsens the prognosis in patients with liver parenchymal disease who undergo LT.⁸ Therefore, identifying PHT in the preoperative period and providing treatment if necessary, as outlined in transplantation guidelines, significantly improves the patient's prognosis.

PHT is a complex systemic disease that affects the entire body and leads to right ventricular failure in its final stage.⁹ The disease is caused by a decrease in vasodilator mediators and an increase in vasoconstrictor mediators.¹⁰ In patients with PHT, endothelial dysfunction occurs first, followed by vascular remodeling, and subsequently by apoptosis, neo-adventitial cell proliferation, and thrombosis in arterioles.¹¹ Regardless of etiology, the imbalance of mediators in patients with PHT leads to vascular remodeling in all systems and, consequently, an increase in mortality and morbidity.¹² Preoperatively, chest X-ray, electrocardiography (ECG), and transthoracic echocardiography (TTE) must be performed. If PHT is suspected, right heart catheterization should be performed to confirm the diagnosis.

TTE is an accurate, reliable, and reproducible tool for estimating systolic pulmonary

arterial pressure (sPAP) in the absence of specific pathologies that elevate right heart pressure, such as pulmonary stenosis or right ventricular outflow tract obstruction.¹³ Due to its non-invasive and easily accessible nature, the tool has high sensitivity as a screening test.¹⁴ The mean PAP (mPAP) value can be largely predicted using the sPAP value obtained with TTE. The formula used for this is as follows: $mPAP: (0.61 * sPAP) + 2$ mmHg. The threshold value of 25 mmHg for mPAP used to define PHT is equivalent to 38 mmHg for sPAP.¹⁵ Patients can be grouped as normal (sPAP < 38 mmHg), mild (sPAP: 38–50 mmHg), moderate (sPAP: 50–70 mmHg), or severe (sPAP > 70 mmHg) based on their sPAP values.¹⁶

Thrombosis generally involves blockage of the lumen of a vessel due to thrombus formation. Stenosis is the narrowing of the vessel's diameter. Occlusion is the complete cessation of blood flow in the vessel.

The primary hypothesis of this study is that thrombosis, hemorrhage, stenosis, occlusion, and biliary system complications may be indirectly affected by these vascular complications. The secondary hypothesis relates to determining the presence of preoperative PHT, superimposed on body mass index (BMI), donor type, cold ischemia time of the graft, presence of early graft dysfunction, need for inotropic support during the perioperative period, total blood loss, and whether there is an increase in the risk of possible vascular and biliary complications when superimposed on the primary etiologies causing liver failure and the presence of preoperative PHT.

Methods

This retrospective study was approved by Başkent University Ethics Committee (protocol number: KA25/256, date: 02.07.2025) and was conducted in accordance with the principles outlined in the Declaration of Helsinki. All patients undergoing LT at our hospital between 2017 and 2024 were retrospectively screened. A total of 282 patients were initially included in this study. Patients with missing data in the archives were excluded from the study (n = 10). Additionally, LT procedures performed due to acute liver failure and cancer were excluded from the study as exclusion criteria, with seven patients undergoing LT due to hepatocellular carcinoma excluded. A total of 265 patients were ultimately included in the study (Figure 1). Preoperative PHT was assessed using TTE based on sPAP measurements, whereas post-transplant computed tomography (CT) examinations were used to evaluate vascular and biliary complications. Preoperative TTE findings, BMI, donor type, graft cold ischemia time, presence of early graft dysfunction, need for inotropic support, total perioperative bleeding volume, primary etiologies of liver failure, and demographic data for these patients were determined.

The sPAP values were determined in the TTE findings of the patients included in the study during the preoperative period, and the patients were grouped as normal, mild, moderate, or severe based on these values. All post-transplant CT scans of the patients were evaluated separately, and all complications in the hepatic artery, portal vein, and

Main points

- Preoperative pulmonary hypertension (PHT) increases the risk of portal vein complications in the liver graft, driven by systemic vascular remodeling and vasoconstriction.
- No independent association was found with hepatic artery, hepatic vein, or biliary complications.
- In patients who are candidates for liver transplantation (LT), detecting PHT preoperatively and planning LT following medical treatment reduces postoperative complications.

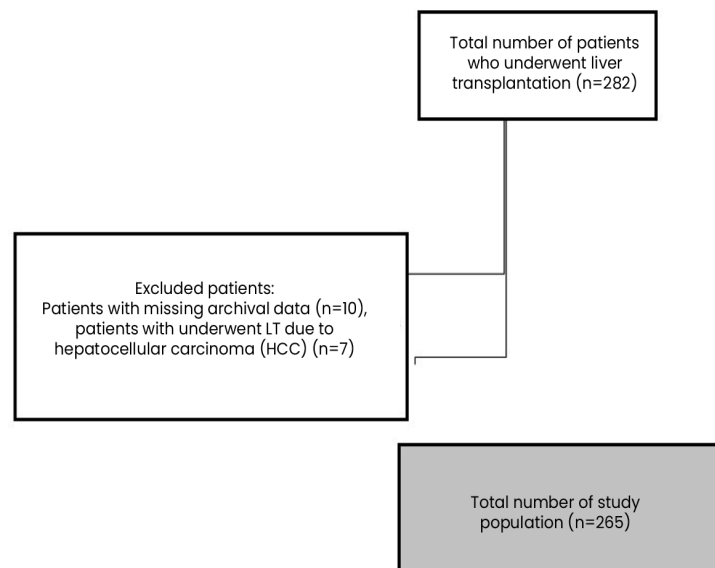


Figure 1. Flow chart of study population. LT, liver transplantation.

hepatic veins (thrombosis, bleeding, stenosis, and occlusion), as well as complications related to the biliary system, were identified. All CT scans were evaluated by a single radiologist with 11 years of experience in abdominal and transplantation radiology.

We included all available variables recorded in our hospital's archives. Operative time, transfusion volume, and intraoperative technical difficulties were inconsistently documented throughout the study and were therefore not included in the analysis.

Statistical analysis

All analyses were performed using IBM SPSS Statistics for Windows, version 26.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics were presented as the median (with quartiles 25%–75%; Q1–Q3) for continuous variables and as number (n) and percentage (%) for categorical variables. The Mann–Whitney U test was used to compare two independent groups that did not show a normal distribution. Categorical variables were compared using Pearson's chi-square test when its assumptions were met, and Fisher's exact test or the Fisher–Freeman–Halton exact test via Monte Carlo simulation when expected cell counts were insufficient. A linear-by-linear association test (Cochran–Armitage trend test) was used to evaluate the association between ordinal PHT severity categories and portal vein thrombosis. For risk assessment, separate multivariable logistic regression models were constructed for each complication type (present/absent) as the dependent variable. The primary independent variable was PHT level (reference: normal). All models were adjusted for recipient age, sex (reference: female), BMI, donor type (reference: deceased donor), and cold ischemia time. Model significance (Omnibus test) and goodness-of-fit (Hosmer–Lemeshow test) statistics were reported. Results were presented as adjusted odds ratios (ORs) with 95% confidence intervals (CIs). A *P* value < 0.05 was considered statistically significant.

Consent to participate and consent to publish

Written consent for participation in the study and for publication of statements was obtained from each patient individually.

Results

A total of 265 individuals were included in the study, with a median age of 8.0 years (Q1–Q3: 1.5–31.0 years), and 66.4% were under 18 years of age. The median BMI was

17.0 kg/m² (14.5–20.5), and 55.1% of participants were male patients; 83.8% of transplants were from living donors. The median cold ischemia time was 70 (60.0–85.0) min, and the median perioperative blood loss was 300 (150–660) mL. Delayed graft dysfunction was detected in 7.9% of cases, and inotropic support was required in 14.0%. In recipients, the distribution of etiologies of liver failure ranked biliary pathologies (28.7%), infection (18.5%), and enzyme deficiency (18.1%) in the top three.

Patients with PHT [median: 40.0 (18.0–55.5) years] were older than those without PHT [6.0 (1.0–16.0) years], and the difference between them was statistically significant (*P* < 0.001). Similarly, 76.2% of those with PHT were over 18 years of age, which was significantly higher than those under 18 years of age with PHT (*P* < 0.001); BMI [18.0 (15.0–24.5)] was significantly higher in patients with PHT than in those without PHT [16.0 (14.0–20.0)] (*P* = 0.013). Male gender was significantly more common in the PHT group (66.7%) than in the non-PHT group (51.5%) (*P* = 0.034). Preoperative PHT was almost twice as common in cadaveric donors (27.0% vs. 12.9%; *P* = 0.008). The amount of bleeding in patients with perioperative PHT [600.0 (175.0–1,000.0) cc] was significantly higher than in those without perioperative PHT [250.0 (100.0–650.0) cc] (*P* = 0.001). Cold ischemia time was similar between patients with and without PHT, with no statistically significant difference [65.0 (60.0–82.5) vs. 70.0 (60.0–90.0) min; *P* = 0.297]. The presence of delayed graft dysfunction (6.3% vs. 8.4%; *P* = 0.791) and the need for inotropic (norepinephrine) infusion (11.1% vs. 14.9%; *P* = 0.455) were similar between the PHT and non-PHT groups. No significant difference was found in the distribution of the etiology of liver failure in recipients (*P* = 0.087). In those with PHT, infection (28.6%) and storage diseases (17.5%) were relatively more common, whereas Budd–Chiari syndrome (0.0%) was not observed. In those without PHT, biliary pathologies (30.2%) and Budd–Chiari syndrome (5.0%) were found to be higher (Table 1).

The distribution of hepatic artery complication types in patients with PHT was similar to that in those without PHT (*P* = 0.762). Although the rate of thrombosis among portal vein complications in the PHT group (17.5%) was higher than that in the non-PHT group (9.4%), the difference between the groups was not statistically significant (*P* = 0.211). No difference was found between the two groups regarding hepatic vein complications

(*P* = 0.494). The incidence of biliary complications in patients with PHT (41.3%) was similar to that in those without PHT (41.1%) (*P* = 0.980). In summary, no significant association was found between preoperative PHT and vascular or biliary complications in the early postoperative period (Table 2).

The relationship between preoperative PHT and postoperative vascular and biliary complications in patients undergoing LT is shown in Table 3. In those without hepatic artery complications, PHT levels were 74.8% in those with normal levels, 66.7% in those with mild levels, and 80.0% in those with moderate/severe levels. In those with hepatic artery thrombosis, the PHT levels were 19.8%, 18.2%, and 20.0%, respectively. In those with hepatic artery bleeding, the levels were 1.0%, 6.1%, and 0.0%, respectively. In those with hepatic artery stenosis, the levels were 4.5%, 9.1%, and 0.0%, respectively. There was no statistically significant relationship between PHT level (normal, mild, moderate/severe) and the distribution of hepatic artery complications (*P* = 0.192).

In the logistic regression analysis (Table 4 and Figure 2), the model for portal vein complications was statistically significant (model: *P* = 0.046; Hosmer–Lemeshow: *P* = 0.793). While no increase in risk was observed for portal vein complications in the mild PHT group (OR: 1.04; 95% CI: 0.39–2.76; *P* = 0.944), a significant increase in risk was found in the moderate/severe PHT group (OR: 3.67; 95% CI: 1.30–10.36; *P* = 0.014). No significant association was found between hepatic artery complications and mild PHT (OR: 1.76; 95% CI: 0.76–4.07; *P* = 0.189) or moderate/severe PHT (OR: 1.25; 95% CI: 0.41–3.80; *P* = 0.700). Similarly, although the overall model for hepatic vein complications was significant (model: *P* = 0.014; Hosmer–Lemeshow: *P* = 0.769), PHT level was not an independent predictor. Hepatic vein complications were not found to be significant in the mild (OR: 0.38, 95% CI: 0.07–1.93, *P* = 0.241) and moderate/severe (OR: 0.45, 95% CI: 0.08–2.67, *P* = 0.380) PHT groups. Regarding biliary complications, no significant association was observed in either the mild (OR: 1.02; 95% CI: 0.47–2.20; *P* = 0.965) or moderate/severe (OR: 0.88; 95% CI: 0.36–2.18; *P* = 0.783) PHT groups.

Postoperative complications such as bleeding and thrombosis are observed as early complications, detected in almost all patients within the first 30 days. On the other hand, stenosis is a slightly longer-term subacute complication, detected between 1 and

Table 1. Demographic and perioperative characteristics according to the presence of PHT in patients [no PHT and PHT present (mild, moderate, and severe)]

	All patients (n = 265)	No PHT (n = 202)	PHT present (n = 63)	P
Age (years) Median (Q1–Q3)	8.0 (1.5–31.0)	6.0 (1.0–16.0)	40.0 (18.0–55.5)	< 0.001^M
Age; n (%)				
<18 years (years)	176 (66.4%)	161 (79.7%)	15 (23.8%)	< 0.001^K
≥18 years (years)	89 (33.6%)	41 (20.3%)	48 (76.2%)	
BMI (kg/m²) Median (Q1–Q3)	17.0 (14.5–20.5)	16.0 (14.0–20.0)	18.0 (15.0–24.5)	0.013^M
Gender; n (%)				
Male	146 (55.1%)	104 (51.5%)	42 (66.7%)	0.034^K
Female	119 (44.9%)	98 (48.5%)	21 (33.3%)	
Donor type; n (%)				
Living donor	222 (83.8%)	176 (87.1%)	46 (73.0%)	0.008^K
Cadaver	43 (16.2%)	26 (12.9%)	17 (27.0%)	
Cold ischemia time (min) Median (Q1–Q3)	70.0 (60.0–85.0)	70.0 (60.0–90.0)	65.0 (60.0–82.5)	0.297 ^M
Bleeding volume (cc) Median (Q1–Q3)	300.0 (150.0–660.0)	250.0 (100.0–650.0)	600.0 (175.0–1,000.0)	0.001^M
Presence of delayed graft function; n (%)				
Absent	244 (92.1%)	185 (91.6%)	59 (93.7%)	0.791 ^F
Present	21 (7.9%)	17 (8.4%)	4 (6.3%)	
Need for inotropic (norepinephrine) infusion; n (%)				
Absent	228 (86.0%)	172 (85.1%)	56 (88.9%)	0.455 ^K
Present	37 (14.0%)	30 (14.9%)	7 (11.1%)	
Etiology of liver failure in recipients; n (%)				
Infection	49 (18.5%)	31 (15.3%)	18 (28.6%)	0.087 ^{MC}
Accumulation	35 (13.2%)	24 (11.9%)	11 (17.5%)	
Cryptogenic	30 (11.3%)	24 (11.9%)	6 (9.5%)	
Enzyme deficiency	48 (18.1%)	39 (19.3%)	9 (14.3%)	
Biliary pathologies	76 (28.7%)	61 (30.2%)	15 (23.8%)	
Alcoholic cirrhosis	4 (1.5%)	2 (1.0%)	2 (3.2%)	
Budd–Chiari syndrome	10 (3.8%)	10 (5.0%)	0 (0.0%)	
Other rare causes	13 (4.9%)	11 (5.4%)	2 (3.2%)	

^M, Mann–Whitney U test; ^K, Pearson's chi-square test; ^F, Fisher's exact test; ^{MC}, Monte Carlo exact test (Fisher–Freeman–Halton); PHT, pulmonary hypertension; BMI, body mass index.

3 months. Occlusion occurs either acutely, due to complete obstruction of the lumen by a thrombus, or chronically, as the stenosis progresses.

In summary, moderate-to-severe PHT levels were only associated with a significant increase in the risk of portal vein complications; no independent association was found with hepatic artery, hepatic vein, or biliary complications.

Discussion

In patients undergoing LT, one of the most common and significant complications during the preoperative period is PHT. In this study, we used sPAP to stratify PHT risk be-

cause patients undergoing LT receive routine preoperative ECG examinations. We believed that risk stratification using sPAP would be more effective because it avoids the invasive nature of cardiac catheterization and does not increase contrast or radiation exposure. Krowka and Edwards⁶ showed that 5% of patients with cirrhosis met the hemodynamic criteria for PoPHT. Humbert et al.¹⁴ reported a PoPHT prevalence of 9.4% based on data from a French PHT Association registry. A study by Kia et al.⁸ demonstrated that high sPAP, as determined by TTE, was associated with a significant increase in mortality and morbidity following LT. Data from the Multicenter Liver Transplant Database showed a 36% mortality rate in patients with PoPHT

who underwent LT.¹⁷ In the present study, patients with PHT were older and had higher BMIs than those without PHT. It is expected that age and obesity, in addition to the primary disease, may cause comorbidity and lead to more complications in patients with liver failure, and our study supports this.

Patients undergoing LT from cadaveric donors have a higher risk of complications and PHT compared with patients undergoing LT from living donors. This is because in the latter, the procedure is performed under more elective conditions, with the patient's clinical and laboratory data closely monitored and intervention performed if necessary, ensuring that the procedure is performed under the best possible conditions. Naturally, the

Table 2. Distribution of PHT status according to vascular and biliary complications				
	All patients (n = 265)	No PHT (n = 202)	PHT present (n = 63)	P
Hepatic artery complications; n (%)				
Normal	197 (74.3%)	151 (74.8%)	46 (73.0%)	0.762 ^{MC}
Thrombosis	52 (19.6%)	40 (19.8%)	12 (19.0%)	
Bleeding	4 (1.5%)	2 (1.0%)	2 (3.2%)	
Stricture	12 (4.5%)	9 (4.5%)	3 (4.8%)	
Portal vein complications; n (%)				
Normal	207 (78.1%)	161 (79.7%)	46 (73.0%)	0.211 ^K
Thrombosis	30 (11.3%)	19 (9.4%)	11 (17.5%)	
Stricture	28 (10.6%)	22 (10.9%)	6 (9.5%)	
Hepatic vein complications; n (%)				
Normal	239 (90.2%)	180 (89.1%)	59 (93.7%)	0.494 ^{MC}
Thrombosis	2 (0.8%)	2 (1.0%)	0 (0.0%)	
Stricture	24 (9.1%)	20 (9.9%)	4 (6.3%)	
Biliary complications; n (%)				
None	156 (58.9%)	119 (58.9%)	37 (58.7%)	0.980 ^K
Present	109 (41.1%)	83 (41.1%)	26 (41.3%)	

^K, Pearson's chi-square test; ^{MC}, Monte Carlo exact test (Fisher–Freeman–Halton; PHT, pulmonary hypertension.

^K, Pearson's chi-square test; ^{MC}, Monte Carlo exact test (Fisher–Freeman–Halton); PHT, pulmonary hypertension.

Table 3. Distribution of vascular and biliary complications according to PHT level				
	PHT			<i>P</i> ^K
	Normal n (%)	Mild n (%)	Moderate/severe n (%)	
Hepatic artery complications				
Normal	151 (74.8%)	22 (66.7%)	24 (80.0%)	0.192 ^{MC}
Thrombosis	40 (19.8%)	6 (18.2%)	6 (20.0%)	
Bleeding	2 (1.0%)	2 (6.1%)	0 (0.0%)	
Stricture	9 (4.5%)	3 (9.1%)	0 (0.0%)	
Portal vein complications				
Normal	161 (79.7%)	27 (81.8%)	19 (63.3%)	0.037 ^{MC}
Thrombosis	19 (9.4%)	2 (6.1%)	9 (30.0%)	
Stricture	22 (10.9%)	4 (12.1%)	2 (6.7%)	
Hepatic vein complications				
Normal	180 (89.1%)	31 (93.9%)	28 (93.3%)	0.810 ^{MC}
Thrombosis	2 (1.0%)	0 (0.0%)	0 (0.0%)	
Stricture	20 (9.9%)	2 (6.1%)	2 (6.7%)	
Biliary complications				
Absent	119 (58.9%)	19 (57.6%)	18 (60.0%)	0.981 ^K
Present	83 (41.1%)	14 (42.4%)	12 (40.0%)	

^K, Pearson's chi-square test; ^{MC}, Monte Carlo exact test (Fisher–Freeman–Halton); PHT, pulmonary hypertension.

^K, Pearson's chi-square test; ^{MC}, Monte Carlo exact test (Fisher–Freeman–Halton); PHT, pulmonary hypertension.

graft's cold ischemia time is also shorter in these patients. However, most of these possibilities are not available in LT surgeries performed with cadaveric donors. Therefore, these patients are more prone to developing complications. In parallel with our study, Morioka et al.¹⁸ compared LT from living donors with that from cadaveric donors and found more postoperative complications in the cadaveric LT group. Furthermore, in addition to donor type, this study emphasized that BMI

and patient age are extremely important independent predictors of complications, consistent with our results.

No significant differences were found between patient groups in the distribution of primary etiological diseases causing ESLD. As stated separately in the studies by Bozbas and Bozbas¹⁹ and Li et al.,²⁰ this finding supports the notion that the severity and etiology of liver disease do not affect the severity of PHT.

Meng et al.²¹ found that vascular complications, including thrombosis, bleeding, and occlusion, were detected at similar rates in patients with and without PHT, with no significant difference between the groups. Our study yielded similar results, with no significant difference detected between the types of vascular complications.

In the internal evaluation of patients with PHT, a statistically significant increase in the risk of portal vein thrombosis was detected

Variable	Portal vein complication		Hepatic artery complication		Hepatic vein complication		Biliary complication	
	<i>P</i>	OR (95% CI)	<i>P</i>	OR (95% CI)	<i>P</i>	OR (95% CI)	<i>P</i>	OR (95% CI)
PHT status								
PHT normal (ref)					0.413	—	0.957	—
PHT mild	0.944	1.036 (0.389–2.761)	0.189	1.756 (0.758–4.068)	0.241	0.377 (0.074–1.926)	0.965	1.017 (0.471–2.196)
PHT moderate and severe	0.014	3.665 (1.296–10.360)	0.700	1.246 (0.408–3.804)	0.380	0.451 (0.076–2.671)	0.783	0.881 (0.356–2.175)
Recipient age (years)	0.162	0.986 (0.967–1.006)	0.077	0.984 (0.966–1.002)	0.048	1.024 (1.000–1.049)	0.709	1.003 (0.988–1.017)
Gender								
Female (ref)								
Male	0.817	0.931 (0.509–1.704)	0.895	1.039 (0.585–1.847)	0.044	0.395 (0.160–0.975)	0.659	0.892 (0.538–1.480)
BMI (kg/m²)	0.863	1.003 (0.968–1.039)	0.208	1.020 (0.989–1.053)	0.823	1.005 (0.959–1.054)	0.643	0.993 (0.963–1.023)
Donor type								
Cadaveric (ref)								
Living donor	0.922	0.946 (0.316–2.835)	0.747	0.828 (0.262–2.613)	0.997	0.000 (0.00–0.00)	0.951	1.029 (0.419–2.524)
Cold ischemia (min)	0.991	1.000 (0.997–1.003)	0.204	0.998 (0.995–1.001)	0.969	1.000 (0.995–1.005)	0.963	1.000 (0.998–1.002)
Model <i>P</i> (Omnibus test)		0.046		0.135		0.014		0.499
Hosmer–Lemeshow <i>P</i> (goodness-of-fit)		0.793		0.442		0.769		0.813

PHT, pulmonary hypertension; BMI, body mass index; OR, odds ratio; CI, confidence interval; Ref, reference; min, minimum.

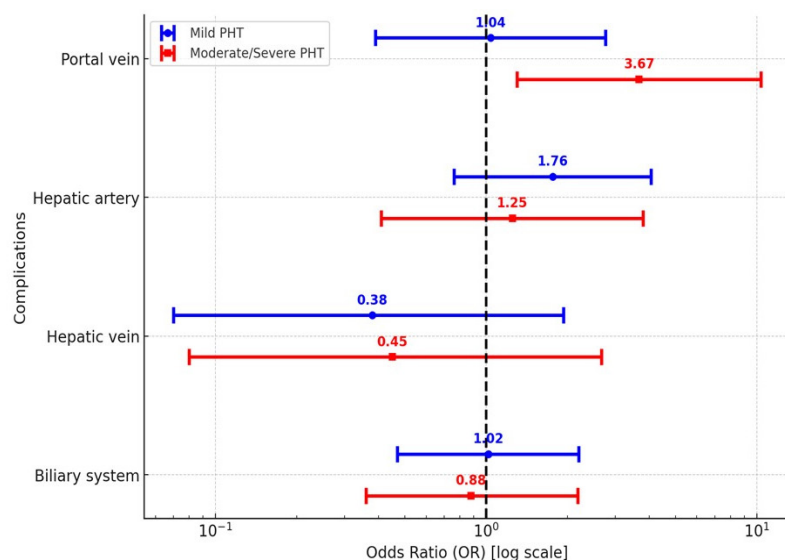


Figure 2. Forest plot showing odds ratios with 95% confidence intervals for vascular and biliary complications according to pulmonary hypertension (PHT) severity.

as the severity of PHT increased. We were unable to find a similar result in our literature review. No significant differences in other types of complications or complications involving other vascular structures were detected according to PHT severity. There are

significant gaps in the literature on this subject. The most significant aspect of our study is that our primary and secondary hypotheses address topics that have received little attention in the existing literature. We believe this study will make a valuable contribution

to the field; however, randomized controlled trials of these hypotheses involving larger groups are needed.

The most important limitations of this study are inherent to retrospective cohort studies. First, the parameters and information we analyzed are based solely on the completeness of medical records and may contain inaccuracies. Second, due to our small patient group, our evaluation is limited. Therefore, we believe it would be beneficial to re-examine this issue with a larger sample. Another important limitation is that right heart catheterization is not performed prior to LT at our institution in patients with ESLD. In our study, PHT groups were determined based on ECG examination. Furthermore, the lack of standardized operative variables (operational time, transfusion volume, technical difficulties) is a limitation of this retrospective study. While we will prioritize this in future studies, we were unable to assess this in this study. The last limitation is that including adult and pediatric patients in the same cohort may introduce residual confounding, even after adjusting for age. Future studies should be designed prospectively with adequate patient numbers to address this issue.

In conclusion, the presence of preoperative PHT causes a significant increase in risk for most complications. In this study, we found that moderate-to-severe PHT levels were associated only with a significant increase in the risk of portal vein complications. No independent association was found with hepatic artery, hepatic vein, or biliary complications. If preoperative PHT is detected during follow-up, we believe that caution should be exercised regarding portal venous system complications in the postoperative period.

Footnotes

Conflict of interest disclosure

The authors declared that there is no conflict of interest.

References

- Mazzaferro V, Regalia E, Doci R, et al. Liver transplantation for the treatment of small hepatocellular carcinomas in patients with cirrhosis. *N Engl J Med*. 1996;334(11):693-699. [\[Crossref\]](#)
- Starkel P, Vera A, Gunson B, Mutimer D. Outcome of liver transplantation for patients with pulmonary hypertension. *Liver Transpl*. 2002;8(4):382-388. [\[Crossref\]](#)
- Singh SA, Shrivastava P, Agarwal A, et al. LTSI Consensus Guidelines: preoperative pulmonary evaluation in adult liver transplant recipients. *J Clin Exp Hepatol*. 2023;13(3):523-531. [\[Crossref\]](#)
- Bozbas SS, Eyuboglu F. Evaluation of liver transplant candidates: a pulmonary perspective. *Ann Thorac Med*. 2011;6(3):109-114. [\[Crossref\]](#)
- Edwards BS, Weir EK, Edwards WD, Ludwig J, Dykoski RK, Edwards JE. Coexistent pulmonary and portal hypertension: morphologic and clinical features. *J Am Coll Cardiol*. 1987;10(6):1233-1238. [\[Crossref\]](#)
- Krowka MJ, Edwards WD. A spectrum of pulmonary vascular pathology in portopulmonary hypertension. *Liver Transpl*. 2000;6(2):241-242. [\[Crossref\]](#)
- Le Pavec J, Souza R, Herve P, et al. Portopulmonary hypertension: survival and prognostic factors. *Am J Respir Crit Care Med*. 2008;178(6):637-643. [\[Crossref\]](#)
- Kia L, Shah SJ, Wang E, et al. Role of pretransplant echocardiographic evaluation in predicting outcomes following liver transplantation. *Am J Transplant*. 2013;13(9):2395-2401. [\[Crossref\]](#)
- Rosenkranz S, Howard LS, Gomberg-Maitland M, Hoepfer MM. Systemic consequences of pulmonary hypertension and right-sided heart failure. *Circulation*. 2020;141(8):678-693. [\[Crossref\]](#)
- Morrell NW, Aldred MA, Chung WK, et al. Genetics and genomics of pulmonary arterial hypertension. *Eur Respir J*. 2019;53(1):1801899. [\[Crossref\]](#)
- Vonk Noordegraaf A, Chin KM, Haddad F, et al. Pathophysiology of the right ventricle and of the pulmonary circulation in pulmonary hypertension: an update. *Eur Respir J*. 2019;53(1):1801900. [\[Crossref\]](#)
- Çekmen N, Gökdemir BN. Anesthesia and pulmonary hypertension: a narrative review. *Heart Mind*. 2024;8:60-73. [\[Crossref\]](#)
- Krowka MJ, Miller DP, Barst RJ, et al. Portopulmonary hypertension: a report from the US-based REVEAL Registry. *Chest*. 2012;141(4):906-915. [\[Crossref\]](#)
- Humbert M, Sitbon O, Chaouat A, et al. Pulmonary arterial hypertension in France: results from a national registry. *Am J Respir Crit Care Med*. 2006;173(9):1023-1030. [\[Crossref\]](#)
- Chemla D, Humbert M, Sitbon O, Montani D, Hervé P. Systolic and mean pulmonary artery pressures: are they interchangeable in patients with pulmonary hypertension? *Chest*. 2015;147(4):943-950. [\[Crossref\]](#)
- Raymond RJ, Hinderliter AL, Willis PW, et al. Echocardiographic predictors of adverse outcomes in primary pulmonary hypertension. *J Am Coll Cardiol*. 2002;39(7):1214-1219. [\[Crossref\]](#)
- Krowka MJ, Mandell MS, Ramsay MA, et al. Hepatopulmonary syndrome and portopulmonary hypertension: a report of the multicenter liver transplant database. *Liver Transpl*. 2004;10(2):174-182. [\[Crossref\]](#)
- Morioka D, Egawa H, Kasahara M, et al. Outcomes of adult-to-adult living donor liver transplantation: a single institution's experience with 335 consecutive cases. *Ann Surg*. 2007;245(2):315-325. [\[Crossref\]](#)
- Bozbas SS, Bozbas H. Portopulmonary hypertension in liver transplant candidates. *World J Gastroenterol*. 2016;22(6):2024-2029. [\[Crossref\]](#)
- Li X, Barbier L, Ferrandière M, Remerand F, Salamé E, Plantier L. Preoperative DLCO predicts severe early complications after liver transplantation. *Respir Med Res*. 2024;85:101089. [\[Crossref\]](#)
- Meng XY, Chen MY, Pan ZY, Lu YF, Wei W, Lu YG. Living donor liver transplantation in biliary atresia children with pulmonary hypertension. *Int J Med Sci*. 2019;16(9):1215-1220. [\[Crossref\]](#)



©Copyright 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Transperineal ultrasonography with a high-frequency probe and a three-dimensional end-fire probe for imaging perianal abscesses and locating the internal openings of anal fistulas

Hongyun Wu
 Juan Chang
 Xiuhong Zhang
 Yunfei Tao

Rugao Hospital of Traditional Chinese Medicine, Clinic
of Ultrasound, Rugao, China

PURPOSE

This study aims to explore the clinical value of transperineal ultrasonography with a high-frequency probe and a three-dimensional (3D) end-fire probe in locating the internal openings of perianal abscesses and anal fistulas.

METHODS

Eighty patients who were diagnosed with perianal abscesses and anal fistulas and treated between June 2023 and December 2024 were enrolled in this study. An observation group (n = 40) and a control group (n = 40) were established, and intergroup comparisons were conducted on the diagnostic efficacy of the intervention in locating internal openings and categorizing abscesses according to the Parks classification system.

RESULTS

The two groups showed no significant differences in baseline data ($P > 0.05$). The observation group showed higher sensitivity, specificity, and overall accuracy in the localization of internal openings than the control group. There was good agreement between the independent image assessments of the two study sonographers (Cohen's kappa = 0.83, $P < 0.001$). The recurrence rate was 7.5% (3/40) in the observation group and 25.0% (10/40) in the control group at 6 months after surgery. The observation group demonstrated higher accuracy in the categorization of abscesses based on the Parks classification system than the control group (92.5% vs. 77.5%).

CONCLUSION

Transperineal ultrasonography using both a high-frequency probe and a 3D end-fire probe can enhance the accurate localization of the internal openings of perianal abscesses and anal fistulas and assist in the accurate classification of abscesses.

CLINICAL SIGNIFICANCE

The combined dual-probe ultrasonography method is a non-invasive, rapid, and highly accurate preoperative imaging approach that aids in precise surgical planning, reduces postoperative recurrence, and offers a practical diagnostic alternative for institutions without access to magnetic resonance imaging.

KEYWORDS

Anal fistula, perianal abscess, probe, surgery, transperineal ultrasonography

Handling editor: Sonay Aydın

Corresponding author: Hongyun Wu

E-mail: wuhyrhtcm@sdsch.cn

Received 11 November 2025; revision requested 30
December 2025; last revision requested 24 January 2026;
accepted 02 February 2026.



Epub: 06.03.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263740

Perianal abscesses are a common emergency condition in anorectal surgery characterized by a high risk of recurrence and postoperative fistula formation. Surgical success largely depends on the accurate preoperative localization of the abscess cavity and its internal openings, as imprecise identification often results in incomplete drainage and disease recurrence.¹ Therefore, reliable imaging assessment is of great importance for improving surgical outcomes in emergency settings.²

You may cite this article as: Wu H, Chang J, Zhang X, Tao Y. Transperineal ultrasonography with a high-frequency probe and a three-dimensional end-fire probe for imaging perianal abscesses and locating the internal openings of anal fistulas. *Diagn Interv Radiol.* 2026;32(5):662-668.

Physical examination remains the most readily available diagnostic approach for locating perianal abscesses but shows limited accuracy in identifying high and complex abscesses and internal openings, as it relies heavily on clinical experience and is especially challenging in primary hospitals.³ This is challenging because it relies heavily on clinical experience, which is lacking in many primary hospitals. Magnetic resonance imaging (MRI) is widely recognized as the gold standard for evaluating perianal infections, as it provides excellent visualization of the spatial relationship between the abscess cavity and the sphincter complex, thus facilitating surgical planning. However, its application in acute clinical practice is restricted by its high cost, long examination times, and limited accessibility in emergency departments.⁴

Ultrasound-based techniques offer a more practical alternative. Intraluminal ultrasonography can accurately detect internal openings and fistula structures, but its clinical utility in the acute abscess stage is limited by poor patient tolerance, severe pain, and the technical difficulty associated with probe insertion.^{5,6} Two-dimensional (2D) transperineal ultrasound (TPUS), by contrast, is non-invasive, convenient, and well-suited for emergency use.⁷ Nevertheless, conventional 2D TPUS alone lacks sufficient capability to depict the three-dimensional (3D) spatial relationship between deep abscess cavities and the anal sphincter complex, which may result in the underdiagnosis of high and complex lesions.⁸

Main points

- The combined use of high-frequency linear array and three-dimensional end-fire probes in transperineal ultrasonography markedly improves the accurate localization of the internal openings of perianal abscesses and anal fistulas.
- The dual-probe approach achieves higher diagnostic sensitivity, specificity, and overall accuracy compared with conventional two-dimensional ultrasonography.
- Accurate preoperative imaging reduces postoperative recurrence rates and enhances the accurate use of the Parks classification system for abscesses.
- Improved imaging guidance contributes to shorter surgical durations, less intraoperative blood loss, milder postoperative pain, and faster wound healing.
- This non-invasive, convenient, and reliable method offers a practical imaging option for preoperative assessment in primary hospitals lacking magnetic resonance imaging resources.

Previous studies have attempted combined ultrasound approaches, most commonly integrating intraluminal ultrasonography with 2D or 3D imaging.^{9,10} Although these techniques improve anatomical visualization, they still depend on intracavitary probe insertion and therefore remain unsuitable for patients in the acute inflammatory phase. In addition, existing TPUS protocols typically rely on a single probe type, limiting examination to either a superficial or deep anatomical assessment.¹¹ As a result, a fully transperineal, pain-tolerant, combined ultrasound strategy has not yet been systematically established.

To address these limitations, the present study introduces a combined TPUS protocol that integrates a high-frequency linear-array 2D probe with a 3D end-fire probe. The high-frequency linear probe provides a high-resolution visualization of superficial fistulas and subcutaneous tissues, whereas the 3D end-fire probe enables the volumetric reconstruction of deep abscess cavities and their relationship to the anal canal, sphincter complex, and levator ani muscle.^{12,13} Unlike previously reported combined ultrasound techniques, this protocol avoids intraluminal probe insertion, thereby improving patient tolerance in the acute stage while achieving a complementary superficial–deep anatomical assessment.

The aim of this study is to systematically characterize the imaging characteristics of perianal abscesses using this combined TPUS approach and to evaluate its accuracy in locating internal openings using surgical findings as the reference standard. This strategy is expected to fill an important clinical gap and provide a non-invasive, accessible, and reliable imaging tool for emergency and primary hospital settings.

Methods

Sample size estimation

Sample size estimation was performed for a two-sample comparison of mean diagnostic accuracy, with accuracy treated as a continuous variable. The estimates of variability and effect size were derived from a pre-experimental pilot study conducted prior to formal enrollment. In this pilot analysis, the mean diagnostic accuracy was 0.78 ± 0.15 in the conventional TPUS group and 0.90 ± 0.12 in the combined TPUS group. Based on these data, the pooled standard deviation (σ) was estimated to be 0.14, and the expected difference in mean accuracy (δ) was 0.12. With a

two-sided significance level (α) of 0.05 and a statistical power ($1 - \beta$) of 0.90, the required sample size was calculated to be 32 patients per group. Allowing for an anticipated loss-to-follow-up rate of approximately 20%, the final target sample size was increased to 40 patients per group, which resulted in a total minimum sample size of 80 participants.

Study type and participants

Study participants ($n = 80$) were patients diagnosed with perianal abscesses and anal fistulas and treated in the study hospital between June 2023 and December 2024. All eligible patients were screened according to the inclusion and exclusion criteria described below. Those meeting the requirements were informed of the study protocol and provided their written consent before participation. After baseline data were collected, the 80 enrolled patients were randomly assigned to either an observation group or a control group using a computer-generated random number table. Allocation concealment was achieved using sealed, opaque, sequentially numbered envelopes prepared by an independent investigator who was not involved in patient recruitment or outcome assessment. Group assignment was revealed only after patient enrollment and the completion of a baseline evaluation. Each group included 40 patients.

All participants completed a standardized preoperative assessment, received their respective interventions, and were followed up postoperatively for outcome evaluation (Figure 1). The study protocol was approved by the Ethics Committee of Rugao Hospital of Traditional Chinese Medicine (approval number: RGSZYLL26002, date: 03.06.2023).

The following inclusion criteria were applied: 1) Patients who were clinically diagnosed with perianal abscesses and anal fistulas and planned to receive surgical treatment in the study hospital; 2) patients with normal anal appearance and function; 3) patients with hepatic and renal function meeting relevant treatment criteria; and 4) patients who required and could undergo examination and diagnosis by ultrasonography.

The following exclusion criteria were applied: 1) Patients with anal fistulas secondary to other diseases; 2) patients with specific anal fistulas caused by Crohn's disease, ulcerative colitis, tuberculosis, among other conditions; 3) patients with severe cardiovascular, cerebrovascular, hepatic, renal, or hematological diseases not suitable for surgery; and 4) patients with other diseases of the anus or rectum.

Ultrasonography methods

The RS85 Prestige ultrasound system (Samsung, Seoul, Korea) was used to perform ultrasonography on all participants. Patients were asked to empty their bladder and bowels prior to examination. No bowel preparation was especially required. All ultrasonographic examinations were performed by two experienced sonographers who were not involved in patient recruitment or postoperative outcome assessment. Due to the inherent differences in examination procedure between the two protocols, the sonographers were aware of the ultrasonographic technique applied; however, they were blinded to surgical findings and postoper-

ative outcomes. To minimize interpretation and verification bias, ultrasonographic findings were recorded in a standardized manner and stored independently, and no feedback regarding surgical results was provided to the sonographers during the study period. Importantly, ultrasonographic findings, including the predicted locations of internal openings, were not disclosed to the operating surgeons prior to surgery.

In the control group, 2D high-frequency transperineal ultrasonography was applied following a standardized scanning protocol. The relevant ultrasonographic features of perianal abscesses and associated fistulous structures were recorded, including abscess

location, size, and morphology and the estimated locations of internal openings. Detailed anatomical descriptions and scanning planes are provided in the Supplementary Material.

In the observation group, a combined transperineal ultrasonography protocol integrating 2D high-frequency linear array imaging and 3D end-fire ultrasonography was applied. 2D transperineal scanning was performed first, as in the control group, and 3D volumetric acquisition was then conducted using an end-fire probe. The combined approach allowed for the complementary assessment of both superficial and deep lesion characteristics. Detailed procedural steps, anatomical landmarks, and image reconstruction methods are described in the Supplementary Material. Predicted internal opening locations and abscess types were documented for subsequent analysis.

Outcome measures

Primary outcome measures

The accuracy of internal opening localization was defined as the proportion of patients whose fistular internal opening locations as predicted by ultrasonography were consistent with those confirmed by surgical findings, which were taken as the gold standard. The anal fistula recurrence rate at 6 months after surgery was defined as the proportion of patients who required a second surgery due to the appearance of new abscesses or fistula tracts at or near the original surgical site within 6 months after the initial surgery.

Secondary outcome measures

The accuracy of the Parks classification system for abscesses was assessed by the proportion of patients whose abscess type (as determined by the system) was consistent with the type finally confirmed during surgery. Total surgical duration from skin incision to wound closure was recorded in minutes. Intraoperative blood loss was assessed by measuring the volume of fluid in the suction container and calculating the weight differences of used surgical gauze. Patients' pain scores on postoperative day 3 were assessed using the visual analog scale, which ranges from 0 to 10 points, with higher scores indicating more severe pain. Wound healing time was defined as the number of days from surgery to complete wound epithelialization without exudate.

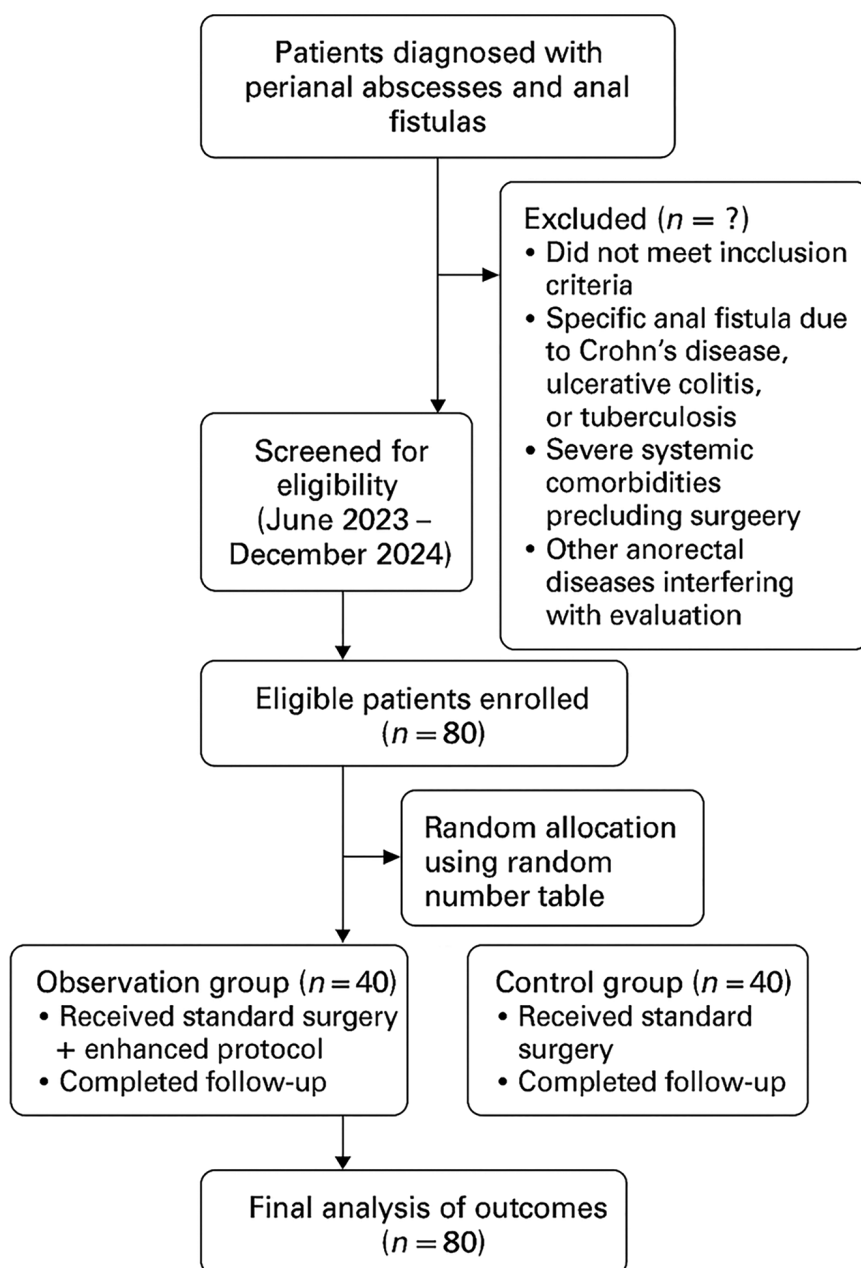


Figure 1. Flow diagram of patient enrollment and study procedures.

Gold standard of surgical findings

Intraoperative findings were used as the gold standard for diagnosis. All surgical procedures were performed independently by two senior anorectal surgeons, each with over 10 years of clinical experience. Both surgeons were blinded not only to group allocation but also to all preoperative ultrasonographic findings, including the predicted locations of internal openings, to prevent incorporation or verification bias in outcome assessments. The intraoperative exploration and identification of internal openings were performed solely based on direct surgical findings. Following the lumbo-sacral or sacral canal block anesthetization of patients, an incision was made, as per standardized procedures, to expose the abscess cavity; pus was thoroughly drained; and a probe was used to gently explore the abscess cavity and any suspicious paths, with the purpose of tracing the course of fistula tracts. For cases with a suspected internal opening, diluted methylene blue or hydrogen peroxide was injected to aid visual identification. Intraoperative records included the specific anatomical location and extent of abscesses, the course and branching of fistula tracts, Parks classification results, and the number and exact locations of internal openings (annotated using the 12-o'clock method, with the height relative to the dentate line specified). To eliminate the effect of postoperative management differences on study results, completely identical standardized protocols for nursing and medication were implemented postoperatively for all participants. After surgery, a drainage gauze strip was routinely placed and removed for 24–72 h, based on the condition of any exudate. Additionally, broad-spectrum antibiotics (covering aerobic and anaerobic bacteria) were administered intravenously for 3–5 days immediately after surgery. Pain was managed with non-steroidal anti-inflammatory drugs.

Follow-up plan

All patients were followed up at 1 and 6 months postoperatively. The follow-up at month 1 was mainly completed in the outpatient clinic and focused on the assessment of complete wound epithelialization, short-term complications (incision infection, exudation, bleeding, etc.), and signs of recurrence. As the 6-month follow-up was the time point for evaluating the primary outcome, a systematic physical examination was performed in the outpatient clinic. For patients unable to attend follow-ups through

clinic visits due to geographical or other objective reasons, assessment was completed by a trained follow-up physician through telephone interviews and supplemented by photos when necessary.

Statistical analysis

Statistical analysis was completed using SPSS software (version 26.0, IBM Corp., Armonk, NY, USA). The normality of continuous variables was assessed using the Shapiro–Wilk test. Normally distributed continuous variables were expressed as mean $\pm$ standard deviation and compared between groups using the independent sample t-test. Categorical variables were expressed as n (%) and compared using the chi-squared test or exact tests, as appropriate. When expected cell counts did not meet the assumptions of the chi-squared test, Fisher's exact test was applied for 2×2 contingency tables, and the Fisher–Freeman–Halton exact test was used for larger contingency tables. All analyses were conducted according to the intention-to-treat principle, and all statistical tests were two-tailed, with $P < 0.05$ denoting a statistically significant difference.

Results

Baseline patient data

There were no statistically significant differences in age, sex ratio, body mass index, duration of symptoms, presence of major clinical symptoms (pain, lumps, and fever), history of perianal abscesses or anal fistulas, diabetes mellitus status, or level of C-reactive protein (an inflammatory indicator) between the two groups ($P > 0.05$), implying comparable baseline characteristics between the two groups (Table 1).

Imaging characteristics

In the two-probe ultrasonography method, superficial abscesses mostly presented as hypoechoic or anechoic zones with relatively clear boundaries, occasionally with liquefied septations. The high-frequency probe accurately demonstrated the extent of each abscess cavity and its relationships to the skin (Figure 2a). For complex or high-grade lesions, 2D end-fire images revealed the anatomical relationship between each abscess cavity and the internal and external sphincters (Figure 2b). Moreover, the 3D end-fire probe enabled the 3D reconstruction of the abscess cavity, visually displaying its space adjacent to the sphincter and levator ani muscle (Figures 2c and 2d). In some typical cases, the fistula tract appeared as a cord-like hypoechoic structure extending from the abscess cavity to the mucosal layer of the anal canal, forming a potential internal opening channel. Atypical cases were characterized by an irregular course of fistula tracts, with branches or concealed internal openings.

Accuracy of internal opening localization

Table 2 displays the comparison between the ultrasonographic predictions for the locations of internal openings and the actual surgical findings in the study cohort. Specifically, 36 cases of true positive, 3 cases of false positive, 37 cases of true negative, and 4 cases of false negative were predicted by ultrasonography. Taking surgical findings as the gold standard, the calculated sensitivity, specificity, positive predictive value, negative predictive value, and overall accuracy of ultrasonography for locating internal openings were 90.0% (36/40), 92.5% (37/40), 92.3% (36/39), 90.2% (37/41), and 91.3% (73/80), respectively, hinting at high diag-

Table 1. Baseline patient data

Characteristic	Observation group (n = 40)	Control group (n = 40)	P
Age (years)	42.6 $\pm$ 11.8	43.4 $\pm$ 12.1	0.765
Gender (M/F)	28/12	27/13	1.000
Body mass index (kg/m ²)	24.1 $\pm$ 3.2	23.8 $\pm$ 3.4	0.686
Duration of symptoms (days)	6.3 $\pm$ 2.4	6.5 $\pm$ 2.8	0.733
Main symptoms			
Pain	33 (82.5%)	34 (85.0%)	1.000
Lumps	25 (62.5%)	27 (67.5%)	0.815
Fever	10 (25.0%)	9 (22.5%)	1.000
History of perianal abscesses/anal fistulas	6 (15.0%)	7 (17.5%)	1.000
Diabetes mellitus morbidity	4 (10.0%)	5 (12.5%)	1.000
Serum C-reactive protein (mg/L)	38.6 $\pm$ 12.4	37.9 $\pm$ 13.1	0.807

M, male; F, female.

nostic efficiency. Additionally, the analysis of agreement between the two independent sonographers yielded a Cohen's kappa value of 0.83 ($P < 0.001$), suggesting favorable inter-observer agreement. The diagnostic performance of ultrasonography in locating internal openings was compared between groups (Table 3), and the observation group

consistently demonstrated higher sensitivity, specificity, and overall accuracy than the control group, though the differences did not reach statistical significance. This trend indicates the potential advantage of combining a high-frequency probe with a 3D end-fire probe for locating the internal openings of fistulas.

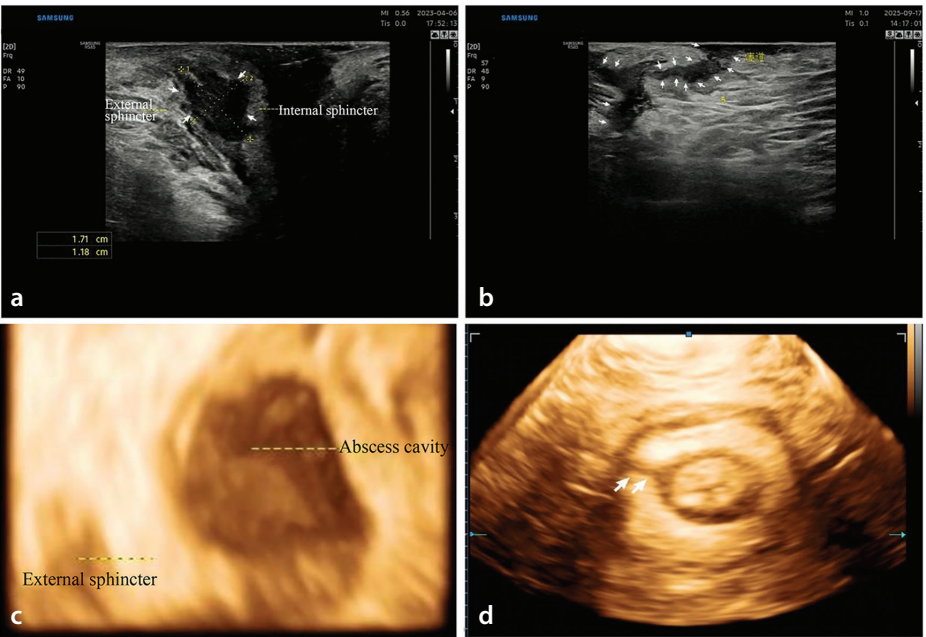


Figure 2. Typical imaging characteristics; (a) Superficial abscesses and subcutaneous fistula tracts displayed using the high-frequency probe; (b) relationship between the abscess cavity and sphincter in two dimensional mode; (c, d) image of three dimensional (3D) reconstruction visually demonstrating the 3D shape of the abscess cavity and the spatial structure of the anal canal. All images are clearly illustrated with arrows and annotations for easy interpretation.

Table 2. Comparison between ultrasonographic diagnosis results and surgical findings (with the latter considered as the gold standard)			
	Presence of internal openings confirmed by surgery	Absence of internal openings confirmed by surgery	Total
Presence of internal openings indicated by ultrasonography	TP = 36	FP = 3	39
Absence of internal openings indicated by ultrasonography	FN = 4	TN = 37	41
Total	40	40	80

TP, true positive; FP, false positive; TN, true negative; FN, false negative.

Table 3. Diagnostic efficacy of ultrasonography in locating inner openings			
Indicator	Observation group (n = 40)	Control group (n = 40)	P
Sensitivity (%)	92.5 (37/40)	75.0 (30/40)	0.066
Specificity (%)	95.0 (38/40)	80.0 (32/40)	0.087
Positive predictive value (%)	94.9	79.0	-
Negative predictive value (%)	92.7	76.2	-
Overall accuracy (%)	93.8	77.5	-

P values were calculated for sensitivity and specificity using a two-sided Fisher's exact test.

Anal fistula recurrence rate at 6 months after surgery

Recurrence was observed in 13 of the 80 study participants at 6 months postoperatively, with an overall recurrence rate of 16.3%, including 3 (7.5%) from the observation group and 10 (25.0%) from the control group. Although the difference between the two groups did not reach statistical significance (Fisher's exact test, $P = 0.066$), the recurrence rate in the observation group was numerically reduced by approximately two-thirds compared with that in the control group. This magnitude of reduction suggests there is potentially a clinically meaningful benefit in combining a high-frequency probe with a 3D end-fire probe to lower the risk of postoperative recurrence. All patients who experienced a recurrence presented with newly formed abscesses or fistula tracts at the original surgical site or adjacent areas and received secondary surgical intervention.

Accuracy of the Parks classification system for abscesses

Compared with classification by surgical findings, the accuracy of classification by ultrasonography was higher in the observation group than in the control group (92.5% vs. 77.5%), though the difference did not reach statistical significance (Fisher's exact test, $P = 0.115$). Notably, a consistent improvement in diagnostic accuracy was observed across multiple fistula subtypes, including intersphincteric, low and high transsphincteric, and ischiorectal (Table 4). This consistent pattern across different anatomical classifications supports the clinical value of combining a high-frequency probe with a 3D end-fire probe, as it may enhance the stability and reproducibility of abscess and fistula tract characterization, particularly in complex anatomical settings.

Other secondary outcome measures

The observation group outperformed the control group in surgical efficiency, intraoperative blood loss, and postoperative recovery outcomes. Specifically, shorter surgical durations, less intraoperative blood loss, milder pain on postoperative day 3, and significantly shortened wound healing times were noted in the observation group, and the differences were statistically significant ($P < 0.05$) (Table 5).

Discussion

Perianal abscesses and anal fistulas are common and frequently encountered dis-

Classification	Number of cases in observation group	Correct cases in observation group (n)	Accuracy in observation group (%)	Number of cases in control group	Correct cases in control group (n)	Accuracy in control group (%)
Intersphincteric type	14	13	92.9	15	11	73.3
Low transsphincteric type	12	11	91.7	11	9	81.8
High transsphincteric type	8	7	87.5	7	5	71.4
Ischiorectal type	5	5	100.0	6	5	83.3
Others (suprasphincteric type, etc.)	1	1	100.0	1	1	100.0
Total	40	37	92.5	40	31	77.5

Measure	Observation group (n = 40)	Control group (n = 40)	P
Surgical duration (minute)	42.3 ± 8.5	49.6 ± 9.1	<0.001
Intraoperative blood loss (mL)	28.4 ± 10.2	36.9 ± 12.7	0.002
VAS pain score on postoperative day 3 (points)	3.2 ± 1.1	4.5 ± 1.3	<0.001
Wound healing time (days)	22.6 ± 5.8	28.3 ± 6.2	<0.001

VAS, visual analog scale.

eases in anorectal surgery, and the cornerstone of their treatment lies in the accurate identification and management of their internal openings, the complete removal of infectious lesions, and the protection, to the greatest extent, of anal sphincter function. Surgical failure and postoperative recurrence are usually ascribed to a failure to accurately identify the locations of internal openings and the course of complex fistulas. In the present study, the combined application of a high-frequency linear array probe and a 3D end-fire probe in the observation group achieved greater accuracy in locating internal openings and Parks classifications, lower postoperative recurrence rates, and greater improvements in surgery-related indicators compared with the control group, demonstrating that a more refined preoperative imaging assessment is capable of effectively guiding surgical decision making and thus improving patient prognoses.

The accurate preoperative assessment of anal fistulas is pivotal for formulating surgical plans.¹⁴ For complex and recurrent anal fistulas, digital examination, probing, and other traditional examination methods achieve low accuracy.¹⁵ In recent years, intraluminal ultrasonography and MRI have become major imaging evaluation methods.⁸ In the present study, the combined examination mode used in the observation group fully leveraged the advantages of two different ultrasound techniques. The high-frequency probe offers excellent superficial tissue resolution for the clear visualization of subcutaneous fistula tracts and superficial abscesses and their re-

lationship to the skin, which is valuable for exploring lesions around external openings. The 3D end-fire probe provides panoramic anatomical images of the anal canal, rectal wall, and surrounding sphincters and, with the help of 3D reconstruction technologies, enables an observation of the shape and size of any abscess cavity or fistula tract and its spatial adjacency to such deep structures as the internal and external sphincters and the levator ani muscle from multiple angles and cross-sections. A systematic review and meta-analysis reports that 3D intraluminal ultrasonography demonstrates high diagnostic accuracy for most anal fistulas and can serve as a first-line diagnostic tool.¹⁶ The combination of shallow-to-deep and surface-to-internal imaging capacity makes the overall judgment of fistula tract course from this method more comprehensive and accurate, thereby enhancing sensitivity and specificity in locating fistular internal openings. Research has shown that intraluminal ultrasonography has a sensitivity of 87.38% and a specificity of 38.46% for diagnosing perianal fistulas¹⁷—markedly inferior to the results in the observation group of the present study. This may be attributable to the technical advantages of the combined examination mode.

Successful anal fistula surgery requires a balance between the radical treatment of lesions and the preservation of anal function. This, in turn, relies heavily on an accurate assessment of the relationship between the fistula tract course and the sphincter—that is, an accurate Parks classification.¹⁴ The Parks classification system serves as the foundation for guiding the selection of surgical approach.

For instance, fistulotomy is adopted for intersphincteric anal fistulas, whereas sphincter-preserving procedures, such as ligation of the intersphincteric fistula tract, are selected for complex high transsphincteric anal fistulas.¹⁸ In the present study, Parks classification accuracy reached 92.5% in the observation group, exceeding that in the control group (77.5%), especially for more complex types such as transsphincteric and ischiorectal anal fistulas. The latest research has shown that 3D pelvic ultrasonography is highly accurate in locating fistula tracts, identifying internal openings, and determining Parks classifications, rendering a reliable basis for safe treatment.¹⁹ This is attributed to the stereoscopic images provided by 3D ultrasonography, which display the course of the fistula tract passing through the sphincter, offering surgeons a clear surgical roadmap.²⁰ Such accurate preoperative assessment translates directly into clinical benefits. The recurrence rate in the observation group of the present study at 6 months after surgery (7.5%) was much lower than that in the control group (25.0%), a result that suggests that accurate preoperative imaging diagnosis is a pivotal factor in decreasing the postoperative recurrence of anal fistulas. 3D intraluminal ultrasonography has been proven to demonstrate higher accuracy than 2D ultrasonography in assessing the height of the main canal of the transsphincteric fistula tract, with both techniques having good agreement with intraoperative findings in identifying main tracts.²¹ With a complete understanding of the internal opening locations, main tract, and branches of a fistula, as well as any secondary abscess cavities, surgeons can achieve more thorough debridement and more appropriate internal opening management during surgery, effectively preventing recurrence resulting from lesion residue.

Furthermore, better secondary outcomes (e.g., surgical duration, intraoperative blood loss, postoperative pain, and wound healing time) were noted in the observation group compared with the control group in the cur-

rent study, further highlighting the value of accurate preoperative planning. As an accurate and non-invasive imaging technique for diagnosis, intraluminal ultrasonography provides precise details of fistula tracts and is thus useful in preoperative management and surgical planning. With surgical plans based on detailed imaging information, surgeons are spared from having to conduct extensive intraoperative probing, reducing the collateral damage to surrounding tissues. A clearer surgical path also shortens surgical durations and reduces intraoperative blood loss,²² resulting in less surgical trauma, milder postoperative pain, and faster wound healing and thus facilitating patient recovery.²³

The present study has some limitations. First, its single-center design and relatively small sample size may affect the generalizability of its results. Second, ultrasonography depends heavily on operator experience and skill—a condition that may pose a challenge to the repeatability of the results herein. Nevertheless, the kappa consistency test demonstrated good inter-observer agreement. The conclusions of this study should be validated with future, multicenter, prospective randomized controlled trials with larger sample sizes.

In conclusion, transperineal ultrasonography that utilizes both a high-frequency linear array probe and a 3D end-fire probe enhances the accuracy of the preoperative diagnosis of perianal abscesses and anal fistulas, especially with respect to internal opening localization and Parks classification. This improved preoperative evaluation method facilitates more precise surgical planning and may contribute to a reduced risk of postoperative recurrence. Moreover, the combined ultrasonographic approach significantly shortens surgical duration, reduces intraoperative blood loss, alleviates postoperative pain, and accelerates wound healing, thereby promoting postoperative recovery and overall surgical outcomes.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Material: <https://d2v96fxpocvxx.cloudfront.net/bda9171a-fae8-4995-8276-2138323f1e16/content-images/05c6b72e-07bd-4811-8be0-a2e13fc7c9d1.pdf>

References

- Yagnik VD, Kaur B, Dawka S, Sohal A, Menon GR, Garg P. Non-locatable internal opening in anal fistula associated with acute abscess and its definitive management by Garg protocol. *Clin Exp Gastroenterol.* 2022;15:189-198. [\[Crossref\]](#)
- Dong S, Chen B, Zhang J. Study on the factors influencing the prognosis after perianal abscess surgery. *BMC Gastroenterol.* 2023;23(1):334. [\[Crossref\]](#)
- García-Botello S, Martín-Arévalo J, Martí-Fernández R, Moro-Valdezate D, Pla-Martí V, Espi-Macías A. Three-dimensional endoanal ultrasound to assess the validity of Goodsall's Rule and the Midline Rule in predicting the path of perianal fistula-in-ano and the location of the internal opening. *Tech Coloproctol.* 2022;26(5):351-361. [\[Crossref\]](#)
- Arkenbosch JHC, van Ruler O, de Vries AC, van der Woude CJ, Dwarkasing RS. The role of MRI in perianal fistulizing disease: diagnostic imaging and classification systems to monitor disease activity. *Abdom Radiol (NY).* 2025;50(2):589-597. [\[Crossref\]](#)
- de la Portilla F, Sojo V, Vázquez-Monchul JM, Pintor-Tortolero J, Dios S, Reyes-Díaz ML. Description of a new ultrasound sign to distinguish Crohn's anal fistula from cryptoglandular fistula: the rosary sign. *Colorectal Dis.* 2023;25(7):1446-1452. [\[Crossref\]](#)
- de la Portilla de Juan F, García-León A, García-Sánchez CJ, et al. Comparative study of diluted hydrogen peroxide and sulfur hexafluoride in the contrast-enhanced ultrasound assessment of anal fistulas. *Dis Colon Rectum.* 2024;67(11):1450-1457. [\[Crossref\]](#)
- Lavazza A, Maconi G. Transperineal ultrasound for assessment of fistulas and abscesses: a pictorial essay. *J Ultrasound.* 2019;22(2):241-249. [\[Crossref\]](#)
- Santoro GA, Fortling B. The advantages of volume rendering in three-dimensional endosonography of the anorectum. *Dis Colon Rectum.* 2007;50(3):359-368. [\[Crossref\]](#)
- Yang D, Yao X. Clinical value of transperineal 3D volume ultrasound combined with 2D high frequency ultrasound in anal fistula. *J Biosci Med.* 2020;8(5):12-19. [\[Crossref\]](#)
- Maconi G, Greco MT, Asthana AK. Transperineal ultrasound for perianal fistulas and abscesses - a systematic review and meta-analysis. *Ultraschall Med.* 2017;38(3):265-272. English. [\[Crossref\]](#)
- Ren J, Bai W, Gu L, Li X, Peng X, Li W. Three-dimensional pelvic ultrasound is a practical tool for the assessment of anal fistula. *BMC Gastroenterol.* 2023;23(1):134. [\[Crossref\]](#)
- Xiong X, Jiang M, He N, et al. Comparative clinical evaluation of combined three-dimensional anorectal ultrasound and high-frequency linear array ultrasound versus MRI for diagnosis and postoperative outcomes of anal fistula. *Am J Transl Res.* 2025;17(7):5054-5062. [\[Crossref\]](#)
- Varsamis N, Kosmidis C, Chatzimavroudis G, et al. Preoperative assessment of perianal fistulas with combined magnetic resonance and tridimensional endoanal ultrasound: a prospective study. *Diagnostics (Basel).* 2023;13(17):2851. [\[Crossref\]](#)
- Akhoundi N, Bozchelouei JK, Abrishami A, et al. Comparison of MRI and endoanal ultrasound in assessing intersphincteric, transsphincteric, and suprasphincteric perianal fistula. *J Ultrasound Med.* 2023;42(9):2057-2064. [\[Crossref\]](#)
- Teymour A, Keshvari A, Ashjaei A, et al. Predictors of outcome in cryptoglandular anal fistula according to magnetic resonance imaging: a systematic review. *Health Sci Rep.* 2023;6(6):e1354. [\[Crossref\]](#)
- Li J, Chen SN, Lin YY, et al. Diagnostic accuracy of three-dimensional endoanal ultrasound for anal fistula: a systematic review and meta-analysis. *Turk J Gastroenterol.* 2021;32(11):913-921. [\[Crossref\]](#)
- Teymour A, Keshvari A, Khorasanizadeh F, et al. External validation of the modified Parks classification of cryptoglandular anal fistula to predict failure of healing: the protocol for a retrospective analysis. *Int J Surg Protoc.* 2025;29(3):118-121. [\[Crossref\]](#)
- Elshamy MT, Emile SH, Abdelnaby M, Khafagy W, Elbaz SA. A pilot randomized controlled trial on ligation of intersphincteric fistula tract (LIFT) versus modified parks technique and two-stage seton in treatment of complex anal fistula. *Updates Surg.* 2022;74(2):657-666. [\[Crossref\]](#)
- Gou BO, Zhang JC, Chen L, et al. Comparison of the diagnostic accuracy of percutaneous fistula contrast-enhanced ultrasound combined with transrectal 360° 3-D imaging and conventional transrectal ultrasound in complex anal fistula. *Ultrasound Med Biol.* 2022;48(10):2154-2161. [\[Crossref\]](#)
- Yang J, Li Q, Li H, et al. Preoperative assessment of fistula-in-ano using SonoVue enhancement during three-dimensional transperineal ultrasound. *Gastroenterol Rep (Oxf).* 2024;12:goae002. [\[Crossref\]](#)
- Garcés-Albir M, García-Botello SA, Espi A, et al. Three-dimensional endoanal ultrasound for diagnosis of perianal fistulas: reliable and objective technique. *World J Gastrointest Surg.* 2016;8(7):513-520. [\[Crossref\]](#)
- Ding YW, Yin HQ, Liang HT, Lu JG, Wang B, Wang C. Can transcutaneous perianal ultrasonography be the first-line diagnostic instrument for evaluating pediatric perianal fistulas? *Gastroenterol Rep (Oxf).* 2022;10:goac071. [\[Crossref\]](#)
- Deng H, Li M, Fang X, et al. Evaluation of the mechanical properties and clinical application of nickel-titanium shape memory alloy anal fistula clip. *Front Surg.* 2023;10:1235666. [\[Crossref\]](#)



Microvascular dysfunction in patients with transthyretin cardiac amyloidosis evaluated by 13N-ammonia positron emission tomography-computed tomography: is it an early marker of the disease?

Aristóteles Neto¹
 Caio Cafezeiro¹
 Bruno Bueno¹
 Cristhian Espinoza Romero¹
 Vitor Emer Rosa¹
 Viviane Hotta¹
 Camila de Godoi Carneiro²
 Carlos Rochitte¹
 Vagner Madrini¹
 Jose Soares¹
 William Chalela¹
 Marcos Lima²
 Carlos Buchpiguel²
 Edward Miller³
 Fabio Fernandes¹

¹Heart Institute (InCor), Hospital das Clínicas HCFMUSP, Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil

²Department of Radiology, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP), São Paulo, Brazil

³Medicine and Radiology in the Section of Cardiovascular Medicine in Yale New Haven Hospital, Connecticut, United States

PURPOSE

To evaluate coronary microvascular function using 13N-ammonia positron emission tomography/computed tomography in individuals with pathogenic *transthyretin* (*TTR*) gene mutations, with and without cardiac involvement. This study is the first to assess coronary flow reserve (CFR) in this population before overt cardiac amyloidosis (CA) is detectable by conventional imaging.

METHODS

We evaluated microvascular impairment by measuring CFR in 20 patients with and 20 patients without cardiac involvement due to TTR amyloidosis (ATTR), all presumed to be free from epicardial coronary artery disease and carrying *TTR* gene mutations.

RESULTS

The study revealed a significantly reduced mean global CFR in the cardiac involvement group (1.849 ± 0.379 vs. 2.952 ± 0.7 , $P < 0.001$). Global CFR inversely correlated with age, functional class, troponin, and B-type natriuretic peptide while positively correlating with the 6-minute walk test distance, mean blood pressure, and global longitudinal strain. Receiver operating characteristic curve analysis identified an optimal cutoff value of global CFR < 2.58 , yielding a sensitivity of 100% and a specificity of 75% for detecting cardiac involvement.

CONCLUSION

In patients with ATTR CA, coronary microvascular dysfunction emerges as a clinically relevant marker of cardiac involvement, even in the absence of structural abnormalities or obstructive coronary disease.

CLINICAL SIGNIFICANCE

CFR assessment may aid in diagnostic suspicion, risk stratification, and understanding of angina symptoms in this population.

KEYWORDS

Amyloidosis, microvascular blood flow, coronary arteries

Handling editor: Mehmet Ruhi Onur

Corresponding author: Cristhian Espinoza Romero

E-mail: cristhian.153@hotmail.com

Received 08 October 2025; revision requested 22 December 2025; accepted 14 January 2026.



Epub: 04.02.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263687

Transthyretin (TTR) amyloidosis (ATTR) affects the heart through the deposition of amyloid fibrils. Early biomarkers, such as troponins, often indicate damage to cardiomyocytes and the coronary microvasculature.^{1,2} ATTR occurs either in a hereditary form, caused by *TTR* gene mutations, or in a wild-type form, typically associated with aging. In both entities, positron emission tomography/computed tomography (PET/CT) has emerged as a valuable tool for assessing microvascular dysfunction, enabling early evaluation of myocardial perfusion and coronary flow reserve (CFR) impairment in this population.²⁻⁵ Cardiac amyloidosis (CA) compromises the microvasculature via fibril-induced thickening, extrinsic vessel

compression, or endothelial and autonomic dysfunction.¹ Patients with CA frequently exhibit chronic troponin elevation, electrically inactive regions on the electrocardiogram, and anginal symptoms, and signs of ischemia have been reported in some individuals with CA without obstructive epicardial coronary artery disease (CAD).²⁻⁵

Given the scarcity of studies evaluating coronary microvascular dysfunction in ATTR, this is the first known study to employ ¹³N-ammonia PET/CT to assess CFR in individuals exclusively carrying TTR mutations without evident cardiac involvement on standard imaging examinations. However, a previous study evaluated 31 patients with CA, including 15 with light-chain amyloidosis (AL) and 6 with ATTR, and compared them with patients with left ventricular hypertrophy secondary to hypertension. In that study, microvascular dysfunction was more prevalent in patients with CA-ATTR.⁶⁻⁹ Another study evaluated 400 patients with CA-ATTR; however, microvascular obstruction, assessed using magnetic resonance imaging, was present in 221 (27.6%) patients and was more common in CA-ATTR than in CA with AL.⁵

The primary aim of this research is to assess microvascular function in individuals with pathogenic TTR gene mutations, both with and without cardiac involvement. Using ¹³N-ammonia PET/CT, we evaluated CFR and its correlation with clinical, laboratory, and imaging variables.

Methods

Study design and population

A total of 40 individuals carrying pathogenic TTR gene mutations were prospectively enrolled (Supplementary Figure 1). Participants were divided into two groups:

- Group A (n = 20): patients with evidence of cardiac involvement
- Group B (n = 20): patients without cardiac involvement.

Main points

- Patients with cardiac involvement exhibit significantly reduced coronary flow reserve (CFR).
- Coronary microvascular dysfunction emerges as a clinically relevant marker of cardiac involvement.
- CFR assessment may aid in diagnostic suspicion, risk stratification, and understanding of angina symptoms in this population.

Additionally, as part of the initial protocol focused on cardiac elastography, another group of 20 patients was included as healthy controls; CFR was not assessed in these individuals. First-degree relatives of patients with hereditary ATTR who elected to undergo genetic testing after genetic counseling were included. Those testing positive for an ATTR-compatible mutation were referred from external services to the cardiomyopathy specialist institutions participating in the study.

Inclusion criteria

- Age ≥18 years
- Confirmed pathogenic mutation in the TTR gene
- Ability to provide informed consent

Cardiac involvement, defined by the presence of at least one of the following:

- Interventricular septal thickness > 12 mm on echocardiography or cardiac magnetic resonance imaging, in the absence of other causes of hypertrophy (e.g., hypertension and aortic stenosis)
- History of heart failure symptoms or signs consistent with CA
- Elevated cardiac biomarkers [troponin or B-type natriuretic peptide (BNP)] without an alternative explanation
- Positive Congo red staining in tissue biopsy or Grade 2 or 3 cardiac uptake on ^{99m}Tc-pyrophosphate scintigraphy, with monoclonal gammopathy of undetermined significance ruled out

The exclusion criteria are provided in the Supplementary appendix. Wild-type TTR CA was excluded.

PET/CT imaging protocol: Coronary microvascular blood flow (MBF) was assessed using dynamic ¹³N-ammonia PET/CT imaging at rest and during pharmacologic stress induced by intravenous dipyridamole. MBF quantification was performed using time-activity curves derived from the blood pool and myocardium, corrected for partial volume effects and spill-over.

CFR calculation: This was calculated as the ratio of stress MBF to rest MBF, according to the 17-segment model proposed by the American Society of Nuclear Cardiology.^{10,11} Rest MBF values were normalized using the rate-pressure product to account for individual variations in cardiac workload.

All 40 patients included in the study underwent calcium score assessment as a screening measure to minimize bias related to CAD, which was low in both groups (Supplementary Table 1). The cut-off values used were 40 ng/L for troponin and 35 pg/mL for BNP. This study was approved by the Ethics Committee of InCor (CAAE 27437019.5.0000.0068) through a substantiated opinion on February 6, 2020. All participants provided written informed consent prior to inclusion, under the supervision and approval of the institutional ethics committee.

Statistical analysis

The sample size was determined pragmatically based on patient availability. Continuous variables are reported as mean ± standard deviation when normally distributed and as median (interquartile range) or median ± median absolute deviation when distributional assumptions were violated. The t-test was used to compare means between two groups when both normality (Shapiro-Wilk) and homoscedasticity (Levene) assumptions were met. When either assumption was violated, the Mann-Whitney U test was applied. The null hypothesis (H₀) for all comparisons stated that no significant difference existed between groups. Where H₀ was rejected, effect size was calculated using Cohen's d (d: 0.2–0.5 = small, 0.5–0.8 = moderate, > 0.8 = large).

Pearson's correlation coefficient (r) was used to quantify linear relationships between parametric variables. Spearman's rank correlation coefficient (ρ) was applied for non-parametric variables. Both coefficients range from –1 to +1, with values approaching either extreme indicating stronger associations. For performance analysis, receiver operating characteristic curve analysis and the Youden index were used to establish the optimal cutoff point. A multivariable regression model was constructed, including covariates such as age, blood pressure, sex, a non-zero coronary artery calcium score, and biomarkers.

All statistical tests were two-tailed, and P < 0.05 was considered statistically significant. Analyses were performed using JASP software, version 0.16.3 JASP software, version 0.16.3 (JASP Team, Amsterdam, The Netherlands).¹²

Results

Baseline characteristics

The main clinical and structural characteristics of the study population are summa-

ized in Table 1. The remaining characteristics are summarized in Supplementary Tables 1, 2, 3 and 4.

Differences between the groups were evident. Patients with cardiac involvement exhibited conditions consistent with longer exposure to amyloid deposition: they were older and had a higher body mass index, greater interventricular septal thickness, atrial enlargement, lower left ventricular ejection fraction, lower left ventricular global longitudinal strain (LVGLS) and right ventricular global longitudinal strain (RVGLS), more advanced degrees of diastolic dysfunction, poorer functional capacity, and worse performance on the 6-minute walk test, with a shorter total distance. In addition, they had higher cardiac biomarker levels, including troponin and BNP. This group also showed a higher prevalence of chronic conditions such as hypertension and diabetes. Despite the diagnosis of hypertension, patients reported not using antihypertensive medications and had normal or reduced blood pressure levels at the time of consultation. No patient with type 2 diabetes mellitus required insulin for glycemic control. There was no history of anginal symptoms, coronary atherosclerotic disease, or use of anti-ischemic medications or prior revascularization in either group. In the groups with pathological variants of the *TTR* gene, the most prevalent mutations were Val142Ile, followed by Val50Met and Thr80Ala, with no statistically significant differences in mutation type between the groups.

Global coronary flow reserve

Mean global CFR was significantly lower in the cardiac involvement group (1.849 ± 0.379) than in the non-involvement group (2.952 ± 0.7; *P* < 0.001) (Figure 1, Table 2). Global CFR showed an inverse correlation with age, functional class according to the New York Heart Association classification, troponin, and BNP levels, and it showed a positive correlation with total distance in the 6-minute walk test and mean arterial pressure. Accordingly, older age, higher functional class according to the NYHA classification, and elevated serum troponin and BNP levels were associated with worse global CFR. Conversely, better performance on the walk test was associated with higher global CFR (Figure 2, Table 3, Supplementary Table 5).

CFR also correlated with multiple clinical, laboratory, and echocardiographic parameters, including markers of left and right ventricular systolic and diastolic function, as well as indi-

ces of structural remodeling in both ventricles and atria (Table 3). Notably, global CFR demonstrated significant associations with predictors of subclinical myocardial involvement, such as LVGLS and RVGLS (Figure 2).

Segmental CFR analysis revealed lower values in the CA group, particularly in apical segments. This reduction was evident both when segments were grouped according to epicardial artery territories and when analyzed individually (Table 4).

Receiver operating characteristic curve analyses

Receiver operating characteristic curve analysis identified a global CFR cut-off of < 2.58 to differentiate cardiac involvement, with a sensitivity of 100% and specificity of 75%, an area under the curve of 0.902 (0.80–1.0), and a *P* value of <0.001 (Supplementary Figure 2). In Supplementary Figure 3 we can see the genetic profile of the patients included in the study.

Coronary flow reserve and ischemic disease equivalent

In the cardiac involvement group, 60% (*n* = 12) demonstrated a global CFR consistent with ischemic disease (≤ 2.5), whereas only

15% (*n* = 3) of the non-involvement group exhibited similar findings.

Discussion

The main findings of this study support the inclusion of anginal symptoms in the absence of obstructive epicardial CAD as a relevant clinical feature when suspecting CA in patients with *TTR* mutations. Additionally, our results emphasize the potential role of coronary microvascular dysfunction as a pathophysiological marker of the disease, revealing a global perfusion impairment pattern that resembles three-vessel CAD, including the apical region.

This study further advances the concept of coronary microvascular dysfunction as a hallmark of CA, a prototypical extracellular storage disorder characterized by insoluble fibril deposition, a hypertrophic phenotype, and extensive structural remodeling.²⁻⁵

The coexistence of microvascular dysfunction and structural abnormalities in the setting of restrictive cardiomyopathy underscores the need for comprehensive functional assessments.⁴⁻⁷ Our findings highlight a potential dissociation between structural and functional impairment, as evidenced by patients with reduced CFR but lacking con-

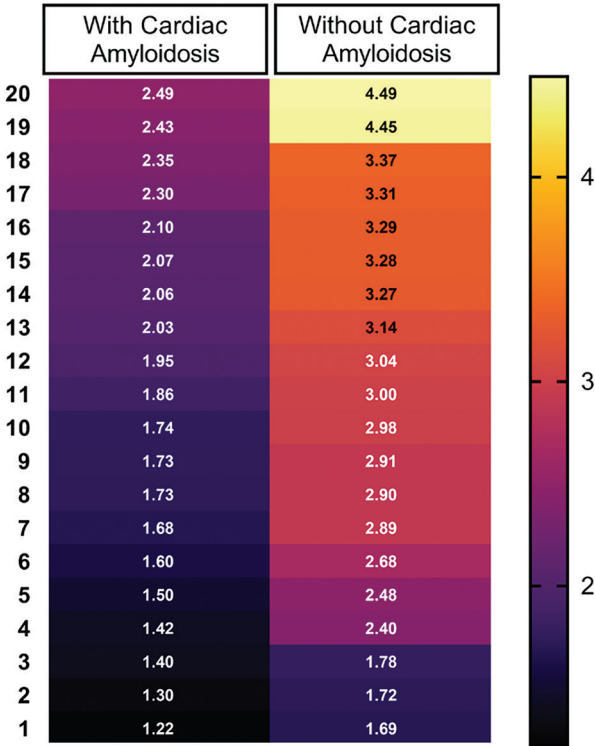


Figure 1. Heat map: global coronary flow reserve (CFR). Heat map representing global CFR distribution across study groups: each column corresponds to a study group, and each row represents an individual patient. Color intensity reflects CFR values, with lighter shades indicating higher (better) CFR and darker shades indicating lower (worse) CFR. This visualization highlights differences in microvascular function between groups.

ventional markers of cardiac involvement. Among patients with established cardiac involvement, 60% exhibited a global CFR pattern consistent with ischemic heart disease, whereas only 15% of those without cardiac involvement showed similar findings. These

results are consistent with previously reported CFR thresholds for epicardial CAD.¹³ In contrast to prior studies, our data demonstrate that CFR impairment may occur independently of anginal symptoms.^{1,14} Notably, 15% of patients without overt car-

diac involvement based on standard diagnostic criteria had CFR values indicative of ischemic disease, suggesting the presence of subclinical microvascular dysfunction.

A novel and clinically relevant observation was the absence of apical microvascular preservation in the cardiac involvement group, despite preserved apical longitudinal strain on speckle-tracking echocardiography. This is the first study to report such a discrepancy, highlighting a unique pattern of microvascular involvement that can be effectively detected using 13N-ammonia PET/CT imaging.

Koyama and Falk¹⁵ demonstrated that reduced longitudinal deformation is associated with poorer survival in patients with AL. However, our findings, derived exclusively from a population with CA-ATTR, suggest that coronary microvascular dysfunction driven by a higher amyloid burden may represent the mechanistic link underlying impaired longitudinal deformation.

Importantly, early-stage amyloidosis without overt myocardial infiltration may still present with anginal symptoms due to microvascular obstruction, which has been associated with worse prognosis in this condition.¹⁶ A prevalence of up to 3% has been reported among patients with amyloidosis and angiographically normal coronary arteries.³ Although the exact frequency of amyloid-related angina remains unclear, the presence of chest pain in patients with suspected amyloidosis warrants careful evaluation.

Finally, our study validated a CFR cut-off value of 2.58, which demonstrated high sensitivity and specificity for detecting cardiac involvement. These findings suggest that early functional assessments, such as CFR evaluation, may complement structural imaging modalities in identifying early cardiac amyloid infiltration.

This study is limited by its small sample size and the lack of invasive coronary artery assessment due to ethical and safety considerations. However, the use of calcium scoring and the diffuse CFR reduction pattern provided reliable insights into microvascular function. Further studies are warranted to validate these findings in larger cohorts. Another limitation is the small and uneven sample, with a significant age disparity between groups (patients without cardiac involvement in hereditary ATTR are typically identified at a young age). Furthermore, the study sample was selected using conveni-

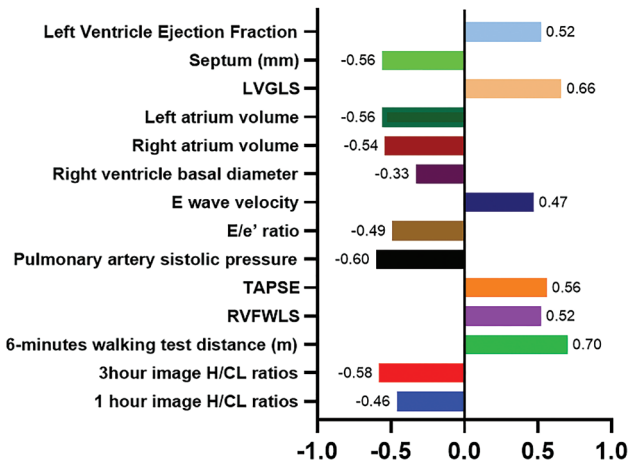


Figure 2. Didactic representation of significant correlations of global coronary flow reserve. Bars represent the strength and direction of correlations, with positive and negative associations indicated accordingly. Numerical values next to each bar denote the correlation coefficients, reflecting the magnitude of the relationships between variables. LVGLS, left ventricular global longitudinal strain; TAPSE, tricuspid annular plane systolic excursion; RVFWLS, right ventricular free wall longitudinal strain; H/CL, heart-to-contralateral ratio.

Table 1. Baseline clinical and laboratory characteristics of patients			
Variable	ATTR with CA (n = 20)	ATTR without CA (n = 20)	P value
Demographics			
Age (years)	69.6 ± 8.2	41.6 ± 10.8	< 0.001
Male sex	16 (80.0)	11 (55.0)	0.177
Arterial hypertension	12 (60.0)	5 (25.0)	< 0.001
Diabetes mellitus	9 (45.0)	3 (15.0)	< 0.001
BMI (kg/m ²)	23.85 ± 3.0	25.8 ± 4.7	0.146
Hemodynamics			
Heart rate (bpm)	74.4 ± 10.5	69.9 ± 9.5	0.17
SBP (mmHg)	117.9 ± 21.3	128.0 ± 17.9	0.12
MBP (mmHg)	89.5 ± 14.4	94.3 ± 25.1	0.45
Clinical features			
Carpal tunnel syndrome (%)	13 (65)	3 (15)	0.005
Extracardiac involvement (%)	14 (70)	7 (35)	0.027
6-minute walk test distance (m)	336.9 ± 83.7	420.4 ± 112.8	< 0.001
Chest pain (%)	4 (20)	2 (10)	0.66
Functional class (NYHA)			
Class I (%)	4 (20)	20 (100)	-
Class II (%)	14 (70)	0 (0)	-
Class III (%)	2 (10)	0 (0)	-
Biomarkers			
Troponin (ng/L)	63.7 ± 41.2	7.15 ± 10.7	< 0.001
BNP (pg/mL)	389.5 (210–545)	14 (12.5–24)	< 0.001

ATTR, transthyretin amyloidosis; CA, cardiac amyloidosis; BMI, body mass index; bpm, beats per minute; SBP, systolic blood pressure; MBP, mean blood pressure; NYHA, New York Heart Association classification; BNP, B-type natriuretic peptide.

Table 2. Coronary flow reserve in patients with and without cardiac involvement

Coronary territory	ATTR with cardiac involvement (n = 20)	ATTR without cardiac involvement (n = 19)	P value	Cohen's d
Global CFR	1.849 ± 0.379	2.952 ± 0.700	<0.001	1.878
Left anterior descending	1.886 ± 0.364	2.970 ± 0.787	<0.001	1.767
Left circumflex	1.911 ± 0.437	3.087 ± 0.873	<0.001	1.704
Right coronary artery	1.886 (1.63–2.13)	2.900 ± 0.804	<0.001	–0.600

Values are reported as mean ± standard deviation, n/N (%), or median (25th–75th percentile), as appropriate. ATTR, transthyretin amyloidosis; CFR, coronary flow reserve.

Table 3. Global CFR clinical correlations

Correlation	r	P value
Age	–0.701	< 0.001
Weight (kg)	0.174	0.309
Height (m)	0.309	0.067
BMI	0.049	0.785
Heart rate (bpm)	–0.285	0.079
SBP (mmHg)	0.149	0.367
MBP (mmHg)	0.344	0.030
Functional class (NYHA)	–0.553	< 0.001
6-minute walk test distance (m)	0.696	< 0.001
Troponin (ng/L)	–0.581	< 0.001
BNP (pg/mL)	–0.734	< 0.001

BMI, body mass index; SBP, systolic blood pressure; MBP, mean blood pressure; bpm, beats per minute; BNP, B-type natriuretic peptide; NYHA, New York Heart Association classification.

Table 4. Comparison of CFR across the 17 cardiac segments in patients with and without cardiac amyloidosis

		ATTR with AC (n = 20)	ATTR without AC (n = 19)	P value	Cohen's d
PET/CT 13N-ammonia					
Basal	Anterior	2.202 ± 0.58	2.802 ± 0.844	0.006	0.843
	Anterolateral	2.213 ± 0.831	3.031 ± 0.981	0.004	0.899
	Inferolateral	1.91 ± 0.501	3.047 ± 0.933	< 0.001	1.518
	Inferior	2.231 ± 1.737	3.133 ± 0.975	0.025	0.641
	Inferoseptal	2.317 ± 1.6	2.714 ± 0.785	< 0.001	0.57
	Anteroseptal	1.959 ± 0.69	2.844 ± 0.696	< 0.001	1.273
Medium	Anterior	1.956 ± 0.409	2.984 ± 0.909	< 0.001	0.730
	Anterolateral	2.071 ± 0.708	3.113 ± 0.962	< 0.001	1.2
	Inferolateral	1.837 ± 0.390	3.26 ± 0.946	< 0.001	1.99
	Inferior	1.929 ± 0.719	3.155 ± 0.95	< 0.001	1.456
	Inferoseptal	2.172 ± 1.652	3.091 ± 0.881	< 0.001	0.755
	Anteroseptal	1.862 ± 0.4	3.061 ± 0.768	< 0.001	1.953
Apical	Anterior	2.0 ± 0.698	2.885 ± 0.71	< 0.001	0.72
	Lateral	2.313 ± 1.89	3.01 ± 0.708	< 0.001	0.765
	Inferior	2.067 ± 1.699	3.012 ± 0.755	< 0.001	0.66
	Septal	1.965 ± 0.69	2.964 ± 0.824	< 0.001	0.7
	Apex	1.603 ± 0.411	3.029 ± 0.741	< 0.001	0.89

ATTR, transthyretin amyloidosis; CA, cardiac amyloidosis; PET/CT, positron emission tomography/computed tomography.

ence sampling, which has the potential to introduce bias. An additional limitation was the use of dipyridamole as the stress agent. Agents such as adenosine or regadenoson have a faster onset of action and more predictable kinetics than dipyridamole; these properties may be more advantageous for the evaluation of absolute myocardial blood flow using PET perfusion tracers. However, dipyridamole is the most widely used, validated, and familiar agent at our institution and, when administered using an appropriate protocol as in the present study, minimizes the probability of errors.

In conclusion, coronary microvascular dysfunction, assessed by 13N-ammonia PET/CT, is a hallmark of CA, with a global involvement pattern mimicking three-vessel CAD. In patients with CA-ATTR, coronary microvascular dysfunction emerges as a clinically relevant marker of cardiac involvement, even in the absence of structural abnormalities or obstructive coronary disease. CFR assessment may aid in diagnostic suspicion, risk stratification, and the understanding of anginal symptoms in this population.

Footnotes

Conflict of interest disclosure

This work was supported financially by Pfizer sponsorship. No other potential conflict of interest relevant to this article was reported.

Supplementary Figures link: <https://d2v96fx-pocvxx.cloudfront.net/cf9d60d6-523c-458a-a2e6-78728d3ffbb0/content-images/2ad30fbc-8742-4fc5-9a13-95775e744b1d.pdf>

Supplementary Tables link: <https://d2v96fx-pocvxx.cloudfront.net/8a9ff4da-541a-42fa-9980-1a9a3ab6d6c5/content-images/5deef236-fd57-4004-a700-8b7b49126a21.pdf>

References

1. Ruberg FL, Grogan M, Hanna M, Kelly JW, Maurer MS. Transthyretin amyloid cardiomyopathy: JACC State-of-the-Art Review. *J Am Coll Cardiol*. 2019;73(22):2872–2891. [\[Crossref\]](#)
2. Ogawa H, Mizuno Y, Ohkawara S, et al. Cardiac amyloidosis presenting as microvascular angina—a case report. *Angiology*. 2001;52(4):273–278. [\[Crossref\]](#)
3. Al Suwaidi J, Velianou JL, Gertz MA, et al. Systemic amyloidosis presenting with angina pectoris. *Ann Intern Med*. 1999;131(11):838–841. [\[Crossref\]](#)
4. Whitaker DC, Tungekar MF, Dussek JE. Angina with a normal coronary angiogram caused by amyloidosis. *Heart*. 2004;90(9):e54. [\[Crossref\]](#)

5. Netti L, Ioannou A, Martinez-Naharro A, et al. Microvascular obstruction in cardiac amyloidosis. *Eur J Heart Fail.* 2025;27(12):2948-2951. [\[Crossref\]](#)
6. Dorbala S, Vangala D, Bruyere J Jr, et al. Coronary microvascular dysfunction is related to abnormalities in myocardial structure and function in cardiac amyloidosis. *JACC Heart Fail.* 2014;2(4):358-367. [\[Crossref\]](#)
7. Rahman JE, Helou EF, Gelzer-Bell R, et al. Noninvasive diagnosis of biopsy-proven cardiac amyloidosis. *J Am Coll Cardiol.* 2004;43(3):410-415. [\[Crossref\]](#)
8. Falk RH, Alexander KM, Liao R, Dorbala S. AL (Light-Chain) Cardiac Amyloidosis: a review of diagnosis and therapy. *J Am Coll Cardiol.* 2016;68(12):1323-1341. [\[Crossref\]](#)
9. Gertz MA, Benson MD, Dyck PJ, et al. Diagnosis, prognosis, and therapy of transthyretin amyloidosis. *J Am Coll Cardiol.* 2015;66(21):2451-2466. [\[Crossref\]](#)
10. Cerqueira MD, Weissman NJ, Dilsizian V, et al. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart. A statement for healthcare professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association. *Circulation.* 2002;105(4):539-542. [\[Crossref\]](#)
11. Dilsizian V, Bacharach SL, Beanlands RS, et al. ASNC imaging guidelines/SNMMI procedure standard for positron emission tomography (PET) nuclear cardiology procedures. *J Nucl Cardiol.* 2016;23(5):1187-1226. [\[Crossref\]](#)
12. Goss-Sampson M. Statistical Analysis in JASP - A Students Guide v0.10.2. [Internet]. figshare; 2019. [\[Crossref\]](#)
13. Herzog BA, Husmann L, Valenta I, et al. Long-term prognostic value of 13N-ammonia myocardial perfusion positron emission tomography added value of coronary flow reserve. *J Am Coll Cardiol.* 2009;54(2):150-156. [\[Crossref\]](#)
14. Fiechter M, Ghadri JR, Gebhard C, et al. Diagnostic value of 13N-ammonia myocardial perfusion PET: added value of myocardial flow reserve. *J Nucl Med.* 2012;53(8):1230-1234. [\[Crossref\]](#)
15. Koyama J, Falk RH. Prognostic significance of strain Doppler imaging in light-chain amyloidosis. *JACC Cardiovasc Imaging.* 2010;3(4):333-342. [\[Crossref\]](#)
16. Mueller PS, Edwards WD, Gertz MA. Symptomatic ischemic heart disease resulting from obstructive intramural coronary amyloidosis. *Am J Med.* 2000;109(3):181-188. [\[Crossref\]](#)



Ultrasound-guided microwave ablation for breast tumors: current status and future perspectives

Jin Wang^{1*}
 Hongjuan Wei^{2*}
 Liming Zhou¹
 Xin Chang¹
 Jing Zhao¹
 Hui Chen¹

¹Department of Ultrasound Medicine, Nanjing Lishui People's Hospital, Zhongda Hospital Lishui Branch, Southeast University, Nanjing, China

²Department of Pediatrics, Nanjing Lishui People's Hospital, Zhongda Hospital Lishui Branch, Southeast University, Nanjing, China

ABSTRACT

Ultrasound (US)-guided microwave ablation (MWA) has emerged as a promising minimally invasive therapy for both benign and malignant breast tumors. This review comprehensively examines the current clinical status, technical principles, and therapeutic outcomes of US-guided MWA in breast tumor management. We discuss the biophysical mechanisms of MWA, its advantages over other ablation techniques—such as rapid temperature elevation, the ability to create more extensive coagulation areas, and diminished impact from heat sink phenomena—and the critical role of real-time US guidance in enhancing procedural precision and safety. Clinical evidence supports the efficacy of US-guided MWA in achieving high rates of complete ablation and significant volume reduction for benign tumors, such as fibroadenomas, with minimal complications and excellent cosmetic results. For early-stage breast cancers, initial studies indicate that US-guided MWA provides local tumor control comparable with surgical resection in the short- to mid-term, while also offering the benefits of shorter operation times, reduced hospitalization, and stimulation of systemic antitumor immune responses. However, challenges remain, including technical limitations in treating tumors near critical structures, the lack of long-term oncological data, and operator dependence. Future directions involve technological refinements, integration with artificial intelligence and advanced imaging, combination with immunotherapy, and standardization of protocols. US-guided MWA represents an important advancement toward personalized, organ-preserving breast tumor therapy, with ongoing innovations poised to expand its clinical applicability.

KEYWORDS

Microwave ablation, breast tumors, ultrasound guidance, minimally invasive therapy, thermal ablation

*Joint first authors

Handling editor: Serap Gultekin

Corresponding author: Liming Zhou

E-mail: 36094610@qq.com

Corresponding author: Xin Chang

E-mail: changxinnj@163.com

Received 24 September 2025; revision requested 16 November 2025; accepted 14 January 2026.



Epub: 04.02.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263639

Breast tumors represent a frequently encountered clinical condition in women, often associated with factors such as dysregulation of endocrine hormones, genetic susceptibility, daily behavioral habits, and external environmental exposure. Common symptoms include detectable lumps in the breast tissue, localized pain or tenderness, and occasional nipple discharge. Based on pathological characteristics, these tumors are primarily classified into benign lesions, such as fibroadenomas, and potentially malignant tumors, including early-stage breast cancers.¹ Although benign tumors typically entail minimal health risks, they may still lead to physical discomfort, such as persistent pain. However, malignant tumors, including breast carcinoma, constitute a serious health concern and require timely intervention to prevent adverse outcomes, thereby considerably impacting overall prognosis and survival.² The widespread adoption of breast imaging screening has resulted in a rising incidence of small and asymptomatic tumors, necessitating effective and less invasive treatment strategies. The long-established standard for definitive diagnosis and treatment of suspicious or symptomatic tumors has been surgical excision, either by lumpectomy or mastectomy. Although effective, surgery is associated with inherent drawbacks, including general anesthesia risks, postoperative pain, scarring, potential deformity of the breast contour, and a prolonged recovery period.³

The last two decades have witnessed a paradigm shift in oncology and interventional radiology towards minimally invasive therapies (MITs). These techniques aim to achieve therapeutic efficacy comparable with surgery while minimizing morbidity, preserving organ func-

tion and cosmesis, and reducing procedural time and cost. Within the realm of breast disease management, several image-guided MITs have been developed and clinically implemented.⁴ Among these, thermal ablation modalities, particularly radiofrequency ablation (RFA) and microwave ablation (MWA), have garnered particular attention due to their ability to induce coagulative necrosis in targeted tissues via precise hyperthermia.⁵ Although RFA is the most thoroughly investigated technique, MWA is gaining broader adoption due to its capacity to elicit more extensive ablation volumes with reduced procedure times.⁶

MWA, which originated in 1962, is categorized as a type of thermal ablation within physical ablation therapies and utilizes heat to inactivate local tissues.⁷ MWA utilizes electromagnetic waves in the microwave energy spectrum to agitate water molecules within tissue, generating frictional heat and inducing coagulative necrosis.⁸ It offers faster heating times than other thermal techniques, higher intratumoral temperatures, larger and more predictable ablation zones, less susceptibility to the heat-sink effect (cooling from adjacent blood vessels), and the ability to perform simultaneous multi-probe ablations.⁹ The evolution of MWA technology, particularly with the integration of real-time imaging guidance, has further enhanced its precision and safety profile. Ultrasound (US), due to its widespread availability, cost-effectiveness, and real-time capabilities, has become the modality of choice for guiding MWA procedures. US guidance allows for accurate needle placement, continuous monitoring of the ablation zone, and immediate evaluation of treatment effectiveness, thereby reducing the risk of incomplete ablation or damage to adjacent normal structures (Figure 1).¹⁰

Main points

- Ultrasound (US)-guided microwave ablation (MWA) is a minimally invasive, highly effective alternative to surgery for treating breast tumors, achieving complete ablation with excellent cosmetic results and minimal complications.
- Real-time US is crucial for precision, enabling accurate targeting, continuous monitoring, and immediate assessment while protecting healthy tissues.
- US-guided MWA not only destroys tumors locally but also stimulates a systemic immune response, offering potential for combined therapies and expanded future applications.

US-guided MWA finds broad application in the treatment of diverse clinical conditions, such as hepatocellular carcinoma,¹¹ thyroid nodular goiter,¹² uterine leiomyomas,¹³ and renal neoplasms,¹⁴ with promising therapeutic efficacy observed in clinical practice. Compared with the liver and kidneys, the breast has a superficial location, less complex anatomical architecture, and relatively sparse vascularity, which contribute to excellent US visualization. These characteristics render it an ideal candidate for US-guided MWA. Several clinical studies have demonstrated the efficacy and safety of US-guided MWA in managing both benign and malignant breast tumors. For benign lesions, such as fibroadenomas, MWA offers a non-surgical alternative that can effectively reduce nodule size and alleviate symptoms, with minimal scarring and rapid recovery.^{15,16} In the context of malignant tumors, particularly early-stage breast cancers, MWA has shown promising outcomes in terms of complete ablation rates, local tumor control, and cosmetic preservation, suggesting its potential role as a valuable alternative in selected cases, particularly for individuals unsuitable for surgery or those seeking less invasive therapeutic options.^{17,18}

This review aims to summarize comprehensively the current state of US-guided MWA for breast tumors, encompassing technical principles, clinical applications, and limitations. To achieve this, a narrative synthesis of the literature was performed, which was conducted using electronic databases, primarily PubMed and Web of Science, with a focus on articles published between 2000 and 2025. Key search terms included “microwave ablation,” “breast,” “ultrasound-guided,” “fibroadenoma,” and “breast cancer,” alone and in combination. Given the specialized and evolving nature of this topic, the selection of included studies was not restricted by study design alone but prioritized clinical relevance, scientific rigor, and the illustration of key technical or clinical concepts. This approach encompassed pivotal clinical trials, prospective and retrospective cohort studies, case series, and seminal review articles. The reference lists of retrieved articles were also screened for additional relevant publications. This strategy was chosen to capture comprehensively the most impactful evidence in a field where the total volume of dedicated literature, although growing, remains manageable for a narrative synthesis. Furthermore, we discuss future perspectives, including technological innovations, potential integration with artificial intelligence

(AI) for treatment planning, and the role of US-guided MWA within multimodal breast cancer management pathways.

Biophysical principles and technology of microwave ablation

Mechanism of microwave ablation

MWA is a thermal ablation technique that employs electromagnetic waves within the microwave frequency spectrum (typically 915 MHz to 2.45 GHz) to generate localized hyperthermia and induce coagulative necrosis in target tissues. The underlying mechanism involves the agitation of water molecules within tissue, resulting in frictional heat due to dielectric hysteresis. This process rapidly achieves cytotoxic temperatures (60 °C–100 °C), causing instantaneous protein denaturation and irreversible cellular damage (Figure 1).^{19,20} The efficacy of MWA depends critically on the inherent dielectric properties of the target tissue, particularly its water content and ionic composition, which determine the extent of energy absorption and heat generation.²¹ Unlike RFA, MWA is theoretically and in many clinical scenarios less susceptible to the heat-sink effect from adjacent blood vessels, allowing for more consistent and predictable ablation zones. Furthermore, MWA supports simultaneous multi-probe activation, enabling the treatment of larger or multifocal tumors efficiently.⁹

System components

A standard MWA system comprises three integral components: a microwave generator, an antenna assembly, and a cooling mechanism. The generator is designed to produce microwave energy within specific frequency bands, commercially available at 915 MHz or 2.45 GHz, and the generated signal is subsequently amplified by a dedicated radiofrequency power amplifier. The antenna design is critical, as it dictates energy distribution and ablation morphology. Commonly used antennas include monopolar, bipolar, array-based, triaxial, choked, and slot designs, each suited to particular tumor sizes, depths, and anatomical contexts. For instance, monopolar antennas are cost-effective and suitable for superficial lesions, whereas array antennas are preferred for larger or deeper tumors due to their synergistic emission and reduced cold spots.²² Recent technological advancements have focused on optimizing antenna design to create more spherical and predictable ablation zones with minimal charring and track seeding.⁹ The third part is

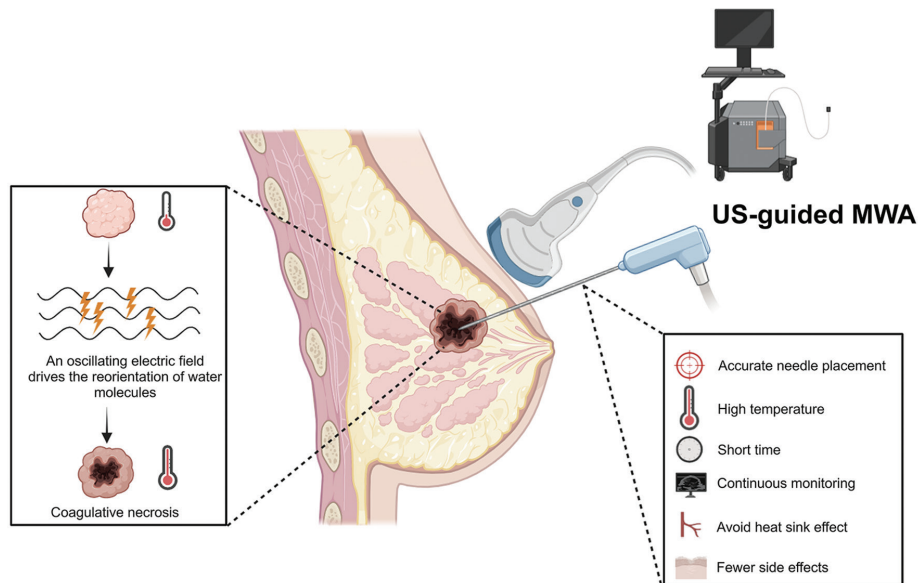


Figure 1. Overview of ultrasound-guided microwave ablation (created using BioRender). US, ultrasound; MWA, microwave ablation.

the cooling system, designed to reduce the probe's temperature using water or gas. This is vital to prevent tissue from carbonizing and to limit heat damage to nearby healthy tissues. In practice, thermal stability during water-cooling is maintained via the circulation of chilled water, whereas gas-cooling (e.g., with CO₂) is chosen for tissues that are sensitive to liquid.²³ Microwave frequency, applied power, and treatment duration are key factors affecting ablation performance. The selection of frequency determines the penetration depth and heating characteristics and should be optimized according to tumor size and anatomical location. The 915 MHz frequency achieves deeper energy penetration (3–4 cm), rendering it appropriate for larger or deeply situated tumors, such as those in the liver or kidneys, and is generally employed at higher power outputs (80–100 W) over extended periods (10–15 minutes) to facilitate larger ablation volumes. In comparison, operating at 2.45 GHz facilitates faster energy delivery, rendering it well-suited for ablating superficial and small lesions. Notably, breast tumors, which are often superficial, are well-suited to ablation at 2.45 GHz using a monopolar antenna. This is commonly achieved by employing moderate power levels of 40–60 W for short durations of 5–10 minutes, ensuring precise and confined treatment.^{9,24,25}

Technical advancements

Recent advances in MWA have markedly extended its clinical utility, driven mainly by three key technological improvements. First, ongoing technical refinements, including

more advanced generator systems and optimized ablation parameters, enable more controlled energy delivery and predictable ablation zones, enhancing treatment safety and efficacy across various tumor types.^{22,26} Second, the integration of real-time imaging guidance, including US, allows precise antenna placement and continuous monitoring of the ablation process, considerably improving tumor targeting and margin assessment.^{10,13} Third, MWA is increasingly recognized for stimulating a durable antitumor immune response, which complements its local effects and supports combination strategies with systemic therapies, such as immunotherapy or chemotherapy.^{9,27,28} These developments together allow more personalized and effective management of tumors, including advanced cases where conventional monotherapies are insufficient.

The role of ultrasound guidance

US guidance is instrumental in optimizing the accuracy, safety, and overall therapeutic outcomes of MWA for breast tumors. As a real-time, radiation-free, and widely accessible imaging modality, US provides exceptional visualization of superficial structures, such as the breast, allowing for accurate needle placement, continuous monitoring of the procedure, and immediate evaluation of therapeutic efficacy.¹⁰ The high-resolution capability of US enables clear differentiation between the target nodule and surrounding tissues, including critical anatomical structures, such as blood vessels and ducts, thereby minimizing the risk of collateral damage (Figure 1).

During the pre-procedural phase, US is used to locate the nodule precisely, determine its size and depth, and plan the optimal trajectory for antenna insertion. This step is crucial for ensuring complete coverage of the target area while avoiding vital structures. Intra-procedurally, real-time US monitoring allows clinicians to observe the formation and expansion of the hyperechoic ablation zone, which corresponds to the region of coagulative necrosis. This dynamic feedback facilitates adjustments in power output or antenna position as needed to achieve adequate margins and avoid under-treatment or overtreatment.¹³ Moreover, color Doppler or contrast-enhanced US (CEUS) can be employed to evaluate vascularity and perfusion changes, providing additional information on treatment efficacy and helping to identify residual viable tissue. Post-procedural US assessment, often supplemented with contrast-enhanced imaging, is essential for confirming the completeness of ablation and detecting any complications, such as hematoma or edema. The ability to perform these evaluations immediately after the procedure contributes to the high clinical acceptability and patient satisfaction associated with US-guided MWA.¹⁶

The integration of US guidance with MWA is particularly advantageous in the breast due to its superficial location and relatively homogeneous parenchyma, which allow for excellent acoustic windows and minimal artifact interference. This synergy not only improves technical success but also supports the broader adoption of MWA as a minimally invasive alternative to surgery for both be-

nign and malignant breast tumors, aligning with the ongoing trend toward personalized and organ-preserving therapies.^{5,15}

Clinical applications of ultrasound-guided microwave ablation for breast tumors

The utilization of US-guided MWA in breast disease management has expanded markedly, encompassing both benign and malignant tumors. Its efficacy is underpinned by the favorable anatomical location of the breast, which allows for excellent US visualization, precise antenna placement, and real-time monitoring of the process.

Treatment of benign breast tumors

Approximately 80% of breast tumors are benign, and with advances in detection technologies, the diagnosis of benign breast tumors has become increasingly common. This trend underscores a growing need for effective treatment strategies. A variety of therapeutic approaches are available for benign breast tumors.^{29,30} As the most common benign breast disease,³¹ fibroadenomas are a common indication for US-guided MWA. This technique offers a minimally invasive alternative to surgical excision for patients seeking to avoid scarring, general anesthesia, or a prolonged recovery period. The primary goals include decreasing tumor volume and relieving co-occurring symptoms, such as pain or palpable discomfort.

In a prospective study, US-guided MWA was applied to 44 benign breast tumors (≤ 3 cm) in 39 patients. The procedure, performed under local anesthesia (mean duration: 74 seconds), achieved complete ablation in 97.5% of cases, as confirmed by CEUS during follow-up. Tumor volume decreased significantly ($P = 0.005$), and no skin burns occurred despite a mean skin-to-tumor distance of 7.5 mm.³² Further supporting evidence comes from a larger cohort study evaluating US-guided MWA in 205 benign breast lesions. Post-ablation, both CEUS and magnetic resonance imaging (MRI) were used to assess treatment efficacy. Results showed a notable decrease in lesion size and volume at 3, 6, or 12 months after the procedure, with 21.5% of lesions completely disappearing by the 12-month follow-up. The CEUS and MRI indicated technical success rates of 87.32% and 82.93%, respectively. The procedure proved safe, with no reports of major complications.³³ Another study focusing on larger benign breast lesions (≥ 2 cm) further supports the efficacy of MWA. In this prospective analysis of 104 lesions treated with US-guid-

ed MWA, complete ablation was achieved in all cases, as confirmed by contrast-enhanced imaging (Figure 2). Over a median follow-up of 12.5 months, the mean volume reduction rate reached 80.2% ($P < 0.001$), with no major complications reported. Multivariate analysis identified lesion location as an independent predictor of volume reduction; adjacency to the areola was associated with greater shrinkage, whereas proximity to the skin correlated with reduced regression. These findings affirm US-guided MWA as a viable and well-tolerated therapeutic approach even for larger benign breast lesions.³⁴ Further evidence from a larger cohort study specifically assessed the therapeutic effects of US-guided MWA in the treatment of breast fibroadenomas. Among 271 lesions in 171 patients, US-guided MWA led to progressive volume reduction and complete regression over time (Figure 3), with complete disappearance observed in 24.3%, 45.5%, and 40.9% of lesions at 1–6, 6–12, and > 12 months, respectively. Smaller lesion diameter (< 1.5 cm) was significantly correlated with more favorable treatment outcomes ($P < 0.05$), whereas factors such as patient age, lesion location, and blood flow showed no significant correlation with outcomes.¹⁶ A comparative analysis was conducted to further evaluate US-guided MWA versus vacuum-assisted excision (VAE) for breast lesions with uncertain malignant potential (B3 lesions). The results indicated comparable technical success rates between the two techniques, with no residual lesions observed in the MWA group (Figure 4), compared with a 3.4% residue rate in the VAE group. Notably, the MWA group experienced significantly fewer postoperative adverse events, particularly in lesions > 2.5 cm, and was associated with lower recurrence, reduced need for re-intervention, and higher patient satisfaction.³⁵ In a comparative retrospective study, US-guided MWA demonstrated marked advantages over both open surgery and US-guided minimally invasive rotary cutting. Patients receiving MWA experienced briefer procedures, less blood loss during surgery, minimal scarring, quicker healing, and reduced postoperative pain at 24–72 hours. Both MWA and rotary cutting resulted in fewer complications and better cosmetic outcomes than open surgery (Figure 5), while showing comparable efficacy between the two minimally invasive approaches. The study highlights US-guided MWA as a particularly effective option, offering rapid recovery, minimal invasiveness, and enhanced patient satisfaction in the management of benign breast lesions.³⁶

For rare benign breast tumors, such as giant leiomyoma, which pose significant cosmetic concerns due to large size or periareolar location, US-guided MWA has also demonstrated promising outcomes. In a case report involving an 8 cm leiomyoma, MWA achieved a volume reduction rate of 69.8% at 10-month follow-up (Figure 6), with excellent cosmetic results and no major complications, highlighting its potential as a tissue-sparing alternative to surgery in selected cases.³⁷

Therefore, accumulating evidence supports US-guided MWA as a feasible and effective minimally invasive treatment for benign breast tumors. This technique is associated with high rates of complete ablation, large volume reduction, and even complete disappearance of tumors over time (Table 1). It offers a favorable safety profile with minimal complications, preserves cosmetic outcomes, and reduces vascular supply to treated areas. Compared with conventional surgical excision, US-guided MWA demonstrates satisfactory therapeutic efficacy, shorter procedure times, reduced intraoperative blood loss, less postoperative pain, and faster recovery. These advantages position US-guided MWA as a promising non-surgical alternative, particularly for selected tumors based on size and location.

Treatment of malignant breast tumors

Globally, breast cancer continues to be the most common cancer and foremost cause of cancer death in women. The disease encompasses a highly heterogeneous group of tumors, classified by histological type, growth patterns, and molecular profiles, all of which are critical for prognostic stratification and therapeutic decision-making. Current treatment strategies include surgery, radiotherapy, chemotherapy, endocrine therapy, targeted agents, and immunotherapy. The selection of treatment modalities is highly individualized and depends on tumor stage, biological subtype, and patient preferences.^{38,39} In addition, various studies have reported comparable overall survival and distant metastasis-free survival between breast-conserving surgery (BCS) and mastectomy, with some evidence suggesting even improved long-term outcomes and lower recurrence rates among patients undergoing conservation therapy.^{40–43} Within this context, techniques such as US-guided MWA are being investigated as promising minimally invasive alternatives that align with the principles of conserving breast tissue while achieving effective tumor control.

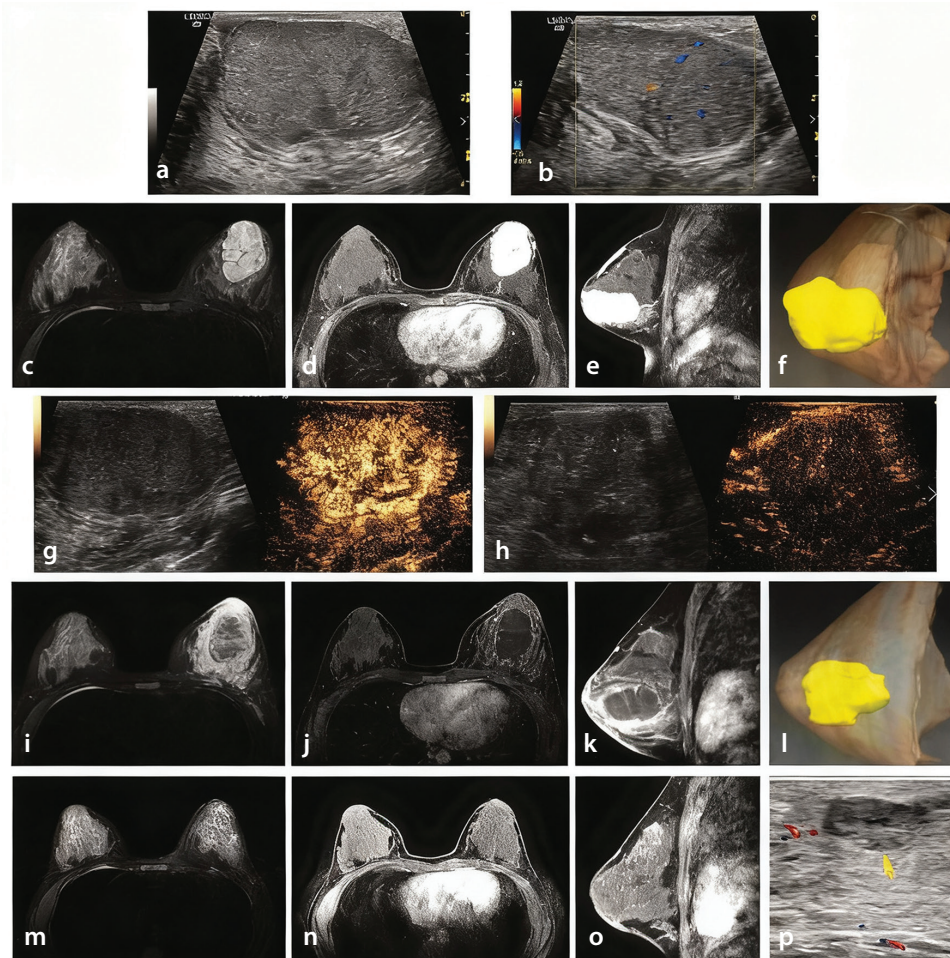


Figure 2. Imaging evaluation before and after microwave ablation (MWA) in one representative case. (a) Benign breast lesion (BBL) with $6.0 \times 3.5 \times 4.9$ cm size showing hypoechogenicity with a clear margin on ultrasound (US); (b) color Doppler flow imaging before MWA showing the arterial blood flow signal in the BBL parenchyma; (c) a hyper-signal in the BBL parenchyma on magnetic resonance imaging (MRI) T2-weighted images before MWA; (d, e) contrast-enhanced MRI in axial and sagittal sections showing a hyper-enhancement in the BBL parenchyma in the arterial phase before MWA; (f) three-dimensional (3D) visualization showing the spatial structure of BBL; (g) BBL before MWA showed hyper-enhancement in the arterial phase of contrast-enhanced ultrasound (CEUS); (h) after MWA, a hypoechoic ablation zone with a size of $5.0 \times 3.6 \times 4.5$ cm is shown in the tumor region by CEUS; (i) hypo-signal in the BBL parenchyma on MRI T2-weighted images after MWA; (j, k) contrast-enhanced MRI in axial and sagittal sections showing no enhancement in the BBL parenchyma in the arterial phase after MWA; (l) 3D visualization showing the spatial structure of the ablation zone; (m–p) after 12 months follow-up, the ablation debris decreased gradually to a size of $1.1 \times 0.4 \times 1.2$ cm on T2-weighted images (m) and arterial phase (n, o) of MRI images and the US image (yellow arrows in p).³⁴

Several recent clinical studies have further validated the therapeutic profile of US-guided MWA for managing malignant breast tumors, particularly in early-stage disease. For instance, a feasibility study involving 41 patients with solitary breast cancers (≤ 3 cm) demonstrated that US-guided MWA achieved complete ablation in 90% of cases (95% confidence interval: 76.9%–97.3%) as confirmed by histochemical staining, with a mean ablation time of only 4.48 minutes and minimal reversible complications.⁴⁴ Another pilot cohort study compared US-guided MWA with nipple-sparing mastectomy in patients with invasive ductal carcinoma (≤ 5 cm) (Figure 7). With a median follow-up of 26.7 months, both groups showed comparable tumor progression rates ($P = 0.16$), no cancer-related deaths, and no major complications. However, the MWA group exhibited

significantly shorter hospitalization times and superior cosmetic outcomes ($P < 0.001$), supporting its suitability, especially for elderly populations ineligible for surgery.⁴⁵ In a multicenter prospective cohort research focusing on elderly individuals with hormone receptor-positive/human epidermal growth factor receptor 2-negative breast cancer, US-guided MWA combined with endocrine therapy yielded disease-free and overall survival outcomes comparable with those of standard surgery with adjuvant therapy (Figure 8), while significantly reducing the hospitalization period (7.1 vs. 13.0 days, $P < 0.001$).⁴⁶ Additionally, a propensity score-matched study comparing US-guided MWA with BCS in patients with early breast cancer found no significant differences in tumor progression, overall survival, or disease-specific survival after a median follow-up of 43

months. US-guided MWA exhibited shorter procedure times, fewer complications, and significantly better cosmetic satisfaction (Figure 9) ($P < 0.001$), highlighting its potential as a minimally invasive alternative to BCS in selected cases.⁴⁷ In a pioneering window-of-opportunity trial focusing on a particularly challenging patient subset where tumors affect the skin or nipple-areola complex (NAC), it demonstrated the viability of US-guided MWA as a local therapeutic option.⁴⁸ The study enrolled 15 inoperable patients with advanced disease, employing innovative techniques such as hydro-dissection and variable power ablation (20–60 W) to protect superficial structures while achieving complete tumor necrosis. Technical success and effectiveness were both 100%, with a median follow-up of 33.5 months. Although instances of skin burns

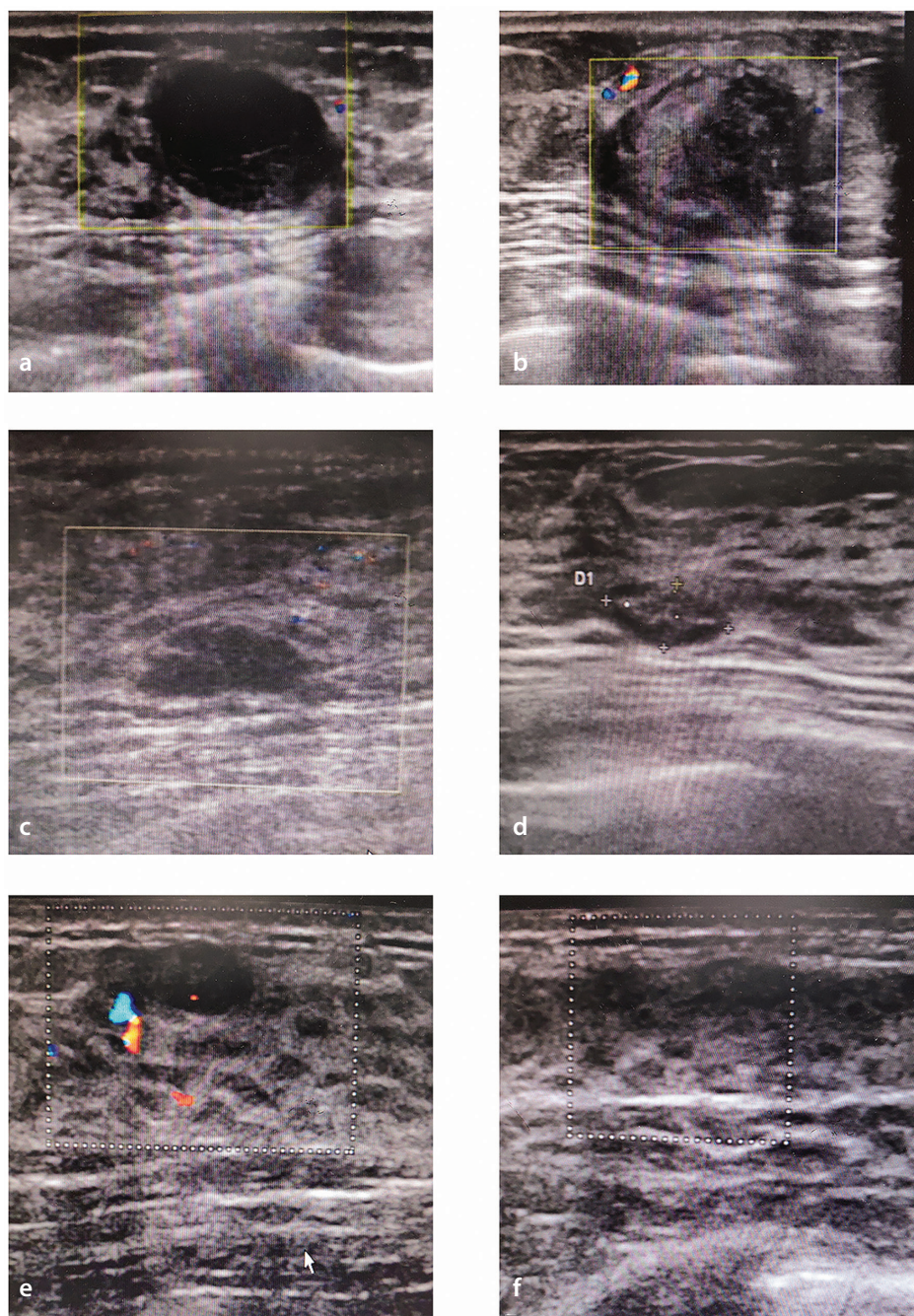


Figure 3. Depicts three cases of patients with different responses to treatment. Ultrasound before microwave ablation (MWA) showed three lesions with clear boundaries (**a**, **c**, **e**). At 8 months, 9 months, and 11 months after the MWA, the figure shows the results of enlargement (**b**), reduction (**d**), and complete regression (**f**), respectively.¹⁶

and nipple loss occurred in some cases, all wounds healed within a median duration of 3.7 months, and > 93% of patients rated their aesthetic results as excellent or good. Importantly, MWA markedly improved the quality of life and even enabled subsequent curative surgery in two patients by reducing tumor burden. This study supports MWA as a promising minimally invasive alternative for complex breast cancers with skin/NAC involvement, particularly in patients ineligible for standard surgical treatment.

Besides ablating tumors locally, MWA can elicit systemic antitumor immune responses, which contribute to sustained therapeutic benefits and may inhibit the growth of distant untreated tumors.⁹ This immunostimulatory effect positions MWA as a valuable component within combinatorial therapeutic approaches for breast cancer. A clinical study in early-stage breast cancer revealed that US-guided MWA not only achieved a high complete ablation rate (91.4%) but also significantly promoted a Th1-polarized immune response, characterized by an increase

in inducible costimulator (ICOS)-activated CD4⁺ T-cells (Figure 10) and elevated IFN- γ levels. T-cell receptor sequencing further indicated clonal expansion of T lymphocytes recognizing breast tumor-specific antigens, suggesting the induction of a durable adaptive immune response.⁴⁹ Single-cell transcriptomic analysis revealed that US-guided MWA induced activation of NK cells and CD8⁺ T-cells (Figure 11), along with enhanced ICOS expression in CD4⁺ T-cells. Immune checkpoint blockade on post-MWA samples resulted in higher T-cell activity, supporting the

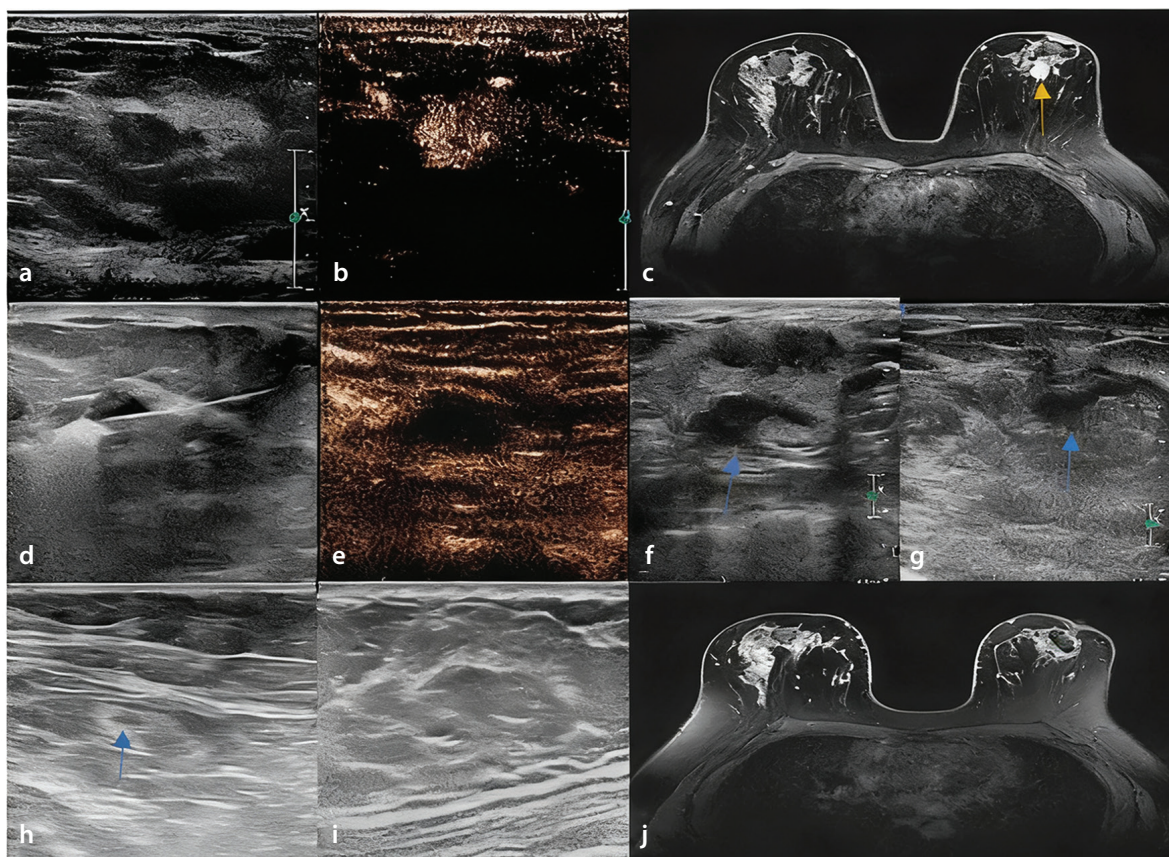


Figure 4. Technical success of microwave ablation (MWA) in one representative case (a 36-year-old woman with a left intraductal papillary pattern of the breast of $17 \times 11 \times 9$ mm in size). (a) Ultrasound (US) before MWA showed a hypoechoic mass with clear boundary; (b) the tumor before ablation with contrast-enhanced US (CEUS) showed homogeneous enhancement on CEUS; (the yellow arrow points to the mass) (c) magnetic resonance imaging (MRI) showed the tumor before ablation with notable enhancement; (d) US showed the tumor was being treated with MWA; (e) the ablated mass without enhancement on CEUS; (f-i) US variance of the breast benign tumor after MWA during follow up at 3, 6, 12, and 24 months, respectively. During the process, the tumor size decreased gradually and finally disappeared on the US image; (the blue arrow points to the mass); (j) at the follow-up of 24 months, no masses were detected on MRI. Reprinted from³⁵ with permission from John Wiley and Sons.

potential synergy between MWA and immunotherapy.²⁷ A window-of-opportunity trial combining preoperative MWA with a single dose of anti-PD-1 antibody (camrelizumab) demonstrated that the regimen was well-tolerated and practicable, without delaying surgery. This approach led to enhanced cytotoxic and memory functions in CD8⁺ T-cells, with monocyte-mediated MHC-I pathway activation contributing to the improved T-cell response.⁵⁰ Moreover, when compared with RFA, US-guided MWA was found to induce superior cytolytic activity in peripheral T-cells and promote memory CD4⁺ T-cell expansion (Figure 12). Single-cell analyses indicated that these differences may stem from distinct antigen presentation mechanisms, with dendritic cells and altered fatty acid metabolism playing a crucial role in the MWA-induced immune environment.¹⁸

Based on the evidence presented, US-guided MWA has emerged as a safe and effective minimally invasive therapeutic option for selected patients with malignant breast tumors, particularly those with early-stage disease. Clinical studies demon-

strate high rates of complete ablation, short-to mid-term local tumor control that appears comparable with surgical interventions, and superior cosmetic outcomes with reduced procedural times and shorter hospital stays (Table 1). Importantly, US-guided MWA not only achieves localized tumor destruction but also stimulates systemic antitumor immune responses, including Th1 polarization and T-cell activation, which may enhance long-term disease control and support combination strategies with immunotherapy. These attributes suggest that US-guided MWA could become a valuable alternative to conventional surgery in selected patient populations, especially in elderly patients or those ineligible for standard operative management, while also offering a promising platform for multimodal treatment approaches in breast oncology.

Limitations and challenges of ultrasound-guided microwave ablation for breast tumors

The application of US-guided MWA, although promising, is constrained by several

technical, biological, and clinical challenges, which must be carefully considered for patient selection and procedural planning.

Technical and procedural limitations

Key technical constraints are often related to tumor characteristics and the physical principles of MWA. Treating tumors near critical structures, such as the skin (< 0.5 cm), NAC, or chest wall, remains challenging due to the inherent risk of thermal injury, including burns or deformity. Techniques such as hydro-dissection can create protective fluid barriers, but they add procedural complexity.⁴⁸ Furthermore, intraprocedural monitoring, although feasible with US, is imperfect. The hyperechoic gas microbubbles generated during ablation can obscure the true ablation margin and the underlying tumor, potentially compromising the assessment of complete coverage, especially for less experienced operators.⁵¹ This is of particular concern with larger tumors, which may require multiple overlapping ablations or moving techniques, increasing the risk of incomplete treatment or collateral damage due to the

cumulative thermal effect and obscured visibility. Finally, as a destructive modality, MWA does not provide a specimen for definitive post-procedural histopathological analysis, which is a major drawback compared with surgical excision or VAE. This underscores the paramount importance of a conclusive histopathological diagnosis via core needle biopsy prior to ablation. The risk of false-negative

biopsies necessitates meticulous sampling from multiple sites within a suspicious lesion to minimize the chance of missing a malignancy before proceeding with ablation.

Methodological heterogeneity and evidence gaps

Beyond technical and biological considerations, the current body of evidence on

US-guided MWA for breast tumors has several methodological limitations that warrant careful interpretation. There is substantial heterogeneity across the available clinical studies, including variability in patient inclusion criteria, tumor subtypes, applied MWA parameters (power, duration), and techniques for assessing treatment efficacy (e.g., CEUS, MRI, histology). This heterogeneity complicates direct cross-study comparisons and meta-analyses. Furthermore, many published studies are limited by small sample sizes and a predominance of single-arm, non-randomized designs. Consequently, although short- and mid-term outcomes reported in cohorts^{45,48} are encouraging, the absence of large-scale, randomized controlled trials (RCTs) with long-term (> 5 years) follow-up currently represents the most critical evidence gap. The lack of Level I evidence precludes definitive conclusions regarding its long-term oncological equivalence to surgery and underscores the need for rigorously designed, prospective RCTs to validate these preliminary findings.

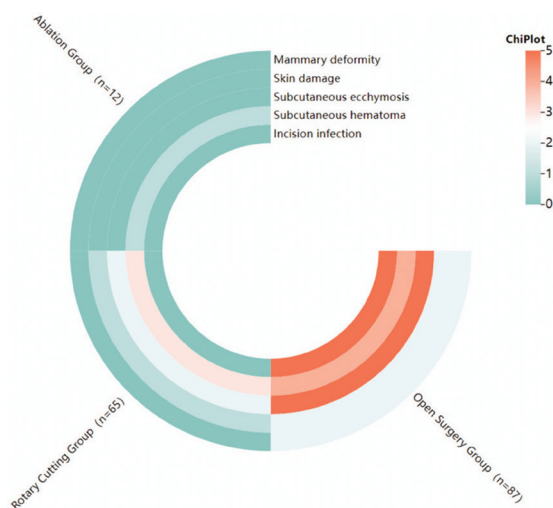


Figure 5. Postoperative complications in the ablation, rotary cutting, and open surgery groups.³⁶

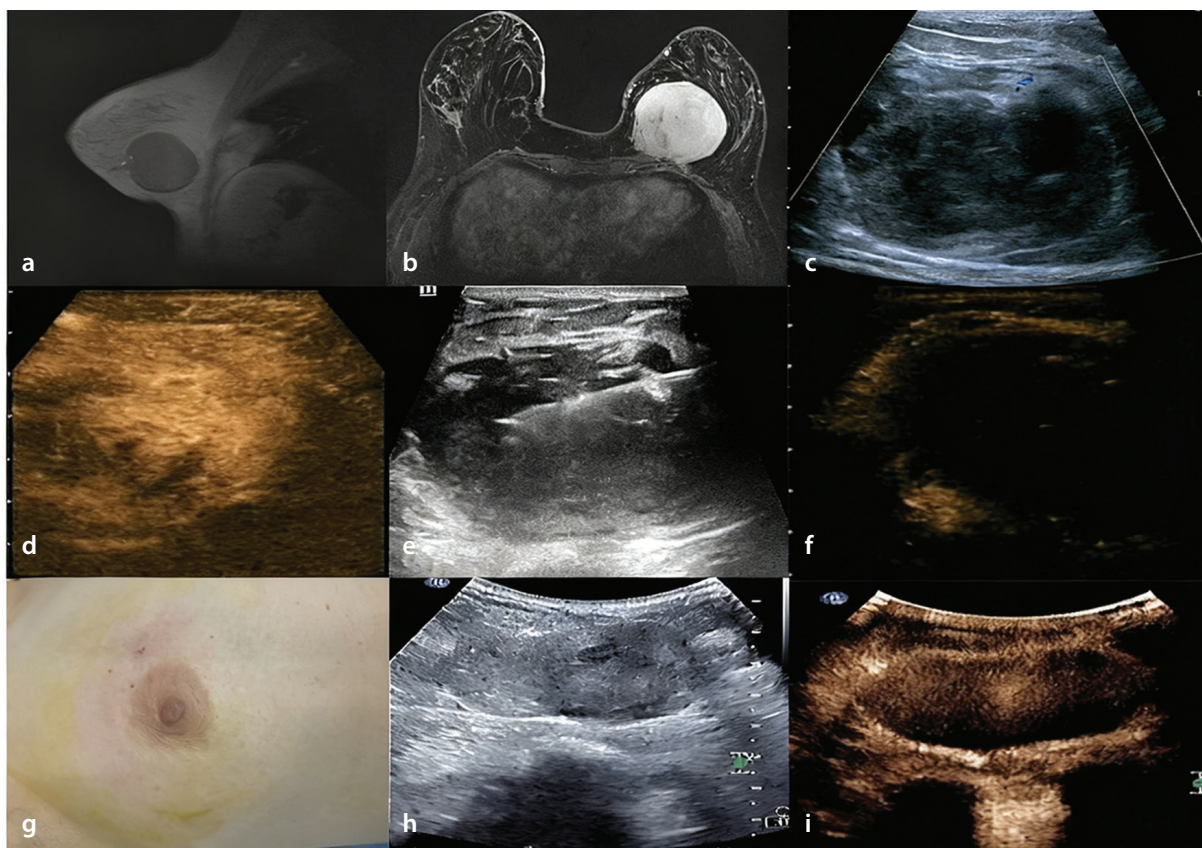


Figure 6. A 37-year-old female with a huge mass in the left breast. (a, b) The location of the tumor on enhanced magnetic resonance imaging; (c) sonographic appearance of the tumor on two-dimensional ultrasound (2D-US); (d) contrast-enhanced US (CEUS) showed homogeneous hyperenhancement of the tumor during the arterial phase; (e) during US-guided microwave ablation; (f) the area of ablation showed no enhancement in arterial phase and venous phase after ablation; (g) the skin of the breast was slightly ecchymosis after ablation, and the appearance of the breast was intact; (h, i) after 10 months of follow-up, 2D-US showed that the mass was considerably reduced, and CEUS showed that there was no enhancement in the arterial phase and venous phase of the ablation area.³⁷

Table 1. Summary of complication profiles from selected US-guided MWA studies

Tumor type	Study ID	Study type	Patients/lesions (n)	Key efficacy	Minor complications (incidence)	Major complications (incidence)	Definitions/notes
Benign tumors	Zhou et al. ³²	Prospective	39 patients/44 lesions (37/41 with follow-up)	Complete ablation: 97.5% (40/41); significant volume reduction ($P = 0.005$)	Slight to moderate pain: 34.1% (14/41); slight skin depression: 5.4% (2/37); oedema and swelling (incidence not reported)	None	Safe (no epidermal burn) even with mean skin-to-tumor distance of 7.5 mm
	Zhang et al. ³³	Prospective	182 patients/205 lesions	Complete ablation: 87.32% (179/205) by CEUS, 82.93% (170/205) by MRI	Local pain and skin erythro swelling: 8.2% (15/182); duct ectasia: 4.9% (9/182); fat liquefaction: 2.2% (4/182)	None	Serious complications: hemorrhage, serious pain and fat necrosis, etc.
	Cui et al. ³⁴	Prospective	80 patients/104 lesions	Complete ablation: 100% (104/104); volume reduction: 80.2% at 12 months ($P < 0.001$)	Pain (slight to moderate): 100% (80/80); skin redness and edema: most patients (incidence not reported)	None	No immediate or delayed complications (skin burn, pectoralis injury, infection or nipple discharge)
	Liu et al. ¹⁶	Retrospective	171 patients/271 lesions	Complete ablation: 24.29% (34/140) at 1–6 months, 45.45% (50/110) at 6–12 months, and 40.91% (18/44) at > 12 months	Slight/moderate pain: 14.6% (25/171); severe pain: 1.7% (3/171); skin scalding: 0.6% (1/171)	None	No fat liquefaction or serious complications occurred during the operation
	Zhang et al. ³⁵	Retrospective	101 lesions	Technical success: 100%; no recurrence: 100% at 24-month follow-up	Localized swelling/pain: 11.9% (12/101); skin ecchymosis: 5.9% (6/101); tape allergy: 1.0% (1/101); infection: 2.0% (2/101); self-absorbable hematoma: 1.0% (1/101)	Skin fistula: 1.0% (1/101)	The technical success rate was defined as no enhancement in any part of the lesion on the first postoperative follow-up imaging
	Lu et al. ³⁶	Retrospective	12 patients	Operation time: 18.39 ± 7.05 minutes; blood loss: 4.12 ± 1.51 mL; healing time: 2.18 ± 1.02 days; complication rate: 8.33%	Subcutaneous hematoma: 8.33% (1/12); no infection, ecchymosis, skin damage, or deformity reported	None	Compared with open surgery and rotary cutting: superior cosmesis (100% excellent/good); minimal invasiveness; low pain scores (VAS 1.02–3.05)
	Zhang et al. ³⁷	Case report	1 patient/1 lesion	Volume reduction: 69.8% at 10-month follow-up; complete ablation confirmed by CEUS	Postoperative pain (VAS 4); slight skin ecchymosis, redness and swelling (resolved in 4 weeks)	None	Giant leiomyoma (7.6 × 6.6 × 7.5 cm); first MWA report for breast leiomyoma; excellent cosmetic outcome

Tumor type	Study ID	Study type	Patients/lesions (n)	Key efficacy	Minor complications (incidence)	Major complications (incidence)	Definitions/notes
Malignant tumors	Zhou et al. ⁴⁴	Prospective	41 patients/41 lesions	Complete coagulation: 90% (37/41) overall; 95% (36/38) for invasive ductal carcinoma; mean ablation time: 4.48 minutes	Skin burn: 2.4% (1/41); pectoralis muscle thermal injury: 4.9% (2/41)	None	No necrosis of the skin flaps, infection, or other adverse effects were noted in all 41 patients during and after the procedure
	Yu et al. ⁴⁵	Retrospective	21 patients/22 lesions	Technical efficacy: 100%; LTP: 4.5% (1/22)	A mild sensation of heat, pain, and local swell in the ablation site: 100%	None	Technical efficacy was defined as the rate of no residual tumor 1 month after the procedure; major complication included severe skin injury, abscess, hematoma, pneumothorax, tumor cell implantation, wound dehiscence, skin flap necrosis and nipple areola complex necrosis
	Zhong et al. ⁴⁶	Prospective	33 patients/33 lesions	Technical effectiveness: 100%; local recurrence: 3.0% (1/33); DFS/OS comparable with surgery	Not specified in detail; only "no ablation-related adverse events" reported	None	Technical effectiveness of MWA was defined as complete ablation at follow-up enhanced imaging 1 month after MWA
	Dai et al. ⁴⁷	Retrospective	21 patients/28 lesions	Technical success: 100%; local recurrence: 4.8% (1/21); comparable OS/DSS with BCS	Slight skin redness and subcutaneous edema (subside within 3 days) (most patients)	None	The technical success of MWA is defined as no enhancement of the entire tumor area during the arterial phase; no skin burns or fat necrosis; excellent cosmetic results: 100%; shorter operation time vs. BCS; suitable for early breast cancer up to 5 cm
	Dai et al. ⁴⁸	Retrospective	15 patients/16 lesions	Technical success: 100% (15/15); technical effectiveness: 100% (15/15); local recurrence: 26.7% (4/15)	Mild skin redness and swelling: 46.6% (7/15); slight bleeding: 26.6% (4/15); mild pain: 20.0% (3/15); mild tenderness: 26.6% (4/15)	Nipple loss: 6.7% (1/15); skin burn and nipple loss: 26.7% (4/15); infection: 6.7% (1/15)	Technical success: tumors treated per protocol and completely covered by ablation zone; technical effectiveness: no enhancement on follow-up imaging at 1 month
	Zhou et al. ⁴⁹	Prospective	35 patients/35 lesions	Complete ablation: 91.4% (32/35) confirmed by pathology or follow-up (median 36 months)	Local swelling at treatment site: 100%, resolved within 1 week	None	Study focused on immune response; MWA induced antitumor immunity
US, ultrasound; MWA, microwave ablation; CEUS, contrast-enhanced ultrasound; MRI, magnetic resonance imaging; LTP, local tumor progression; VAS, visual analog scale; DFS, disease-free survival; OS, overall survival; DSS, disease-specific survival; BCS, breast-conserving surgery.							

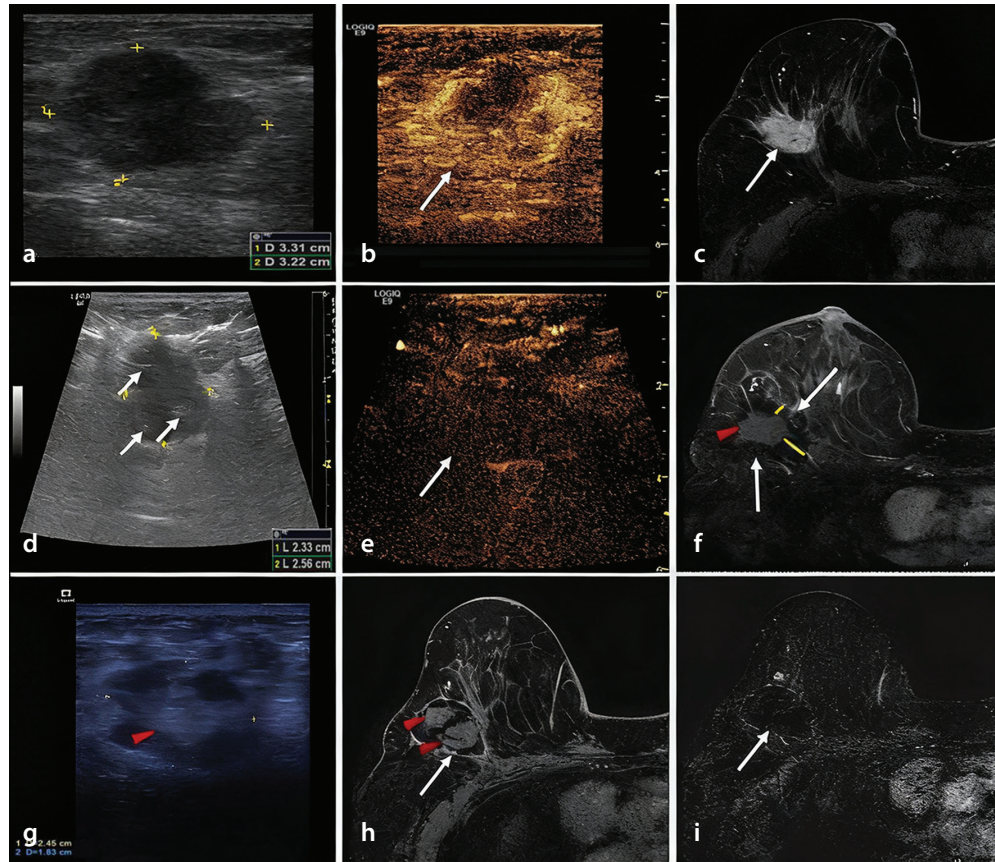


Figure 7. A 68-year-old woman with invasive ductal carcinoma of the right breast. (a) Ultrasound (US) scan before microwave ablation (MWA) shows the hypoechoic mass (arrow) with a size of 3.3×3.2 cm; (b) contrast-enhanced US (CEUS) before MWA shows the mass is hyper-enhanced (arrow) in arterial phase; (c) transverse contrast-enhanced magnetic resonance imaging (MRI) shows hyperintensity masses (arrow) before MWA in arterial phase; (d) US scan shows the heterogeneously hypoechoic mass (marker) with a size of 2.6×2.3 cm immediately after MWA (ghost size). Hyperechoic needle tracts can be seen in the ablated mass (arrow); (e) CEUS immediately after MWA shows the mass is non-enhanced (arrow) in arterial phase; (f) contrast-enhanced MRI image shows hyperintensity ghost of mass (red arrow) and the peripheral hypointensity treatment zone (white arrow) in the arterial phase 3 days after MWA. The ablation margin is from 1.2 to 2.2 cm (yellow lines), which can be measured in the hospital information system; (g) US scan shows the heterogeneously ablation zone (marker) shrinks to a size of 2.5×1.8 cm at 18 months after MWA. The ghost of mass (arrow) is surrounded by hypoechoic adipose tissue; (h) contrast-enhanced MRI image shows the treatment zone (white arrow) is non-enhanced with a clear capsule and the central hyperintensity ghost of mass (red arrow) in arterial phase at 18 months after MWA; (i) MRI silhouette shows no signal for the ablation zone with clear fibrous capsule and margin (arrow).⁴⁵

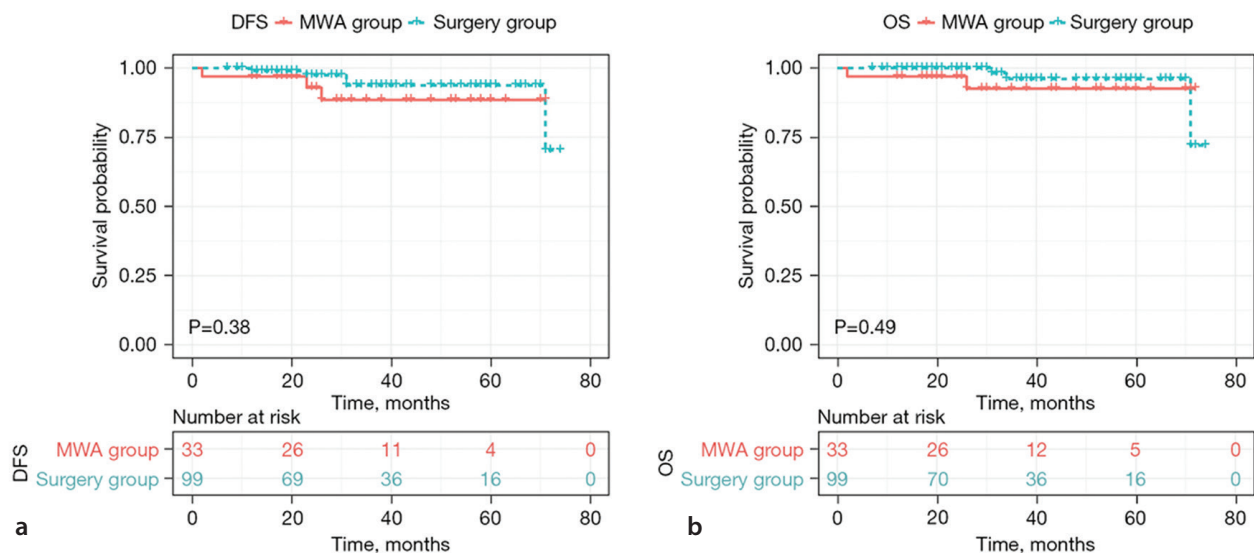


Figure 8. Kaplan–Meier curves for (a) disease-free survival and (b) overall survival.⁴⁶ DFS, disease-free survival; MWA, microwave ablation; OS, overall survival.

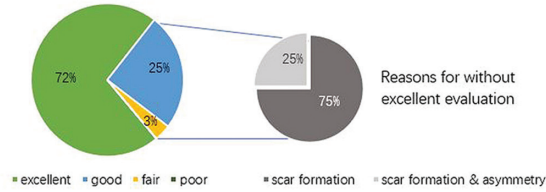
Cosmetic Results in Total cohort

MWA



■ excellent ■ good ■ fair ■ poor

BCS



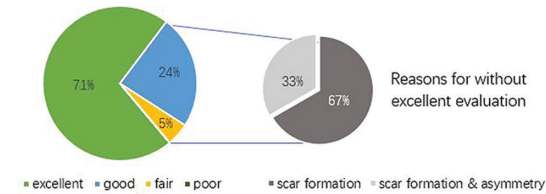
Cosmetic Results in PSM cohort

MWA



■ excellent ■ good ■ fair ■ poor

BCS



a

b

Figure 9. Cosmetic results and reasons for without excellent evaluation between selected patients with early breast cancer who underwent MWA and breast-conserving surgery. (a) Cosmetic results in total cohort; (b) cosmetic results in propensity score matching cohort.⁴⁷ MWA, microwave ablation; BCS, breast-conserving surgery; PSM, positive surgical margin.

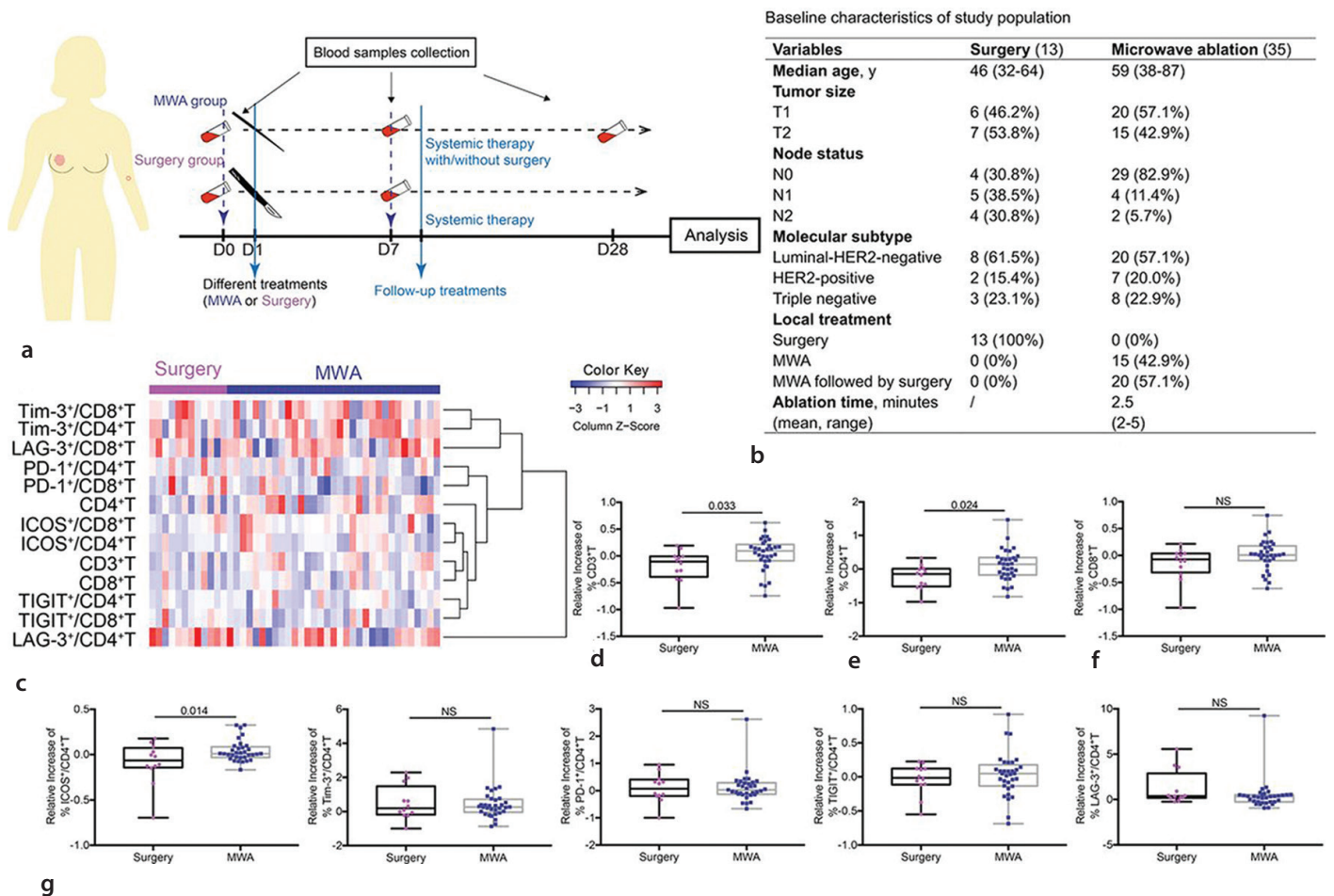


Figure 10. Phenotypical characterization of peripheral T-cells in patients treated with microwave ablation (MWA). (a) Schematic illustrating the study design; (b) basic characteristics of enrolled patients; (c) heat map of the changes in peripheral T-cells, CD4⁺, and CD8⁺ T-cell subsets in the MWA group (n = 33) and the surgery group (n = 12). The increased frequencies of peripheral T-cells (d); CD4⁺ (e); but not CD8⁺ (f) T-cells in patients treated with MWA were compared with those in surgery; (g) scatter plots showing the changes in the activated (inducible costimulator) and exhausted (LAG-3, TIGIT, TIM-3, and PD-1) CD4⁺ T-cells. Data are presented as mean ± standard deviation.⁴⁹

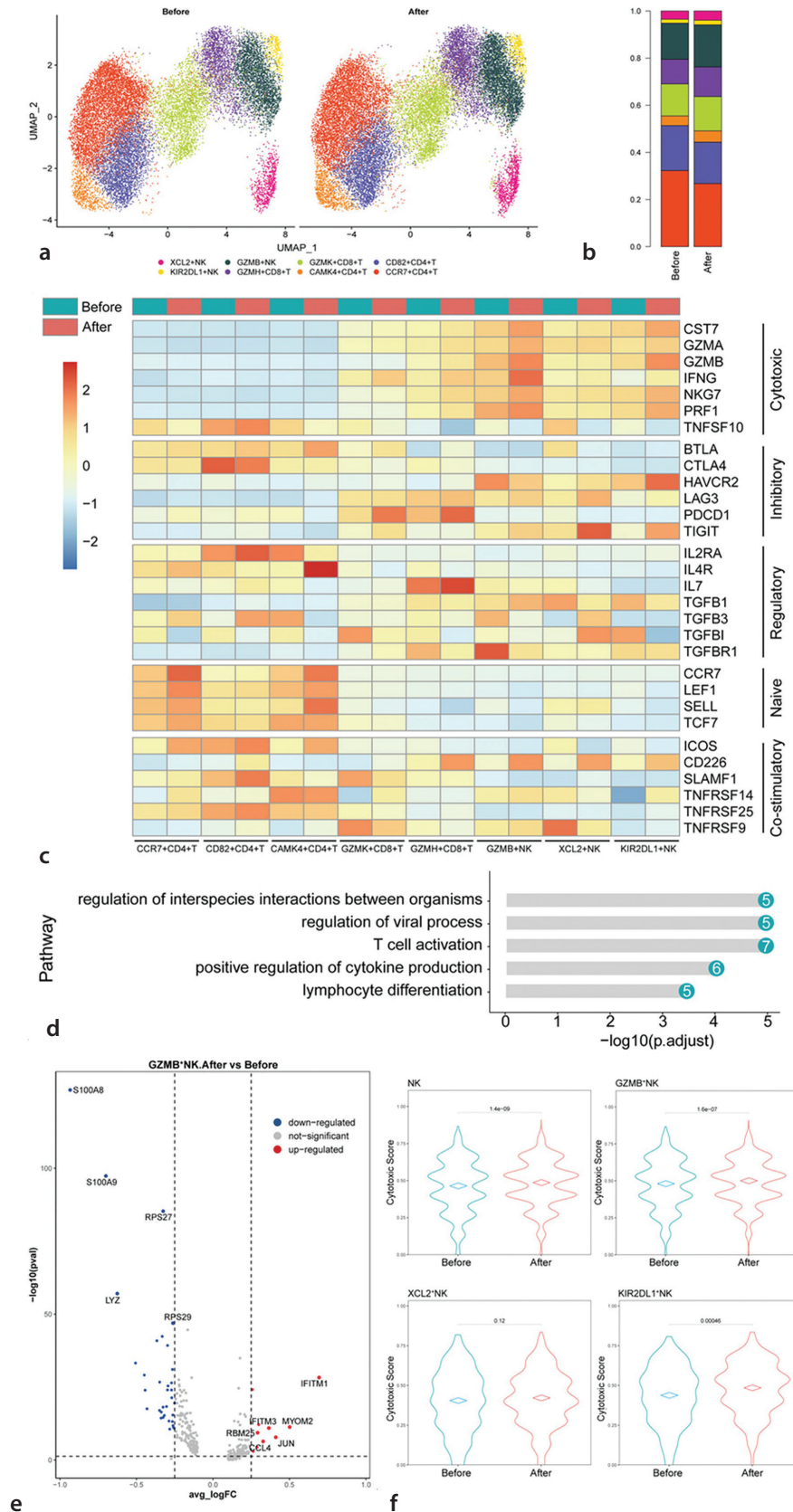


Figure 11. Overview of changes in the NK and T-cells, and the activated phenotypes of NK cells induced by microwave ablation (MWA) of breast cancer (n = 6). (a) UMAP analysis of peripheral NK and T-cells showing eight clusters before and after MWA; (b) the proportions of each cell cluster before and after ablation; (c) heatmap of the gene sets of cytotoxicity, exhaustion/inhibitory, regulatory, naive and co-stimulation of these eight cell clusters; (d) gene ontology enrichment pathway analysis of genes preferentially upregulated in XCL2+ NK cells; (e) volcano plot showing upregulated genes of peripheral GZMB+ NK cells induced by MWA; (f) the cytotoxic scores of different peripheral NK cell clusters before and after MWA.²⁷

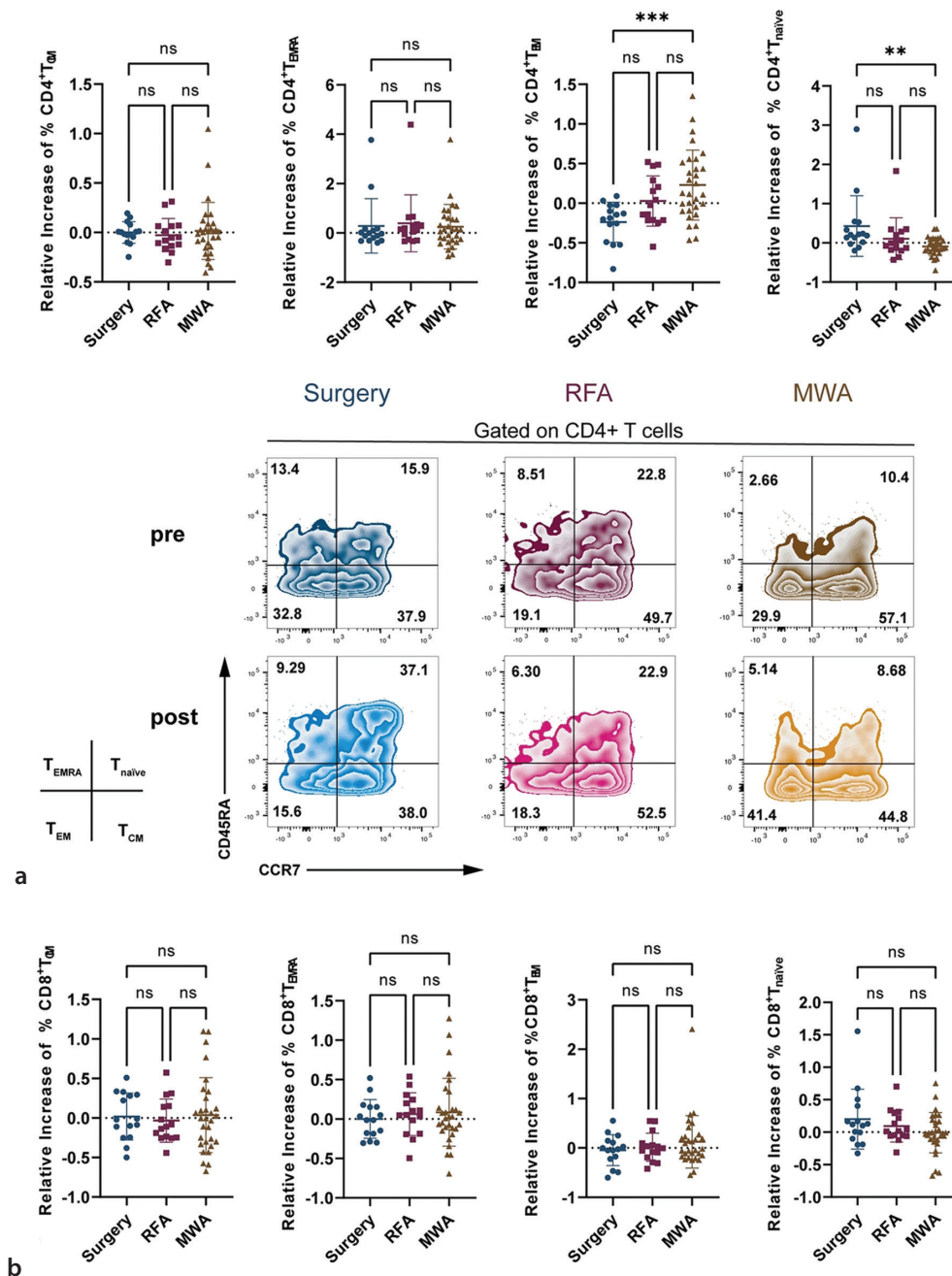


Figure 12. Changes in peripheral memory CD4+ and CD8+ T-cell subsets among different treatment groups. Flow cytometry gating graphs of CD4+ T-cells and relative increases in CD4+ (a) and CD8+ (b) TCM, TEMRA, TEM, and Tnaive subsets in patients treated with RFA (n = 15) or MWA (n = 30) compared with surgery (n = 15). Data are shown as mean $\pm$ standard deviation. ** $P < 0.01$; *** $P < 0.001$. ns, not significant; RFA, radiofrequency ablation; MWA, microwave ablation.

Biological uncertainties and standardization barriers

Considerable biological and diagnostic uncertainties persist. For benign disease, there is a lack of consensus on whether to intervene in certain conditions, such as hyperplastic nodules (excluding severe or atypical hyperplasia), often leading to management based on patient symptoms or preference rather than standardized guidelines. For malignant applications, the biological response to thermal ablation may vary based on tumor biology (e.g., molecular subtypes, presence of an extensive intraductal component)

and microenvironmental factors, influencing the risk of residual disease or recurrence.⁵² Finally, widespread adoption is hindered by a lack of universally standardized protocols and guidelines regarding power settings, ablation time, and endpoint determination. Success remains operator-dependent, requiring significant expertise in both ultrasonography and interventional technique.

Future perspectives

The continued evolution of US-guided MWA for breast tumors will be shaped by interdisciplinary innovations aimed at enhancing

precision, expanding clinical applicability, and integrating within broader oncological frameworks. Based on current evidence and emerging technological trends, the following key directions are envisioned:

Technological refinements and smarter ablation systems

Future MWA systems will likely feature miniaturized and adaptive antenna designs (e.g., triaxial, slot, or multi-tine configurations) capable of conforming to complex tumor morphologies while sparing critical adjacent structures. Coupled with real-time

temperature monitoring through non-invasive methods, such as MRI thermography or integrated fiber-optic sensors, these advances will allow dynamic modulation of energy delivery, reducing the risk of incomplete ablation or collateral damage. Furthermore, the integration of AI for pre-procedural planning and intraprocedural guidance could standardize treatment precision;⁵³ AI algorithms trained on multi-parametric imaging and radiomic data may predict optimal ablation parameters, antenna trajectories, and even early signs of treatment response.

Advanced imaging integration and radiomics

The fusion of real-time US with pre-procedural MRI or CT will enhance spatial accuracy and margin assessment, particularly for tumors adjacent to critical anatomy. US radiomics,⁵⁴ which involves extracting high-throughput quantitative features from grayscale, Doppler, and CEUS, holds significant promise for personalizing US-guided MWA. Radiomic signatures may predict dielectric properties, ablation susceptibility, and immune microenvironment features, enabling patient-specific therapy planning. Post-ablation, delta-radiomics could serve as a non-invasive biomarker for treatment efficacy and residual disease.⁵⁵

Immunomodulation and combination strategies

US-guided MWA-induced immunogenic cell death provides a strong rationale for combining ablation with immunotherapy. Future studies should focus on optimizing the timing and sequencing of immune checkpoint inhibitors (e.g., anti-PD-1) with US-guided MWA to amplify systemic antitumor responses. Nanotechnology represents a research avenue that could further enhance this synergy through thermally responsive nanocarriers co-loaded with immunomodulators or chemotherapeutics, enabling targeted delivery to the ablation zone and residual tumor niches.^{9,56} Research into spatial immune dynamics via single-cell transcriptomics will help identify novel biomarkers and combination targets.

Standardization and validation

There is a critical need for standardized international guidelines on technical parameters, endpoint definitions, and patient selection criteria for both benign and malignant applications. Large-scale prospective registries and RCTs with long-term follow-up

(> 5 years) are essential to validate oncological outcomes, including local recurrence, survival, and cosmetic results. Incorporating patient-reported outcomes and cost-effectiveness analyses will further support the integration of US-guided MWA into clinical care pathways.

Cost-effectiveness, economic considerations, and accessibility

The economic viability of US-guided MWA, as an emerging technology, is a key factor influencing its broader adoption. The initial costs associated with the acquisition of MWA equipment and specialized operator training present major barriers. Moreover, variability in medical insurance coverage for ablation procedures remains a challenge. However, the inherent advantages of US-guided MWA, such as its ability to be performed under local anesthesia, its shorter operative times, and the reduced need for hospitalization, suggest a potential for long-term cost savings compared with traditional surgical resection, though this requires formal validation. These savings would stem from decreased use of operating room resources, shorter inpatient stays, and faster patient recovery. Concurrently, expanding the clinical indications for US-guided MWA to complex scenarios (e.g., tumors involving the skin or NAC) and patients ineligible for surgery can enhance its clinical value. To improve global accessibility, future efforts should focus on developing more portable and affordable systems, particularly for resource-limited settings. Coupled with the implementation of tele-proctoring and simulation-based training platforms to disseminate expertise efficiently,⁵⁷ these strategies can help lower economic and training barriers. Nonetheless, comprehensive health-economic analyses and large-scale cost-effectiveness studies are urgently needed to evaluate and quantify these economic benefits formally against the current standard of care.

US-guided MWA represents a promising minimally invasive modality for treating both benign and malignant breast tumors. Supported by robust clinical evidence, US-guided MWA demonstrates high rates of complete ablation, considerable tumor volume reduction, and favorable cosmetic outcomes, with a low incidence of complications. For benign lesions, such as fibroadenomas, it offers a viable non-surgical alternative that avoids general anesthesia, minimizes scarring, and facilitates rapid recovery. For early-stage breast cancer, preliminary studies

indicate that US-guided MWA can achieve local tumor control comparable with surgical resection in the short- to mid-term, while offering notable clinical benefits, such as shorter procedure times, reduced hospitalization, and enhanced patient satisfaction.

Beyond its local ablative effects, US-guided MWA stimulates systemic antitumor immune responses, including Th1 polarization and T-cell activation, which may synergize with immunotherapy and other systemic treatments, opening new avenues for multimodal cancer management. However, the adoption of US-guided MWA is not without challenges. Technical limitations related to tumor proximity to critical structures, intraprocedural monitoring difficulties, and the lack of long-term oncological data from randomized trials necessitate cautious patient selection and procedural planning. Future advancements in antenna design, real-time imaging integration, AI-assisted treatment planning, and combination immunotherapies hold great potential to enhance the precision, efficacy, and accessibility of US-guided MWA. Standardized international guidelines and large-scale prospective studies are urgently needed to validate long-term outcomes and establish US-guided MWA as a mainstream option in breast tumor management.

In summary, US-guided MWA represents a major advancement in the shift toward minimally invasive, organ-preserving therapies. With ongoing technological refinement and clinical validation, it is poised to meet the demand for personalized, minimally invasive care in breast oncology, particularly for patients seeking alternatives to traditional surgery.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

This study was supported by Special Technology Project for High-Quality Development of Nanjing Lishui People's Hospital (No. LWG202404); Medical Education Collaborative Innovation Fund of Jiangsu University (No. JDYY2023101); Research Project on Hospital Management Innovation of the Jiangsu Provincial Hospital Association (No. JSYGY-3-2024-193).

References

- Łukaszewicz S, Czezelewski M, Forma A, Baj J, Sitarz R, Stanisławek A. Breast cancer-epidemiology, risk factors, classification, prognostic markers, and current treatment strategies-an updated review. *Cancers (Basel)*. 2021;13(17):4287. [\[Crossref\]](#)
- Li L, Deng H, Ye X, Li Y, Wang J. Comparison of the diagnostic efficacy of mathematical models in distinguishing ultrasound imaging of breast nodules. *Sci Rep*. 2023;13(1):16047. [\[Crossref\]](#)
- Veronesi U, Cascinelli N, Mariani L, et al. Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. *N Engl J Med*. 2002;347(16):1227-1232. [\[Crossref\]](#)
- Mactier M, McIntosh SA, Sharma N. Minimally invasive treatment of early, good prognosis breast cancer-is this feasible? *Br J Radiol*. 2024;97(1157):886-893. [\[Crossref\]](#)
- van de Voort EMF, Struik GM, Birnie E, Moelker A, Verhoef C, Klem TMAL. Thermal ablation as an alternative for surgical resection of small (≤ 2 cm) breast cancers: a meta-analysis. *Clin Breast Cancer*. 2021;21(6):e715-e730. [\[Crossref\]](#)
- Xu C, Yu Q, Wang M, et al. Efficacy and safety of microwave ablation for benign breast lesions: a systematic review and meta-analysis. *Wideochir Inne Tech Maloinwazyjne*. 2022;17(3):418-429. [\[Crossref\]](#)
- CRILE G Jr. Selective destruction of cancers after exposure to heat. *Ann Surg*. 1962;156(3):404-407. [\[Crossref\]](#)
- Radmilović-Radjenović M, Bošković N, Radjenović B. Computational modeling of microwave tumor ablation. *Bioengineering (Basel)*. 2022;9(11):656. [\[Crossref\]](#)
- Dong F, Wu Y, Li W, et al. Advancements in microwave ablation for tumor treatment and future directions. *iScience*. 2025;28(4):112175. [\[Crossref\]](#)
- Ke J, Cheng S, Yang T, Zhang C, Wang M, Zhang Y. Feasibility of controlling metastatic osseous pain using three kinds of image-guided procedures for thermal microwave ablation: a retrospective study. *Orthop Surg*. 2021;13(1):116-125. [\[Crossref\]](#)
- Li R, Yu Y. Clinical study on ultrasound-guided percutaneous microwave ablation for hepatocellular carcinoma. *Am J Transl Res*. 2025;17(5):3640-3648. [\[Crossref\]](#)
- Tang Y, Sun X, Tian J, et al. Clinical efficacy analysis of ultrasound-guided microwave ablation for nodular goiters. *Am J Transl Res*. 2024;16(9):4526-4533. [\[Crossref\]](#)
- Jonsdottir G, Beermann M, Lanz E, et al. Ultrasound guided microwave ablation treatment of uterine fibroids: clinical response and patient acceptability. *Acta Obstet Gynecol Scand*. 2025;104(2):350-356. [\[Crossref\]](#)
- Cheng Z, Yu X, Han Z, et al. Percutaneous microwave ablation under ultrasound guidance for renal cell carcinomas at clinical staging T1 in patients aged 65 years and older: a comparative study. *J Cancer Res Ther*. 2022;18:509-515. [\[Crossref\]](#)
- Zhang J, Li H, Lin L, Lu J, Wang H. Ultrasound-guided microwave ablation for multiple benign breast lesions: a prospective study. *J Obstet Gynaecol Res*. 2021;47(9):3362-3369. [\[Crossref\]](#)
- Liu G, Zhang Y, Hu E, et al. Feasibility and efficacy of microwave ablation for treating breast fibroadenoma. *Int J Hyperthermia*. 2021;38(1):471-478. [\[Crossref\]](#)
- Zhu J, Yu M, Chen L, et al. Enhanced antitumor efficacy through microwave ablation in combination with immune checkpoints blockade in breast cancer: a pre-clinical study in a murine model. *Diagn Interv Imaging*. 2018;99(3):135-142. [\[Crossref\]](#)
- Yu M, Wang B, Qu Y, et al. Comparison of local effects and systemic T-cell responses in patients with breast cancer treated by radiofrequency ablation versus microwave ablation. *Cancer Cell Int*. 2025;25(1):261. [\[Crossref\]](#)
- Vogl TJ, Nour-Eldin NA, Albrecht MH, et al. Thermal ablation of lung tumors: focus on microwave ablation. *Rofo*. 2017;189(9):828-843. [\[Crossref\]](#)
- Lubner MG, Brace CL, Hinshaw JL, Lee FT Jr. Microwave tumor ablation: mechanism of action, clinical results, and devices. *J Vasc Interv Radiol*. 2010;21(8 Suppl):S192-203. [\[Crossref\]](#)
- Moore C, Salas N, Zaia J, Shields J, Bird V, Leveille R. Effects of microwave ablation of the kidney. *J Endourol*. 2010;24(3):439-444. [\[Crossref\]](#)
- Huang H, Zhang L, Moser MAJ, Zhang W, Zhang B. A review of antenna designs for percutaneous microwave ablation. *Phys Med*. 2021;84:254-264. [\[Crossref\]](#)
- Pfannenstiel A, Iannuccilli J, Cornelis FH, Dupuy DE, Beard WL, Prakash P. Shaping the future of microwave tumor ablation: a new direction in precision and control of device performance. *Int J Hyperthermia*. 2022;39(1):664-674. [\[Crossref\]](#)
- Rahpeima R, Soltani M, Moradi Kashkooli F. Numerical study of microwave induced thermoacoustic imaging for initial detection of cancer of breast on anatomically realistic breast phantom. *Comput Methods Programs Biomed*. 2020;196:105606. [\[Crossref\]](#)
- Simo KA, Tsirlina VB, Sindram D, et al. Microwave ablation using 915-MHz and 2.45-GHz systems: what are the differences? *HPB (Oxford)*. 2013;15(12):991-996. [\[Crossref\]](#)
- Park WK, Maxwell AW, Frank VE, et al. Evaluation of a novel thermal accelerant for augmentation of microwave energy during image-guided tumor ablation. *Theranostics*. 2017;7(4):1026-1035. [\[Crossref\]](#)
- Zhou W, Yu M, Mao X, et al. Landscape of the peripheral immune response induced by local microwave ablation in patients with breast cancer. *Adv Sci (Weinh)*. 2022;9(17):e2200033. [\[Crossref\]](#)
- Zhang L, Zhang M, Wang J, et al. Immunogenic change after percutaneous microwave ablation in pulmonary malignancies: variation in immune cell subsets and cytokines in peripheral blood. *Front Immunol*. 2022;13:1069192. [\[Crossref\]](#)
- Paepke S, Metz S, Brea Salvago A, Ohlinger R. Benign breast tumours - diagnosis and management. *Breast Care (Basel)*. 2018;13(6):403-412. [\[Crossref\]](#)
- Pleasant V. Benign breast disease. *Clin Obstet Gynecol*. 2022;65(3):448-460. [\[Crossref\]](#)
- Ahire P, Nagwani P, Waghmare R, Gandhi A. Unraveling benign breast conditions: a comprehensive study of diagnosis, treatment, and care. *Cureus*. 2025;17(6):e85183. [\[Crossref\]](#)
- Zhou W, Wang R, Liu X, et al. Ultrasound-guided microwave ablation: a promising tool in management of benign breast tumours. *Int J Hyperthermia*. 2017;33(3):263-270. [\[Crossref\]](#)
- Zhang W, Li JM, He W, et al. Ultrasound-guided percutaneous microwave ablation for benign breast lesions: evaluated by contrast-enhanced ultrasound combined with magnetic resonance imaging. *J Thorac Dis*. 2017;9(11):4767-4773. [\[Crossref\]](#)
- Cui R, Wu H, Xu J, et al. Volume reduction for ≥ 2 cm benign breast lesions after ultrasound-guided microwave ablation with a minimum 12-month follow-up. *Int J Hyperthermia*. 2021;38(1):341-348. [\[Crossref\]](#)
- Zhang S, Yang J, Wang L, Li J, He C, Lu M. Comparison of ultrasound-guided microwave ablation and ultrasound-guided vacuum-assisted excision for treating breast lesions of uncertain malignant potential (B3 Lesions): a retrospective study. *J Obstet Gynaecol Res*. 2023;49(4):1283-1290. [\[Crossref\]](#)
- Lei L, Qing SH, Wei L, et al. A comparative analysis of the clinical effect of multiple treatments for benign breast tumors. *J Int Med Res*. 2025;53(8):3000605251367623. [\[Crossref\]](#)
- Zhang S, Wang L, Yang J, Lu M. Ultrasound-guided microwave ablation for giant breast leiomyoma: a case report. *Front Oncol*. 2023;13:1095891. [\[Crossref\]](#)
- Sharma DK, Saripilli R. Recent strategies in diagnosis, screening, prevention, and treatment of breast cancer in young women. *Discov Oncol*. 2025;16(1):1532. [\[Crossref\]](#)
- Anton SC, Lazan A, Grigore M, et al. Current trends in breast cancer genetics, risk factors, and screening strategies. *J Biol Methods*. 2025;12(2):e99010054. [\[Crossref\]](#)

40. van den Broek AJ, Schmidt MK, van 't Veer LJ, et al. Prognostic impact of breast-conserving therapy versus mastectomy of Brca1/2 mutation carriers compared with noncarriers in a consecutive series of young breast cancer patients. *Ann Surg.* 2019;270(2):364-372. [\[Crossref\]](#)
41. Yu P, Tang H, Zou Y, et al. Breast-conserving therapy versus mastectomy in young breast cancer patients concerning molecular subtypes: a seer population-based study. *Cancer Control.* 2020;27(1):1073274820976667. [\[Crossref\]](#)
42. Luo Y, Chen X, Lv R, et al. Breast-conserving surgery versus modified radical mastectomy in T1-2n3m0 stage breast cancer: a propensity score matching analysis. *Breast Cancer.* 2024;31(5):979-987. [\[Crossref\]](#)
43. De la Cruz Ku G, Karamchandani M, Chambergo-Michilot D, et al. Does Breast-Conserving Surgery with radiotherapy have a better survival than mastectomy? A meta-analysis of more than 1,500,000 patients. *Ann Surg Oncol.* 2022;29(10):6163-6188. [\[Crossref\]](#)
44. Zhou W, Zha X, Liu X, et al. Us-guided percutaneous microwave coagulation of small breast cancers: a clinical study. *Radiology.* 2012;263(2):364-373. [\[Crossref\]](#)
45. Yu J, Han ZY, Li T, et al. Microwave ablation versus nipple sparing mastectomy for breast cancer ≤5 cm: a pilot cohort study. *Front Oncol.* 2020;10:546883. [\[Crossref\]](#)
46. Zhong Z, Ling L, Tao J, et al. Percutaneous microwave ablation combined with endocrine therapy versus standard therapy for elderly patients with Hr-positive and Her2-negative invasive breast cancer: a propensity score-matched analysis of a multi-center, prospective cohort study. *Gland Surg.* 2023;12(7):940-951. [\[Crossref\]](#)
47. Dai YQ, Liang P, Wang J, et al. Microwave ablation without subsequent lumpectomy versus breast-conserving surgery for early breast cancer: a propensity score matching study. *Int J Hyperthermia.* 2023;40(1):2186325. [\[Crossref\]](#)
48. Dai Y, Jiang J, Liang P, et al. Percutaneous microwave ablation: a viable local therapy for breast cancer involving the skin/nipple-areola complex? *Curr Probl Surg.* 2024;61(6):101483. [\[Crossref\]](#)
49. Zhou W, Yu M, Pan H, et al. Microwave ablation induces Th1-type immune response with activation of icos pathway in early-stage breast cancer. *J Immunother Cancer.* 2021;9(4):e002343. [\[Crossref\]](#)
50. Pan H, Yu M, Tang X, et al. Preoperative single-dose camrelizumab and/or microwave ablation in women with early-stage breast cancer: a window-of-opportunity trial. *Med.* 2024;5(4):291-310.e295. [\[Crossref\]](#)
51. Varghese T, Techavipoo U, Zagzebski JA, Lee FT Jr. Impact of gas bubbles generated during interstitial ablation on elastographic depiction of in vitro thermal lesions. *J Ultrasound Med.* 2004;23(4):535-544; quiz 545-546. [\[Crossref\]](#)
52. Shi L, Wang J, Ding N, et al. Inflammation induced by incomplete radiofrequency ablation accelerates tumor progression and hinders Pd-1 immunotherapy. *Nat Commun.* 2019;10(1):5421. [\[Crossref\]](#)
53. Tran HH, Thu A, Twayana AR, et al. Artificial intelligence in cardiac electrophysiology: enhancing mapping and ablation precision. *Cardiol Rev.* 2025. Epub ahead of print. [\[Crossref\]](#)
54. Guo Y, Li N, Song C, Yang J, Quan Y, Zhang H. Artificial intelligence-based automated breast ultrasound radiomics for breast tumor diagnosis and treatment: a narrative review. *Front Oncol.* 2025;15:1578991. [\[Crossref\]](#)
55. Ma QP, He XL, Li K, et al. Dynamic contrast-enhanced ultrasound radiomics for hepatocellular carcinoma recurrence prediction after thermal ablation. *Mol Imaging Biol.* 2021;23(4):572-585. [\[Crossref\]](#)
56. Xu J, Cheng X, Tan L, et al. Microwave responsive nanoplatfrom via p-selectin mediated drug delivery for treatment of hepatocellular carcinoma with distant metastasis. *Nano Lett.* 2019 ;19(5):2914-2927. [\[Crossref\]](#)
57. Takemasa I, Okuya K, Okita K, et al. Tele-proctoring for minimally invasive surgery across japan: an initial step toward a new approach to improving the disparity of surgical care and supporting surgical education. *Ann Gastroenterol Surg.* 2024;8(2):356-364. [\[Crossref\]](#)



©Copyright 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Establishment of an inferior vena cava filter retrieval program: the effect on trauma and non-trauma patient populations

 Katelyn Gill¹
 Hector Ferral^{2,3}
 Alexandra H. Fairchild^{2,3}
 Bahri Üstünsöz^{2,3}
 Dan F. Laney IV^{2,3}
 John Hunt^{3,4}
 Jonathan E. Schoen^{3,4}
 Alan Marr^{3,4}
 Patrick Greiffenstein^{3,4}
 Lance Stuke^{3,4}
 Alison A. Smith^{3,4}

¹Louisiana State University Health Sciences Center,
School of Medicine, New Orleans, USA

²Louisiana State University Health Sciences Center,
Department of Radiology, New Orleans, USA

³University Medical Center, New Orleans, USA

⁴Louisiana State University Health Sciences Center,
Department of Surgery, New Orleans, USA

PURPOSE

This study aimed to compare the effect of establishing an inferior vena cava filter (IVCF) retrieval program (IVCFRP) on the IVCF retrieval rates in trauma and non-trauma patients.

METHODS

This was an institutional review board-approved retrospective study. IVCF retrieval in trauma and non-trauma patients was compared before and after the establishment of an IVCFRP in a single Level I Trauma Center. The IVCFRP at our institution was established in April 2020. All patients who underwent IVCF placement between January 2016 and March 2020 were compared with patients who had an IVCF placed between April 2020 and June 2024. A medical record review included the collection of demographic information, indication for IVCF insertion and retrieval, date of IVCF insertion and retrieval, IVCF retrieval rate, clinical outcomes, and complications. The patients were stratified into trauma and non-trauma groups. Univariate analyses were performed with a *P* value of <0.05 considered statistically significant.

RESULTS

A total of 164 patients underwent IVCF placement between January 2016 and June 2024. Fifty-two IVCFs were implanted before and 112 after the establishment of the IVCFRP. The overall rate of IVCF retrieval was significantly higher following the establishment of an IVCFRP (33.3% vs. 51%, *P* = 0.047). In non-trauma patients, the retrieval rate was significantly higher after the establishment of an IVCFRP (37.5% vs. 61.3%, *P* = 0.03). The retrieval rate in trauma patients (22.2% vs. 21.4%) was not significantly changed by the establishment of an IVCFRP. Clinical outcomes and complications were similar between groups.

CONCLUSION

IVCF retrieval rates significantly improved after the establishment of an IVCFRP. This increase in IVCF retrieval rate was driven by an increase in filter retrievals in the non-trauma patient population. The rate of IVCF retrieval in trauma patients was not affected by the implementation of an IVCFRP. Special considerations and changes in practice may need to be established to improve IVCF retrieval rates in trauma patients.

CLINICAL SIGNIFICANCE

Implementation of a structured IVCFRP significantly improved retrieval rates in non-trauma patients but did not yield similar results in trauma patients, highlighting the need for alternative strategies in this population.

KEYWORDS

Inferior vena cava filter, trauma, retrieval program, complex retrieval, inferior vena cava filter complication

Handling editor: Barbaros Çil

Corresponding author: Katelyn Gill

E-mail: kgill1@lsuhsc.edu

Received 25 June 2025; revision requested 01 August 2025; accepted 01 September 2025.



Epub: 29.09.2025

Publication date: 01.09.2026

DOI: 10.4274/dir.2025.253409

You may cite this article as: Gill K, Ferral H, Fairchild AH, et al. Establishment of an inferior vena cava filter retrieval program: the effect on trauma and non-trauma patient populations. *Diagn Interv Radiol.* 2026;32(5):692-699.

Venous thromboembolism (VTE), which comprises deep vein thrombosis (DVT) and pulmonary embolism (PE), represents an important morbidity in hospitalized patients.¹ PE secondary to DVT carries a high risk of mortality and remains the third leading cause of death in trauma patients who survive the first day of hospital admission.² Inferior vena cava filters (IVCFs) are indicated in patients with VTE or PE who have contraindications to anticoagulation.³ Trauma patients are at high risk of developing VTE and often have contraindications to anticoagulation secondary to injuries and bleeding complications.^{3,4}

Current IVCF guidelines recommend removal of retrievable IVCFs once the filters are no longer indicated.⁵ The implementation of structured IVCF follow-up programs has been proven to increase filter retrieval rates and decrease the number of patients lost to follow-up;⁶⁻⁸ however, this practice concept has not been validated in trauma patients.^{9,10} The working hypothesis of this retrospective study was that creation of an IVCF database and establishment of an IVCF retrieval program (IVCFRP) would increase IVCF retrieval rates both in non-trauma and trauma patients. The purpose of this study was to compare IVCF retrieval rates in trauma and non-trauma patients before and after establishing an IVCFRP at a single Level I Trauma Center in the United States.

Methods

Patient population

This is a retrospective, institutional review board-approved, Health Insurance Portability and Accountability Act of 1996-compliant study with a waiver of consent. The study group included patients who underwent IVCF insertion and retrieval by interventional radiology (IR) at a single Level I Trauma Center before and after implementation of an IVCFRP. The IVCFRP was established in April

2020 (Figure 1). This retrospective study was approved by the Louisiana State University Health Sciences Center New Orleans Institutional Review Board (approval: IRB #4222, date: May 29, 2023). The inclusion criteria for this study were as follows: patients aged ≥ 18 years with IVCFs placed at the University Medical Center New Orleans by the IR service between January 2016 and June 2024. The exclusion criteria were as follows: patients aged < 18 years, pregnant women, patients with IVCF not inserted at the institution, or prisoners. Patients who underwent IVCF insertion by other services (e.g., cardiology or vascular surgery) were excluded, as only patients with IVCF placed by the IR service were entered in the database and were part of the IVCFRP. Patients who had IVCF placement at outside institutions and were referred to the center for IVCF retrieval were identified but not included in the analysis. One hundred ninety-six patients who met the study criteria were included in the medical record review.

The electronic records of patients who had IVCF insertion between January 2016 and March 2020 were identified using the Slicer-Dicer program in the EPIC (version 11.3.0.11, Verona, WI, USA) software platform. The records of the patients who underwent IVCF placement after April 2020 were identified through the IVCFRP database. The collected clinical variables included the following: indication for IVCF insertion and removal, IVCF retrieval rates, IVCF dwell time, IVCF complications [tilting, fractures, filter migration, filter leg penetration through the inferior vena cava (IVC) wall, filter thrombosis]. Other

variables included the following: in-hospital mortality, 6-month mortality, hospital length of stay (HLoS), and intensive care unit length of stay (ICULoS).

Filter tilt was defined as the deviation of the cephalon-caudal axis of the IVCF by > 30 degrees compared with the IVC axis. Inferior vena cava thrombosis was defined as direct visualization of thrombus during the retrieval cavagram. Inferior vena cava occlusion was defined as the inability to pass a wire caudal to the IVCF during the retrieval procedure. Protrusion of struts was defined as perforation > 5 mm past the IVC wall or extension into surrounding structures on computed tomography.

Inferior vena cava filter retrieval program

At the University Medical Center New Orleans, IVCFs are placed by fellowship-trained interventional radiologists in accordance with current guidelines.⁵ Following the implementation of IVCFRP in April 2020, details of patients who undergo IVCF insertion at the institution are entered into a database, and the patients are enrolled in a structured IVCF follow-up program. Details of IVCFs inserted by other services are not entered in the IR IVCF database and are not included in the IR IVCF surveillance and follow-up program. At the time of IVCF insertion, patients are scheduled for a visit to the IR clinic to discuss eligibility for filter retrieval. The clinic visit for IVCF removal evaluation is scheduled 8-12 weeks after the date of the IVCF insertion. Patients eligible for filter retrieval are scheduled for the procedure. Patients who

Main points

- Timely inferior vena cava filter (IVCF) retrieval is necessary to prevent filter complications associated with prolonged dwell time.
- IVCF retrieval programs (IVCFRPs) have been shown to increase IVCF retrieval rates, though this intervention did not significantly increase filter retrieval in trauma patients.
- The implementation of IVCFRPs facilitates referrals for complex filter removals secondary to prolonged filter dwell time.

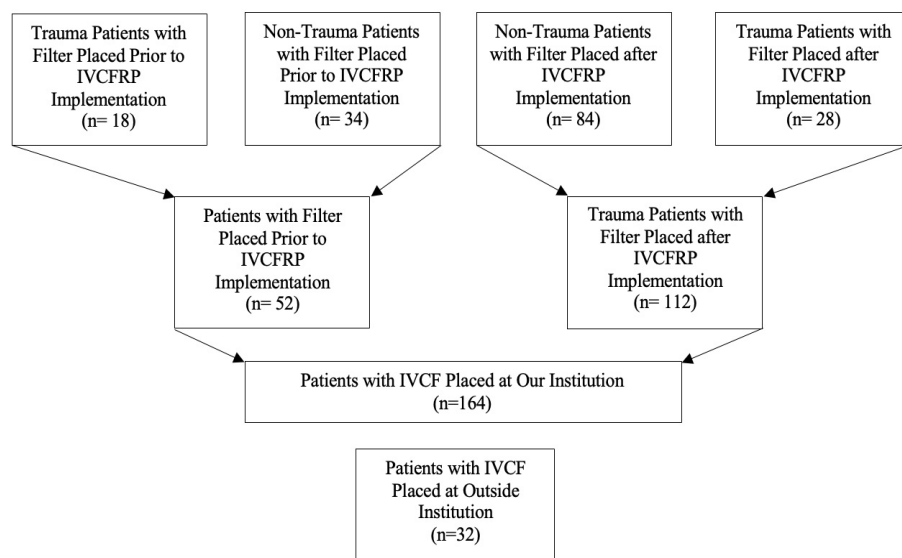


Figure 1. Flowchart of patient inclusion. IVCF, inferior vena cava filter; IVCFRP, inferior vena cava filter retrieval program.

are evaluated and considered ineligible for retrieval enter active surveillance and are scheduled for an IR clinic visit with a kidney, ureter, and bladder X-ray every 6 months to assess for filter complications and reevaluate retrieval eligibility. Patients who do not show up to IR clinic appointments are contacted by phone by the IR service navigators to reschedule a visit. If contact with the patient cannot be established, or the patient does not show up for additional appointments, they are deemed lost to follow-up.

Statistical analysis

Patients were stratified into groups based on their time of presentation (before vs. after IVCFRP implementation) and clinical presentation (trauma vs. non-trauma). In total, four groups were compared: (1) trauma patients before program implementation, (2) non-trauma patients before program implementation, (3) trauma patients after program implementation, and (4) non-trauma patients after program implementation. Patients with IVCF placed between January 2016 and March 2020 were stratified into before-program implementation groups. Patients with IVCF placed between April 2020 and June 2024 were stratified into after-program implementation groups. The IVCF retrieval rates were adjusted for filter retrieval eligibility and intent to retrieve. Patients with permanent filters or filters placed within 6 months of study initiation and patients who died within 6 months of implantation were excluded from the IVCF retrieval analysis. Univariate analysis was performed using Fisher's exact test for categorical variables or Student's t-test for continuous variables. Data were analyzed using IBM SPSS (version 27, Armonk, NY, USA) and GraphPad Prism (version 10.0.0, La Jolla, CA, USA). A *P* value of < 0.05 was considered statistically significant.

Results

Inferior vena cava filter indications and complications

One hundred ninety-six patients were included in the study, of whom 164 had an IVCF placed at the institution. The indications for IVCF insertion were as follows: (1) contraindication to anticoagulation due to active bleeding (*n* = 43, 26.2%), upcoming or recent surgery (*n* = 30, 18.3%), bleeding disorders (*n* = 6, 3.7%), unspecified contraindications to anticoagulation (*n* = 12, 7.3%); (2) massive PE with cor pulmonale and residual DVT (*n* = 29, 17.7%); (3) failure of anticoagulation (*n* = 3, 1.8%); (4) prophylactic with no concurrent

DVT or PE (*n* = 4, 2.4%); and (5) non-specified (*n* = 55, 33.5%). A total of 14 IVCF complications were identified in patients with IVCFs placed at the institution. Identified complications included IVC thrombosis (*n* = 6, 3.7%), protrusion of struts through the IVC (*n* = 5, 3%), filter tilting (*n* = 2, 1.2%), and PE after IVCF insertion (*n* = 1, 0.6%) (Table 1).

Inferior vena cava filter retrieval before and after program implementation

Before IVCFRP implementation, 52 patients underwent IVCF insertion, and 112 patients had an IVCF placed after implementation of the program. Baseline demographic [age, gender, body mass index (BMI)] and clinical outcomes (mortality, HLoS, ICULoS) were similar between groups (*P* > 0.05). The percentage of patients seen in the IR clinic for filter retrieval significantly increased from before-program implementation to after-program implementation (40.4% vs. 58.9%, *P* = 0.03). Eight (9.5%) patients had a permanent filter placed.

Insertion and retrieval procedure lengths and fluoroscopy times were similar between groups (*P* > 0.05). Patients who had their filter placed after IVCFRP implementation had a significantly longer filter dwell time (142.9 ± 133.4 days) than patients who had their IVCF placed before program implementation (50.1 ± 83.4 days) (*P* = 0.002). The IVCF retrieval rates were adjusted for retrieval eligibility and intent to retrieve. In the group who had filters placed before IVCFRP implementation, 4 (7.7%) patients died within 6 months of filter implantation, leaving 48 patients eligible for filter retrieval. In the group who had filters placed after IVCFRP implementation, 19 (16.9%) of patients died within 6 months, and 8 (7.1%) patients had permanent filters placed, leaving 85 patients eligible for IVCF retrieval. When adjusted for IVCF retrieval eligibility, patients who had IVCFs placed after program implementation had a significantly higher rate of IVCF retrieval (44/85, 51.8%) than patients who had IVCFs placed before program implementation (16/48, 33%) (*P* = 0.047). There was no significant difference in IVCF complications between groups (*P* > 0.05) (Table 2).

Retrieval of inferior vena cava filters in trauma patients

Forty-six trauma patients had an IVCF placed at the institution. Eighteen patients (39.1%) had their filter placed before IVCFRP implementation, and 28 (60.9%) after. The overall average HLoS for trauma patients

who had an IVCF implanted was 30.6 ± 23.9 days. Fifty percent (23/46) of trauma patients with IVCFs placed had an HLoS >21 days, and 37% (17/46) had an HLoS >30 days. Baseline demographics (age, gender, BMI), clinical outcomes (mortality, HLoS, ICULoS), and IVCF complications were similar between groups (*P* > 0.05). The percentage of trauma patients seen in the IR clinic for IVCF retrieval evaluation was similar before and after program implementation (38.8% vs. 46.4%, *P* = 0.76). Two patients who had an IVCF placed before the program implementation died within 6 months of IVCF insertion, leaving 16 patients eligible for filter retrieval. Five patients who had an IVCF placed after the program implementation died within 6 months of IVCF insertion, leaving 23 patients eligible for IVCF retrieval. There was no significant difference in the rate of IVCF retrieval in eligible trauma patients who had IVCF placement before IVCFRP implementation (4/16, 25%) or after program implementation (6/23, 26.1%) (*P* > 0.05). The mean IVCF dwell time was 127.8 ± 159.3 days for patients with IVCFs placed before the program initiation and 85.4 ± 46.4 days for patients after the program initiation. Two patients (4.3%) had a failed IVCF retrieval attempt due to complete IVC thrombosis (Table 3).

Retrieval of inferior vena cava filters in non-trauma patients

One hundred eighteen non-trauma patients had an IVCF placed at the institution; 34 patients (28.8%) had an IVCF placed before and 84 patients (71.2%) after the program implementation. There was no significant difference in baseline demographics, clinical outcomes (mortality, HLoS, ICULoS), or IVCF complications between groups (*P* > 0.05). The percentage of patients seen in the IR clinic for IVCF retrieval evaluation significantly increased from before-IVCFRP to after-IVCFRP implementation (41.2% vs. 63.1%, *P* = 0.04). There was no significant difference in IVCF insertion or retrieval procedure length or fluoroscopy time between groups (*P* > 0.05). The mean filter dwell time for non-trauma patients who had an IVCF placed after program implantation (142.9 ± 133.4 days) was significantly longer than those who had an IVCF placed before program implementation (50.1 ± 83.4 days) (*P* = 0.0003). Two non-trauma patients who had an IVCF implanted before the program implementation died within 6 months of the procedure, leaving 32 patients eligible for filter retrieval. Fourteen patients who had an IVCF placed after the retrieval program implementation

Table 1. Indications for inferior vena cava (IVC) filter (IVCF) placement at our institution; IVCF filter complications; IVCF placed at outside institutions and referred for filter retrieval	
Indications for IVCF implantation, n (%)	Filters placed at the institution (2016–2024) n = 164
Contraindication to anticoagulation in the setting of current deep vein thrombosis (DVT) and/or pulmonary embolism (PE)	91 (55.5)
Active bleeding	43 (26.2)
Gastrointestinal	18 (11.0)
Brain	12 (7.3)
Trauma site	6 (3.7)
Hematuria	2 (1.2)
Other	5 (3.0)
Upcoming/recent surgery	30 (18.3)
Bleeding disorder	6 (3.7)
Hemophilia	2 (1.2)
Anemia	2 (1.2)
Thrombocytopenia	1 (0.6)
Non-adherence to anticoagulants	1 (0.6)
Other/non-specified contraindication	12 (7.3)
Massive PE + cor pulmonale with residual DVT	29 (17.7)
Failure of anticoagulation	3 (1.8)
Prophylactic (no current DVT and/or PE)	4 (2.4)
Polytrauma	2 (1.2)
Prior to surgery	2 (1.2)
Other/unknown	37 (22.6)
IVCF complications, n (%)	
Total	14 (8.5)
Thrombosis	6 (3.7)
Protrusion of struts through IVC	5 (3.0)
Tilting	2 (1.2)
PE after IVCF insertion	1 (0.6)
Patients referred for IVCF retrieval, n (%)	Filters placed at outside institutions n = 32
Filters retrieved	17 (53.1)
Failed retrieval	2 (6.3)
Average dwell time, mean days (SD)	2,928.4 (2,915.5)
Retrieval not attempted	13 (40.6)
Never seen in clinic	5 (15.6)
Patient refused retrieval	3 (9.4)
No show for procedure	2 (6.3)
Deemed not eligible for retrieval	2 (6.3)
Filter deemed stable	1 (3.1)
IVCF complication	41
Protrusion of struts through IVC wall	10 (31.2)
Extension of struts into other structures	6 (18.8)
Duodenum	4 (12.5)
Other	4 (12.5)
Filter tilt with tip embedded in lumen	6 (18.8)
Filter fracture	6 (18.8)
	5 (15.6)
	1 (3.1)
	1 (3.1)
Filters with multiple complications	8 (25)

died within 6 months, and 8 patients had a permanent IVCF placed, leaving 62 patients eligible for IVCF retrieval. When adjusted for filter retrieval eligibility, non-trauma patients who had an IVCF placed after retrieval program implementation had a significantly higher rate of IVCF retrieval (38/62, 61.3%) than patients who had an IVCF placed before the program implementation (12/32, 37.5%)

($P = 0.03$). There was no significant difference in IVCF complications between groups ($P > 0.05$) (Table 4).

Filters placed at outside institutions

Thirty-two patients had their filters placed at an outside institution and were referred for IVCF retrieval. A total of 17 IVCF filters placed at outside institutions were retrieved at the

University Medical Center New Orleans after the implementation of an IVCFRP. One patient had an IVCF removed surgically due to erosion of the filter into the right renal vein; the remaining 16 patients had their IVCFs removed using advanced endovascular techniques. Removal of the IVCF was unsuccessful in 2 (6.3%) patients. One patient had a failed IVCF retrieval attempt at another institution.

	Before (n = 52)	After (n = 112)	P value
Demographics			
Age, mean years (SD)	53.4 (16.2)	56.1 (16.6)	0.33
Male, n (%)	35 (67.3)	63 (56.3)	0.23
Body mass index, mean (SD)	30.7 (11.5)	29 (9.3)	0.31
Clinical outcomes			
In-hospital mortality, n (%)	4 (7.7)	16 (14.3)	0.31
Mortality within 6 months, n (%)	4 (7.7)	19 (16.9)	0.15
Hospital length of stay (LoS), mean days (SD)	19.2 (16.5)	27.6 (39.7)	0.14
Intensive care unit LoS, mean days (SD)	7.6 (12.6)	11.4 (24.2)	0.29
Interventional radiology clinic follow-up, n (%)	21 (40.4)	66 (58.9)	0.03
Procedural information			
Insertion procedure length, mean mins (SD)	24.9 (10.1)	28.9 (18.3)	0.14
Insertion fluoroscopy, mean mins (SD)	2.2 (1.9)	3.5 (5)	0.07
Permanent filters placed, n (%)	0	8 (7.1)	0.06
Time from insertion to removal, mean days (SD)	69.5 (106.9)	131.5 (124.2)	0.002
Removal procedure length, mean mins (SD)	29.6 (15.1)	28.1 (7.9)	0.4
Removal fluoroscopy, mean mins (SD)	5.4 (6.5)	4.8 (1.6)	0.36
IVCF retrieval, n (%)	16 (30.8)	44 (39.3)	0.27
Adjusted IVCF retrieval, n (%)	16 (33.3)	44 (51.8)	0.047
Failure of IVCF retrieval, n (%)	0	2 (1.8)	1
IVCF complications, n (%)			
Total	3 (5.8)	11 (9.8)	0.55
Thrombosis	2 (3.8)	4 (3.6)	1
Protrusion of struts through IVC	1 (1.9)	4 (3.6)	1
Tilting	0	2 (1.8)	1
Pulmonary embolism after IVCF insertion	0	1 (0.9)	1

SD, standard deviation.

One patient was discovered to have a fractured filter that was embedded in a chronic clot that could not be removed, and the procedure was aborted. The average dwell time of IVCFs placed at outside institutions was $2,928.4 \pm 2,915.5$ days (8 years). IVCF retrieval was not attempted in 13 (40.6%) patients referred for retrieval for the following reasons: 5 (15.6%) patients did not show up for the initial appointment and were never evaluated in clinic, 3 (9.4%) patients refused IVCF retrieval, 2 (6.3%) patients did not show up for a scheduled retrieval, 2 (6.3%) patients were deemed not eligible for retrieval, and 1 (3.1%) filter was deemed stable at the time of evaluation and the risk of filter removal was deemed to be higher than the clinical benefit.

A total of 41 IVCF complications were identified in patients with IVCFs placed at outside institutions. Eight (25%) patients had multiple filter complications identified. Complications included protrusion of struts

through the IVC wall (n = 10, 31.2%), extension of struts into other structures (n = 6, 18.8%), filter tilt with the tip embedded in the IVC lumen (n = 6, 18.8%), filter fracture (n = 6, 18.8%), IVC thrombosis (n = 5, 15.6%), filter migration to the right common iliac vein (n = 1, 3.1%), and filter collapse within the IVC (n = 1, 3.1%). Surrounding structures affected by protruded struts included the duodenum, right renal vein, right spermatic artery, L3 lumbar artery, and lumbar vertebral body (Table 1).

Discussion

An IVCFRP was implemented at the institution in April of 2020. The goal of this study was to assess the effect of an IVCFRP on the retrieval rate in trauma and non-trauma patients. We hypothesized that the establishment of an IVCFRP would increase IVCF retrieval rates in both trauma and non-trauma patients. This study showed that implementation of the IVCFRP was effective

in the non-trauma population, increasing the rate of follow-up and IVCF retrieval. The percentage of non-trauma patients seen in the IR clinic for IVCF retrieval evaluation increased significantly from 41.2% to 63.1% after the IVCFRP implementation. The adjusted IVCF retrieval rates increased significantly from 37.5% to 61.3%. Nationally, the rate of IVCF retrieval is estimated to be as low as 12%–18%.¹¹ With the implementation of structured follow-up programs, the rate of IVCF retrieval has been noted to increase to 28.3%–72.9%, with retrieval rates as high as 92.5% observed at select institutions.^{6,7,12,13} This increase in retrieval rates after the establishment of a dedicated retrieval program is consistent with the findings of our study.

This study showed that the establishment of an IVCFRP did not increase the filter retrieval rates in trauma patients. The IVCF retrieval rate in trauma patients was 22.2% before and 21.4% after the program implementation. A commonly reported problem

	Before (n = 18)	After (n = 28)	P value
Demographics			
Age, mean years (SD)	45.8 (15.2)	44.6 (17.3)	0.67
Male, n (%)	16 (88.9)	22 (78.6)	0.45
Body mass index, mean (SD)	26.9 (4.3)	26 (5)	0.26
Clinical outcomes			
In-hospital mortality, n (%)	2 (11.1)	5 (17.9)	0.69
Mortality within 6 months, n (%)	2 (11.1)	5 (17.9)	0.69
Hospital length of stay (LoS), mean days (SD)	24.2 (14.8)	36.2 (63.9)	0.18
Intensive care unit LoS, mean days (SD)	12.6 (13)	15.3 (14)	0.24
IR clinic follow-up, n (%)	7 (38.8)	13 (46.4)	0.76
Procedural information			
Insertion procedure length, mean mins (SD)	25.9 (8.9)	27.1 (10.1)	0.46
Insertion fluoroscopy, mean mins (SD)	2 (0.9)	2.3 (1.4)	0.16
Permanent filters placed, n (%)	0	0	1
Time from insertion to removal, mean days (SD)	127.8 (159.3)	85.4 (46.4)	0.22
Removal procedure length, mean mins (SD)	41.5 (24.8)	39.3 (13.9)	0.47
Removal fluoroscopy, mean mins (SD)	9.2 (8.6)	7.9 (3.6)	0.17
IVCF retrieval, n (%)	4 (22.2)	6 (21.4)	1
Adjusted IVCF retrieval, n (%)	4 (25)	6 (26.1)	1
Failure of IVCF retrieval, n (%)	0	2 (7.14)	0.51
IVCF complications, n (%)			
Total	0	4 (14.3)	0.14
Thrombosis	0	3 (10.7)	0.27
Protrusion of struts through IVC	0	0	1
Tilting	0	0	1
Pulmonary embolism after IVCF insertion	0	1 (3.6)	1

SD, standard deviation.

in the trauma patient population from other similar Level 1 Trauma Centers is the low compliance with follow-up appointments, which is consistent with the results of our study.^{9,10} The results of this study suggest that attempting to improve IVCF retrieval rates by increasing adherence to follow-up appointments may not be an effective solution for this patient population. The retrieval of IVCFs with prolonged dwell times is more difficult, typically requiring advanced retrieval techniques and possibly leading to retrieval-related complications.^{14,15} For this reason, timely filter retrieval is necessary to reduce the risk of filter-related complications associated with prolonged dwell time.⁵

The findings of this study emphasize the need for increased focus on improving the rate of IVCF retrieval in trauma patients.^{3,4,16,17} The use of IVCFs in trauma patients has been investigated and documented in the literature in recent years.¹⁸⁻²¹ Considerable focus has been placed on the efficacy of and

indications for IVCF implantation in trauma patients, ultimately resulting in recent guideline changes.⁵ However, this fixation on the indications for IVCF implantation in trauma patients has not been met with a similar focus on ensuring that implanted IVCFs are retrieved. As current guidelines regarding IVCF use in trauma patients evolve, measures to improve retrieval rates in this patient population should be included. One proposed strategy to increase the rate of IVCF retrieval has been the removal of filters before patient discharge.²² In our study, 50% (23/46) of trauma patients with IVCFs placed had a HLoS >21 days, and 37% (17/46) had a HLoS >30 days. For trauma patients with prolonged lengths of stay, it may be possible to retrieve IVCFs before discharge as long as the patients no longer have contraindications to anticoagulation. This intervention could be an effective way to improve IVCF retrieval rates in trauma patients without the need for increased follow-up visits. The premature retrieval of

IVCFs may increase the risk of complications from thromboembolic events;^{23,24} however, our study did not identify any subsequent PEs following IVCF retrieval. An additional strategy cited as having an 86.9% retrieval rate in trauma patients was a reduction in the number of IVCFs placed in trauma patients combined with the use of a dedicated trauma nurse practitioner or physician's associate to ensure strict patient follow-up.^{17,25}

This study showed an overall complication rate of 12.2%, which is consistent with the reported 10%–22.2% complication rate.²⁶⁻²⁹ The incidence of IVC thrombosis was significantly higher in the trauma population; this complication, in particular, has been associated with prolonged dwell times, as the risk-to-benefit ratio becomes less favorable.⁴ To reduce the risk of filter-related complications associated with prolonged dwell times and improve patient outcomes, timely IVCF retrieval in trauma patients is necessary.

Table 4. Comparison of inferior vena cava (IVC) filter (IVCF) retrieval in non-trauma patients before and after IVCF retrieval program implementation

	Before (n = 34)	After (n = 84)	P value
Demographics			
Age, mean years (SD)	57.4 (15.5)	59.9 (14.6)	0.32
Male, n (%)	19 (55.9)	40 (47.6)	0.54
Body mass index, mean (SD)	32.7 (13.6)	29.7 (10.2)	0.12
Clinical outcomes			
In-hospital mortality, n (%)	2 (5.9)	11 (13.1)	0.34
Mortality within 6 months, n (%)	2 (5.9)	14 (16.7)	0.15
Hospital length of stay (LoS), mean days (SD)	17.3 (17)	24.9 (42.6)	0.22
Intensive care unit LoS, mean days (SD)	5.4 (12.2)	9.1 (26.2)	0.33
IR clinic follow-up, n (%)	14 (41.2)	53 (63.1)	0.04
Procedural information			
Insertion procedure length, mean mins (SD)	25.1 (10.7)	29.5 (20.2)	0.14
Insertion fluoroscopy, mean mins (SD)	2.5 (2.2)	3.7 (5.4)	0.13
Permanent filters placed, n (%)	0	8 (9.5)	0.10
Time from insertion to removal, mean days (SD)	50.1 (83.4)	142.9 (133.4)	0.0003
Removal procedure length, mean mins (SD)	31.8 (17)	29.4 (8.2)	0.22
Removal fluoroscopy, mean mins (SD)	4.3 (5.3)	3.8 (1.6)	0.36
IVCF retrieval, n (%)	12 (35.3)	38 (45.2)	0.41
Adjusted IVCF retrieval, n (%)	12 (37.5)	38 (61.3)	0.03
Failure of IVCF retrieval, n (%)	0	0	1
IVCF complications, n (%)			
Total	3 (8.8)	7 (8.3)	1
Thrombosis	2 (5.9)	1 (1.2)	0.2
Protrusion of struts through IVC	1 (3.0)	4 (4.8)	1
Tilting	0	2 (2.4)	1
Pulmonary embolism after IVCF insertion	0	0	1

SD, standard deviation.

An interesting collateral finding in this study was that the establishment of an IVCFRP resulted in patient referrals for complex filter retrievals. A total of 32 patients were referred to our center for complex filter retrieval. Retrieval was possible in 53.1% of these patients. The mean indwelling time in this group of patients was approximately 8 years. All retrievals required advanced techniques, as most filters with prolonged dwell times had a filter-associated complication, and many filters were discovered to have multiple complications. The establishment of an IVCFRP may increase the retrieval rates of filters in patients who have been lost to follow-up and experience a complication with the IVCF.^{13,30} The limitations of this study are that it is a retrospective, single-center study with a relatively low number of trauma patients who underwent IVCF insertion. In addition, the implementation of the IVCFRP occurred during the coronavirus disease-2019

pandemic, possibly introducing residual confounding that was not adjusted for. The results of this study confirm that there is a need to establish specific filter retrieval guidelines for trauma patients.

In conclusion, this study showed that in our center, the implementation of an IVCFRP improved the overall rate of IVCF retrieval and the rate of IVCF retrieval in non-trauma patients but did not significantly increase the rate of IVCF retrieval in trauma patients. With current guidelines focusing on increasing IVCF retrieval rates through improved patient follow-up, additional strategies may be necessary to increase IVCF retrieval in the trauma population, which has been associated with low follow-up adherence. Practice strategies, including retrieval before discharge or the establishment of a dedicated trauma nurse practitioner, may increase retrieval rates in trauma patients.

Footnotes

Conflict of interest disclosure

A.A. Smith is a paid consultant for Aroa Biosurgery and on the advisory board for Prytime Medical Devices. P. Greiffenstein is a paid consultant for Zimmer Biomet and MedExpert. J.E. Schoen is a paid consultant for Polynovo, Avita Medical, Lifesciences Plus, and Livanova. The other authors have no disclosures. Bahri Üstünsöz, MD, serves as Section Editor for Diagnostic and Interventional Radiology. He had no involvement in the peer review of this article and had no access to information regarding its peer review.

References

1. Cook AD, Gross BW, Osler TM, et al. Vena cava filter use in trauma and rates of pulmonary embolism, 2003-2015. *JAMA Surg.* 2017;152(8):724-732. [\[Crossref\]](#)
2. Gunning AC, Maier RV, de Rooij D, Leenen LPH, Hietbrink F. Venous thromboembolism (VTE) prophylaxis in severely injured patients: an international comparative assessment. *Eur J Trauma Emerg Surg.* 2021;47(1):137-143. [\[Crossref\]](#)
3. Karmy-Jones R, Jurkovich GJ, Velmahos GC, et al. Practice patterns and outcomes of retrievable vena cava filters in trauma patients: an AAST multicenter study. *J Trauma.* 2007;62(1):17-25. [\[Crossref\]](#)
4. Aryafar H, Kinney TB. Optional inferior vena cava filters in the trauma patient. *Semin Intervent Radiol.* 2010;27(1):68-80. [\[Crossref\]](#)
5. Kaufman JA, Barnes GD, Chaer RA, et al. Society of Interventional Radiology Clinical Practice Guideline for inferior vena cava filters in the treatment of patients with venous thromboembolic disease: developed in collaboration with the American College of Cardiology, American College of Chest Physicians, American College of Surgeons Committee on Trauma, American Heart Association, Society for Vascular Surgery, and Society for Vascular Medicine. *J Vasc Interv Radiol.* 2020;31(10):1529-1544. [\[Crossref\]](#)
6. Klinken S, Humphries C, Ferguson J. Establishment of an inferior vena cava filter database and interventional radiology led follow-up - retrieval rates and patients lost to follow-up. *J Med Imaging Radiat Oncol.* 2017;61(5):630-635. [\[Crossref\]](#)
7. Salei A, Raborn J, Manapragada PP, Stoneburner CG, Aal AKA, Gunn AJ. Effect of a dedicated inferior vena cava filter retrieval program on retrieval rates and number of patients lost to follow-up. *Diagn Interv Radiol.* 2020;26(1):40-44. [\[Crossref\]](#)
8. Stevens H, Bortz H, Chao S, et al. Improving the rate of inferior vena cava filter retrieval through multidisciplinary engagement. *Res Pract Thromb Haemost.* 2023;7(1):100040. [\[Crossref\]](#)
9. Leukhardt WH, Golob JF, McCoy AM, Fadlalla AM, Malangoni MA, Claridge JA. Follow-up disparities after trauma: a real problem for outcomes research. *Am J Surg.* 2010;199(3):348-353. [\[Crossref\]](#)
10. Overton TL, Shafi S, Gandhi RR. Local access to care programs increase trauma patient follow-up compliance. *Am J Surg.* 2014;208(3):476-479. [\[Crossref\]](#)
11. Morris E, Duszak R Jr, Sista AK, Hemingway J, Hughes DR, Rosenkrantz AB. National trends in inferior vena cava filter placement and retrieval procedures in the medicare population over two decades. *J Am Coll Radiol.* 2018;15(8):1080-1086. [\[Crossref\]](#)
12. Sheehan M, Coppin K, O'Brien C, et al. A single center 9-year experience in IVC filter retrieval - the importance of an IVC filter registry. *CVIR Endovasc.* 2022;5(1):15. [\[Crossref\]](#)
13. Sterbis E, Lindquist J, Jensen A, et al. Inferior vena cava filter retrieval rates associated with passive and active surveillance strategies adopted by implanting physicians. *JAMA Netw Open.* 2023;6(3):e233211. [\[Crossref\]](#)
14. Laws JL, Lewandowski RJ, Ryu RK, Desai KR. Retrieval of inferior vena cava filters: technical considerations. *Semin Intervent Radiol.* 2016;33(2):144-148. [\[Crossref\]](#)
15. Quencer KB, Smith TA, Deipolyi A, et al. Procedural complications of inferior vena cava filter retrieval, an illustrated review. *CVIR Endovasc.* 2020;3(1):23. [\[Crossref\]](#)
16. Smoot RL, Koch CA, Heller SF, et al. Inferior vena cava filters in trauma patients: efficacy, morbidity, and retrievability. *J Trauma.* 2010;68(4):899-903. [\[Crossref\]](#)
17. Leeper WR, Murphy PB, Vogt KN, et al. Are retrievable vena cava filters placed in trauma patients really retrievable? *Eur J Trauma Emerg Surg.* 2016;42(4):459-464. [\[Crossref\]](#)
18. Lee SJ, Fan S, Guo M, et al. Prophylactic IVC filter placement in patients with severe intracranial, spinal cord, and orthopedic injuries at high thromboembolic event risk: a utilization and outcomes analysis of the National Trauma Data Bank. *Clin Imaging.* 2022;91:134-140. [\[Crossref\]](#)
19. Kim M, Lee SY, Cha JG, et al. Single center experience of inferior vena cava filter retrieval in trauma patients: contrast-enhanced CT-based retrieval within hospital stay. *Clin Imaging.* 2021;79:43-47. [\[Crossref\]](#)
20. Trung Tran T, Bjarnason H, McDonald J, et al. Prophylactic placement of inferior vena cava filters and the risk of death or venous thromboembolism in severe trauma patients: a retrospective study comparing two hospitals with different approaches. *Acta Radiol Open.* 2021;10(3):2058460121999345. [\[Crossref\]](#)
21. Alshaqqaq HM, Al-Sharydah AM, Alshahrani MS, Alqahtani SM, Amer M. Prophylactic inferior vena cava filters for venous thromboembolism in adults with trauma: an updated systematic review and meta-analysis. *J Intensive Care Med.* 2023;38(6):491-510. [\[Crossref\]](#)
22. Robbins JM, Garwe T, McCarthy CK, et al. Removal of retrievable inferior vena cava filters before discharge: is it associated with increased incidence of pulmonary embolism? *J Trauma Acute Care Surg.* 2019;87(5):1113-1118. [\[Crossref\]](#)
23. Ho KM, Tan JA, Burrell M, Rao S, Misur P. Venous thrombotic, thromboembolic, and mechanical complications after retrievable inferior vena cava filters for major trauma. *Br J Anaesth.* 2015;114(1):63-69. [\[Crossref\]](#)
24. Pan Y, Zhao J, Mei J, Shao M, Zhang J, Wu H. Retrievable inferior vena cava filters in trauma patients: prevalence and management of thrombus within the filter. *Eur J Vasc Endovasc Surg.* 2016;52(6):830-837. [\[Crossref\]](#)
25. Matsumoto MM, Cun A, DeSanto C, Paycardo A, Stavropoulos SW, Trerotola SO. Managing an effective system for retrieving IVC filters: outcomes of a prospective patient database, 2012-2023. *CVIR Endovasc.* 2025;8(1):35. [\[Crossref\]](#)
26. Wassef A, Lim W, Wu C. Indications, complications and outcomes of inferior vena cava filters: a retrospective study. *Thromb Res.* 2017;153:123-128. [\[Crossref\]](#)
27. Shabib AB, Alsayed F, Aldughaythir S, et al. Indications, retrieval rate, and complications of inferior vena cava filters: single-center experience in Saudi Arabia. *Ann Thorac Med.* 2018;13(2):108-113. [\[Crossref\]](#)
28. Nazzal M, Chan E, Nazzal M, et al. Complications related to inferior vena cava filters: a single-center experience. *Ann Vasc Surg.* 2010;24(4):480-486. [\[Crossref\]](#)
29. Bajda J, Park AN, Raj A, Raj R, Gorantla VR. Inferior vena cava filters and complications: a systematic review. *Cureus.* 2023;15(6):e40038. [\[Crossref\]](#)
30. Swietlik JF, Rose AE, Meram E, Schwartz BS, Matsumura JS, Laeseke PF. Reengaging patients with forgotten filters through an institutional multidisciplinary approach. *J Vasc Surg Venous Lymphat Disord.* 2023;11(5):995-1003. [\[Crossref\]](#)



Single-center retrospective comparative study of 40–120 μm versus 100–300 μm trisacryl gelatin microspheres as the main embolic agents for prostatic artery embolization

Hippocrates Moschouris¹

Çağın Şentürk²

Konstantinos Stamatiou³

Apostolos Mouroukis¹

Chrysovalantis Stylianou¹

Dimitrios Kontothanasis³

¹University Hospital of Alexandroupolis, Clinic of Radiology, Alexandroupolis, Greece

²Tinaztepe Galen Hospital, Clinic of Radiology, İzmir, Türkiye

³General Hospital "Tzanio", Clinic of Urology, Piraeus, Greece

Handling editor: Barbaros Çil

Corresponding author: Hippocrates Moschouris

E-mail: hipmosch@gmail.com

Received 13 August 2025; revision requested 10 September 2025; last revision requested 03 November 2025; accepted 16 December 2025.



Epub: 23.02.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263602

PURPOSE

To compare the safety and efficacy of trisacryl gelatin microspheres (TGM) with diameters of 40–120 μm (TGM 40–120) vs. 100–300 μm (TGM 100–300) as the primary embolic agents for prostatic artery (PA) embolization (PAE).

METHODS

This single-center retrospective comparative study included patients with symptomatic benign prostatic hyperplasia treated with PAE using TGM 40–120 (group A) or TGM 100–300 (group B) over a two-year period. Evaluation included International Prostate Symptom Score (IPSS), quality of life score (QoL), prostatic volume (PV), post-void residual (PVR), and percentage of prostatic infarction (pPI) measured by contrast-enhanced ultrasound. Clinical success was defined as post-PAE reduction of IPSS ≤ 15 points with a decrease of at least 25% from the baseline, QoL score ≤ 3 points or a decrease of at least 1 point from baseline, and successful bladder catheter removal (for catheter-dependent patients). Adverse events were recorded and graded using the modified CIRSE classification system for complication reporting.

RESULTS

Most baseline characteristics did not differ significantly between group A ($n = 21$) and group B ($n = 31$). At 3 months post-PAE, there were no statistically significant differences between groups in improvement of IPSS (group A: $49.5 \pm 26.1\%$, group B: $51.8 \pm 23.3\%$, $P = 0.403$) or QoL (group A: $54.7 \pm 3.1\%$, group B: $51.2 \pm 14.1\%$, $P = 0.589$), nor in reduction of PV (group A: $35.5 \pm 16.9\%$, group B: $28.4 \pm 11.0\%$, $P = 0.104$) or PVR (group A: $59.8 \pm 23.9\%$, group B: $60.4 \pm 28.5\%$, $P = 0.934$).

However, pPI (calculated 1–4 days post-PAE) was significantly higher in group A ($48.3 \pm 16.2\%$ vs. $30.3 \pm 11.9\%$, $P = 0.007$). Mid-term PV reduction was also significantly greater in group A ($31.4 \pm 21.4\%$ vs. $12.2 \pm 10.2\%$, $P = 0.030$).

There were more complications in group A (7/21 vs. 4/31, $P = 0.095$). Three complications in group A were grade 3a events (bladder wall ischemia, $n = 2$; persistent severe anal and perineal pain requiring hospitalization and opioids, $n = 1$). No grade 3 events were observed in group B ($P = 0.060$).

Clinical success rates at 6, 12, 24, and 36 months post-PAE were 76.2%, 76.2%, 71.1%, and 71.1%, respectively, for group A, and 87.1%, 83.9%, 79.9%, and 79.9%, respectively, for group B ($P = 0.451$).

CONCLUSION

Compared with TGM 100–300, PAE with TGM 40–120 appears to provide similar clinical benefit but results in significantly higher pPI and greater mid-term prostate shrinkage. Ischemic complications may be more frequent and more severe with the smaller microspheres.

CLINICAL SIGNIFICANCE

This is the first systematic comparison between the smallest and the second smallest commercially available TGM sizes in the clinical context of PAE. These findings may help guide ongoing research to identify the most appropriate embolic material for PAE.

KEYWORDS

Prostatic artery embolization, trisacryl gelatin microspheres, prostatic infarction, complications

You may cite this article as: Moschouris H, Şentürk Ç, Stamatiou K, Mouroukis A, Stylianou C, Kontothanasis D. Single-center retrospective comparative study of 40–120 μm versus 100–300 μm trisacryl gelatin microspheres as the main embolic agents for prostatic artery embolization. *Diagn Interv Radiol*. 2026;32(5):700–708.

Although liquid embolic agents have recently emerged as a promising material for prostatic artery embolization (PAE) in symptomatic benign prostatic hyperplasia (BPH),¹ the vast majority of procedures are still performed with polyvinyl alcohol (PVA) particles or microspheres.² Among the latter, trisacryl gelatin microspheres (TGM) with diameters of 100–300 μm and 300–500 μm (Embosphere, Merit Medical Systems Inc., South Jordan, UT, USA), or a combination of these sizes, have been widely used for PAE worldwide since the early applications of this procedure,³ and several studies have demonstrated their satisfactory safety and efficacy profiles.^{4,6}

Currently, the smallest commercially available diameter of TGM is 40–120 μm (TGM 40–120), and experience with this very small TGM in PAE remains limited.⁷ Experimental studies have demonstrated that TGM 40–120 can cause complete occlusion of arteries with a mean diameter as small as 137.8 μm ,⁸ and that TGM can cause significantly larger infarction than other types of microspheres.⁹ A human cadaveric study¹⁰ found that arteries surrounding the hyperplastic adenomatous prostatic nodules (perinodal arteries) had a mean diameter of 150 μm (range: 59–266 μm), whereas arteries entering the prostatic nodules (terminal, intranodal arteries) had a mean diameter of 56 μm (range: 24–104 μm).

It may be hypothesized that TGM 40–120 could maximize the efficacy of PAE through more distal penetration and tighter packing within perinodal and even some intranodal arteries, resulting in more complete and irreversible ischemia. Notably, ischemic necrosis of hyperplastic prostatic tissue induced by embolization is considered one of the proposed mechanisms of therapeutic action of PAE.¹¹ However, in addition to these potential benefits, the intense ischemia caused by very

small microspheres may also be associated with a higher prevalence and greater severity of ischemic complications compared with larger microspheres.

To evaluate these hypotheses regarding the safety and efficacy of PAE using the smallest commercially available TGM, a retrospective comparative study was conducted in patients with BPH who underwent PAE with TGM 40–120 and another group treated with the more commonly used TGM 100–300.

Methods

Patients

A retrospective review was conducted of patients treated with PAE for symptomatic BPH at a single tertiary center during a 2-year period (April 2021–April 2023). Inclusion criteria were largely consistent with CIRSE standards of practice for PAE:¹² moderate-to-severe lower urinary tract symptoms [International Prostate Symptom Score (IPSS) ≥ 8]; quality of life (QoL) ≥ 3 ; failure of medical treatment (at least 6 months of administration of 5 α -reductase inhibitors or selective $\alpha 1$ blockers) or urinary retention managed with an indwelling bladder catheter (IBC), with at least three failed catheter removal attempts prior to PAE; and a prostate volume > 40 mL measured by ultrasound.

Exclusion criteria were previous surgical or interventional prostate treatment, urinary tract infection, prostate or bladder cancer, neurogenic bladder, large (> 2 cm) bladder diverticula or bladder stones, contraindications to angiographic procedures, and vascular pathologies that precluded safe arterial access. A minimum follow-up period of three months was also required.

All patients had been informed in detail regarding the procedure and potential benefits and risks (including complications specifically associated with very small microspheres) and had provided their written informed consent. The study was approved by the Scientific and Ethics Committee of General Hospital Tzanio (decision number:17; date: October 10th, 2024). Patients treated with TGM 40–120 comprised “group A,” and those treated with TGM 100–300 comprised “group B”. Figure 1 shows how the two study groups were formed.

Baseline evaluation and treatment planning

All patients underwent clinical evaluation by the referring urologist prior to PAE, and

the IPSS and QoL scores were recorded. A transabdominal ultrasound was performed a few days before PAE to measure baseline prostatic volume (PV) and post-void residual (PVR). Contrast-enhanced ultrasonography (CEUS) was performed at the same session to assess prostatic parenchymal enhancement. Computed tomographic angiography (CTA) of the pelvic arteries was performed 5–10 days prior to PAE for vascular planning in all patients.

PAE procedure

Thirty minutes before the procedure, the patients received an intravenous antibiotic (cefradine, 2 g), analgesic (parecoxib, 40 mg), and gastric protection. Vascular access was obtained via the right or left common femoral artery using the Seldinger technique, and a 5 French (Fr) reverse-curve angiographic catheter (usually the MOT catheter, Merit Medical Systems Inc., South Jordan, UT, USA) was introduced into the ipsilateral internal iliac artery. An ipsilateral oblique angiogram was then performed to clearly demonstrate the origin of the PA.

The PA was catheterized with a 1.98 Fr microcatheter (PARKWAY SOFT, Asahi Intecc Co.) and a double-angled 0.016” microguide-wire (ASAHI Meister). If catheterization failed, the microcatheter was exchanged for a steerable one (2.4 Fr, SwiftNINJA, Merit Medical Systems). Successful catheterization was confirmed with a superselective angiogram using manual contrast injection through the microcatheter.

Embolization commenced with the microcatheter tip positioned in the extraprostatic segment of the PA. Syringes containing 2.0 ml of microspheres were used for both TGM sizes. Microspheres were diluted according to the manufacturer's instructions, with a total volume of 22 mL of diluted microspheres prepared per syringe. Slow, flow-directed manual injection was performed under fluoroscopy until complete flow stasis was observed in the targeted PA.

Distal advancement of the microcatheter and additional embolization of intraprostatic branches were then attempted according to the “PERFecTED” technique.¹³ Finally, the microcatheter was withdrawn to the proximal PA, and a minimal amount of diluted PVA particles (Bearing, 150–250 μm , Merit Medical) was injected to enhance proximal flow stasis in all the patients. The contralateral PA was then embolized after a crossover maneuver.

Main points

- Microspheres measuring 40–120 μm represent the smallest commercially available trisacryl gelatin embolic microspheres (Embospheres), and their use in prostatic artery embolization remains limited.
- Compared with the second-smallest size (100–300 μm), 40–120 μm trisacryl gelatin microspheres cause significantly more extensive prostatic necrosis and greater mid-term prostate shrinkage; however, these differences are not clearly reflected in the clinical efficacy of PAE.
- Microspheres measuring 40–120 may cause more frequent and more severe ischemic complications than larger microspheres.

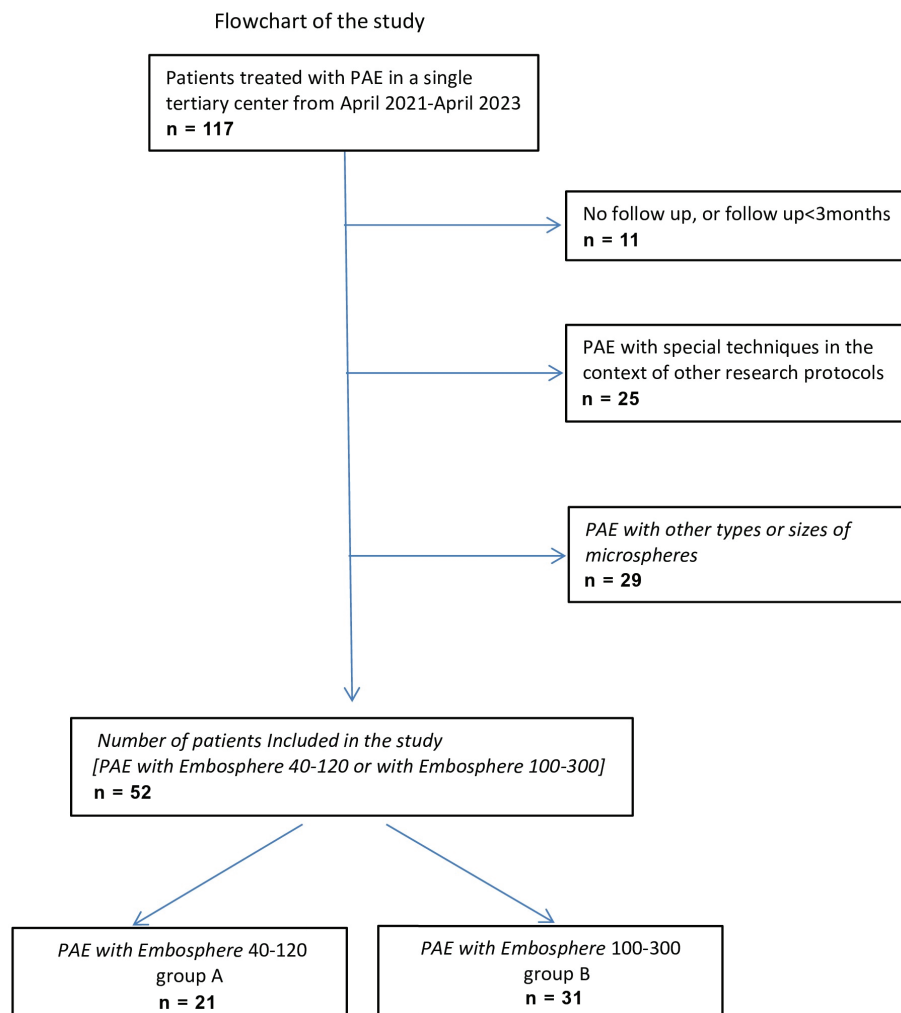


Figure 1. Flowchart of the study population. PAE, prostatic artery embolization.

All PAE procedures were performed using a flat-panel angiographic unit (Axiom Artis Zee, Siemens Healthineers, Erlangen, Germany) by the same operator, who had 6 years of prior experience performing PAE with TGM. Intraprocedural cone-beam computed tomography (CBCT), unenhanced ultrasound, and CEUS were available but used only when deemed necessary by the operator for detailed vascular assessment or monitoring of embolic agent distribution.

Clinical assessment and diagnosis of clinical success and of adverse events

Patients were clinically evaluated within the first 4 days after PAE, at 1 and 3 months post-PAE, and then every 3 months; IPSS and QoL scores were updated at each visit. After the first 6 months, some clinical visits could be replaced by telephone or online communication if all the required information could be consistently and reliably provided by the patient.

Clinical success post-PAE was diagnosed if all the following criteria were met: (a) IPSS ≤ 15 points with a decrease of at least 25% from baseline; (b) QoL score ≤ 3 points or a decrease of at least 1 point from baseline; (c) no need for additional medical or surgical treatment. For patients with urinary retention and IBC, clinical success also required successful and permanent catheter removal with spontaneous micturition and PVR < 100 mL.

Complications were recorded and graded according to the modified CIRSE classification system.¹⁴ Post-embolization syndrome (PES) was defined as dysuria, urgency, urinary frequency, nausea, fever (without evidence of sepsis), pelvic or prostatic pain, or any combination of these symptoms within the first 7 days following PAE.^{12,15} Because PES was considered an adverse event, rather than a complication, it was recorded separately. Rectal bleeding, hematuria, and hemospermia were also considered as adverse events

(not true complications) and components of PES if mild and self-limiting.^{12,14}

Intra- or post-procedural pain was not systematically recorded. However, cases of significant pain affecting technical or clinical outcomes, or pain associated with a complication, were documented, and pain intensity in these cases was assessed using the visual analogue scale (VAS).

Imaging follow-up

Unenhanced ultrasound was performed at the first post-procedure visit (within 4 days) and at 3 months post-PAE. PV was calculated using the ellipsoid formula based on transabdominal ultrasound measurements, and prostate shrinkage was determined by comparing baseline PV with PV at 3 months post-PAE.

CEUS was also performed at the first post-procedure visit to assess prostatic infarcts. Infarcts were defined as newly appear-

ing non-enhancing areas within the prostatic parenchyma. The volume of each infarct was calculated using the ellipsoid formula, and the volumes of all infarcts within the treated prostate were summed. The percentage of prostatic infarction (pPI) was calculated by dividing total infarct volume by PV, consistent with previous work.¹⁶

Additional imaging was not routinely scheduled during follow-up. However, if PV measurements were obtained during later visits (more than 3 months post-PAE), they were recorded. Imaging performed for other reasons (e.g., pelvic CT for complication assessment or unrelated imaging months or years after PAE) was also considered if they yielded findings relevant to the purposes of this work.

Statistical analysis

Descriptive statistics were used to present baseline characteristics and imaging and clinical outcomes. The Shapiro–Wilk test was used to assess the normality of the data distribution. Comparisons between group A and group B were performed for baseline and outcome parameters. Depending on parameter type and distribution, appropriate statistical tests were applied (Fisher’s exact

test, Welch’s t-test, Mann–Whitney U test).

The Kaplan–Meier method was used to calculate the clinical success rates of PAE over time, and the log-rank test was used to evaluate differences between the groups. Statistical significance was defined as $P < 0.05$. Statistical analyses were performed using the Statistical Package for Social Sciences (Version 26.0, SPSS Inc., Chicago, Illinois, USA).

Results

General and clinical outcomes

Fifty-two patients were included in the study: 21 in group A and 31 in group B. Apart from age, baseline clinical features were comparable between the groups (Table 1). Most anatomic features did not differ significantly between the two groups, either; however, there were significantly more patients with mild prostatic hyperplasia ($PV \leq 50$ mL) in group A (Supplementary Tables 1 and 2). Most technical aspects of PAE were also comparable between the two groups (Supplementary material Table 3). Notably, a significantly larger volume of diluted TGM suspension was injected in group A compared with group B (25.1 ± 6.9 mL vs. 20.0 ± 5.7 mL, $P = 0.008$). The volume of diluted PVA

injected at the end of embolization of each pelvic side did not exceed 2 mL in any case.

Follow-up ranged from 3 to 48 months [mean $\pm$ standard deviation (SD): 23.4 ± 15.5 months] for group A and from 3 to 45 months (mean $\pm$ SD: 23.3 ± 13.4 months) for group B. At 3 months post-PAE, there were no statistically significant differences in early clinical improvement (expressed by reduction of IPSS and QoL) between the two groups (Table 2).

Clinical success rates at 6, 12, 24, and 36 months post-PAE were: 76.2%, 76.2%, 71.1%, and 71.1%, respectively, for group A and 87.1%, 83.9%, 79.9%, and 79.9%, respectively, for group B; these differences were not statistically significant ($P = 0.451$, Figure 2). Among patients with mild prostatic hyperplasia ($n = 5$, all in group A), three experienced early clinical failure; in two of these, further investigation revealed pathologies other than BPH as potential causes of LUTS. The remaining two patients with mild prostatic hyperplasia reported only moderate symptomatic improvement, with IPSS reductions of 25% and 27%, respectively.

Table 1. Demographic and clinical features of the two patient groups at baseline

Variable	Group A (n = 21)	Group B (n = 31)	Significance of differences (P)
Age (years, mean $\pm$ SD)	75.3 $\pm$ 7.0	68.9 $\pm$ 10.1	0.024
PV (mL, mean $\pm$ SD)	95.3 $\pm$ 55.8	107.6 $\pm$ 43.4	0.472
PVR	184 $\pm$ 162	118 $\pm$ 84	0.097
IPSS	26.2 $\pm$ 5.8	26.5 $\pm$ 4.0	0.984
QoL	4.2 $\pm$ 0.8	4.0 $\pm$ 0.6	0.284
Indication for PAE (proportion of patients)			
-Moderate LUTS	2/21	3/31	0.984
-Severe LUTS	10/21	19/31	0.332
-IBC	9/21	9/31	0.303

For moderate LUTS, IPSS = 8–19; for severe LUTS, IPSS = 20–35. Welch’s t-test was used for comparisons of age, PV, and PVR. Mann–Whitney U test was used for comparisons of IPSS and QoL. Fisher’s exact test was used for comparisons of proportions of patients with moderate/severe LUTS and IBC. SD, standard deviation; PV, prostate volume; PVR, post-void residual; IPSS, International Prostate Symptom Score; QoL, quality of life; PAE, prostatic artery embolization; LUTS, lower urinary tract symptoms; IBC, indwelling bladder catheter.

Table 2. Comparison of short-term outcome parameters for the two patient groups

Variable	Group A (n = 21)	Group B (n = 31)	Significance of differences (P)
pPI ^a (mean $\pm$ SD, %)	48.3 $\pm$ 16.2	30.3 $\pm$ 11.9	0.007
PV reduction ^b (mean $\pm$ SD, %)	35.5 $\pm$ 16.9	28.4 $\pm$ 11.0	0.104
PVR reduction ^b (mean $\pm$ SD, %)	59.8 $\pm$ 23.9	60.4 $\pm$ 28.5	0.934
IPSS reduction ^b (mean $\pm$ SD, %)	49.5 $\pm$ 26.1	51.8 $\pm$ 23.3	0.403
QoL improvement ^b (mean $\pm$ SD, %)	54.7 $\pm$ 31.6	51.2 $\pm$ 14.1	0.589

^aCalculated 1–4 days post-PAE; ^bCalculated 3 months post-PAE and compared with respective baseline measurements. All values for outcome parameters represent percentages (%), not absolute values. Welch’s t-test was used for comparisons. pPI, percentage of prostatic infarction; SD, standard deviation; PV, prostate volume; PVR, post-void residual; IPSS, International Prostate Symptom Score; QoL, quality of life.

Imaging outcomes

Early CEUS (1–4 days post-PAE) revealed almost 60% greater pPI in group A. At 3 months post-PAE, prostate shrinkage was also greater in group A, although the difference did not reach statistical significance. Reduction in PVR was similar between the two groups. Detailed comparisons of the early outcomes are provided in Table 2.

Mid-term PV measurements (12–48 months post-PAE) were available for 9 patients in group A (these 9 patients formed subgroup A) and for 11 patients in group B (subgroup B). In this subset, prostate shrinkage was significantly greater in subgroup A (Table 3, Figure 3). Almost complete prostate regrowth (0%–5% difference compared with baseline PV) occurred in 1 of 9 patients in subgroup A and 4 of 11 patients in subgroup B. Interestingly, two of the four subgroup B patients with almost complete prostate regrowth reported sustained symptomatic improvement. Sustained symptomatic improvement was also reported by all patients with documented prostate shrinkage on

mid-term PV measurements.

Adverse events

Post-procedure PES was by far the most common adverse event in both groups (Table 4).

Two patients in group A experienced significant intraprocedural pain (VAS scores 8 and 7), occurring approximately 5 minutes after completion of embolization on one pelvic side. In the first patient, the pain (VAS 8) persisted despite additional intravenous analgesia and intra-arterial lidocaine and, combined with patient anxiety and discomfort, led to early termination of the procedure before embolization of the contralateral side. According to the updated CIRSE classification system,¹⁴ this event was recorded as a grade 1b complication.

Additionally, three grade 2 complications occurred in group A (acute urinary retention) and four in group B (acute urinary retention, n = 3; small ischemic penile ulcers with moderate pain, n = 1). All grade 2 complications

resolved completely with conservative treatment (and/or Foley catheter placement for less than 7 days).

Three grade 3a complications occurred in group A. Two patients (both with type I PA origin, or “common vesicoprostatic trunk”) experienced hematuria and moderate pain (VAS 5) 2–4 days post-PAE. In both patients, CT revealed bladder wall ischemia with no signs of bladder perforation, and cystoscopy was performed to assess the extent of injury and remove superficial necrotic bladder wall tissue. In one of these patients, a necrotic portion of the protruding median lobe was also cystoscopically removed (Figure 4). Both patients recovered with no permanent sequelae.

A third patient in group A (with a history of sigmoidectomy for cancer 4 years earlier) developed severe anal and perineal pain (VAS 8) beginning the day after PAE. The patient required hospitalization for 3 days, and the pain persisted for several months despite oral opioid analgesia. CT performed 2 days post-PAE indicated inadvertent embolization

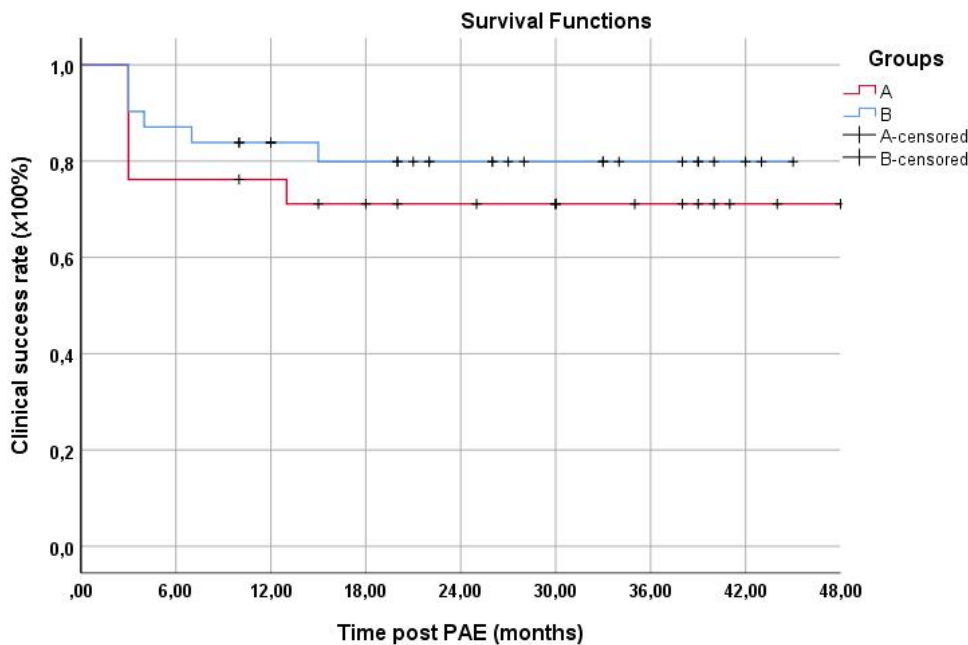


Figure 2. Kaplan–Meier curves showing no statistically significant difference ($P = 0.451$) in the clinical success rates between group A (PAE with TGM 40–120) and group B (PAE with TGM 100–300). PAE, prostatic artery embolization; TGM, trisacryl gelatin microspheres.

Table 3. Comparison of mid-term outcome parameters for the two patient subgroups			
Variable	Subgroup A (n = 9)	Subgroup B (n = 11)	Significance of differences (P)
Time post-PAE (min/max/mean/SD, months)	12/48/34/11.6	12/36/22/9.2	-
PV reduction (mean ± SD, %)	31.4 ± 21.4	12.2 ± 10.2	0.030
IPSS reduction (mean ± SD, %)	47.1 ± 24.6	37.3 ± 20.7	0.414
Welch's t-test was used for comparisons.			
PAE, prostatic artery embolization; SD, standard deviation; PV, prostate volume; IPSS, International Prostate Symptom Score; min/max, minimum/maximum.			

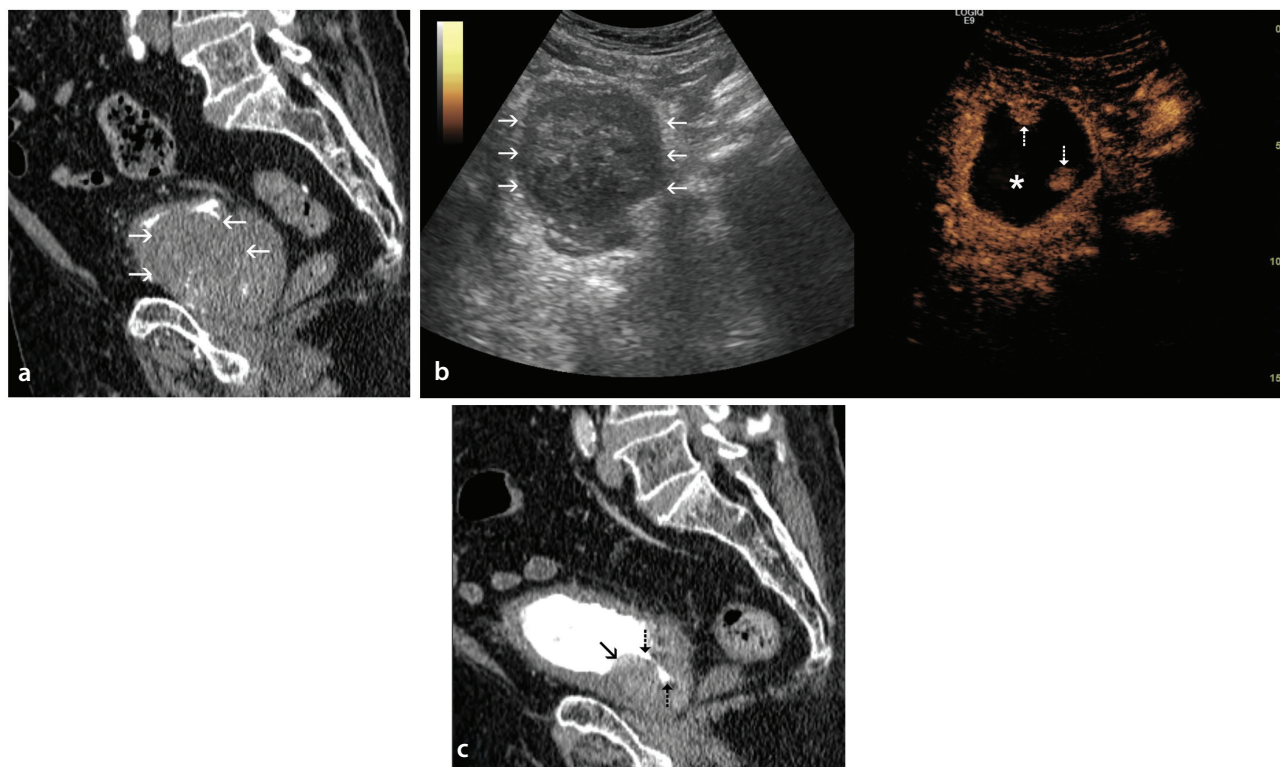


Figure 3. Representative images from a clinically successful case of prostatic artery embolization (PAE) with TGM 40–120. Sagittal reformatted planning computed tomography (CT) image prior to PAE; (a) shows prostatic enlargement and massive intravesical prostatic protrusion (arrows). Axial dual sonographic image obtained 2 days post-PAE (b); with reference gray-scale image on the left showing the bulky intraprostatic protrusion (arrows) and corresponding contrast-enhanced ultrasonography image on the right showing extensive prostatic infarction (*) and minimal residual enhancing tissue (dotted arrows). Sagittal reformatted CT image (c); obtained 12 months post-PAE for an unrelated reason, shows marked prostate shrinkage (arrow) and relief of pressure on the prostatic urethra (dotted arrows).

TGM, trisacryl gelatin microspheres.

Adverse event	Group A (n = 21)	Group B (n = 31)	Significance of differences (P)
Post-embolization syndrome	10	11	0.404
Severe intraprocedural pain	2	0	0.158
Mild rectal bleeding*	1	0	0.404
Mild hemospermia*	0	1	1.000
Complication			
Grade 1b	1	0	0.404
Grade 2	3	4	1.000
Grade 3a	3	0	0.060
Total number of complications	7	4	0.095

Fisher's exact test was used for comparisons. *Not considered true complications, despite specific reporting.

of the left middle rectal artery. This vessel had not been identified on planning CTA or DSA, and intraprocedural CBCT had not been performed. No rectal wall thickening or dehiscence was observed on CT post-PAE, and colonoscopy was not performed. Despite favorable long-term urologic outcomes and substantial, durable prostate debulking, this case was registered as a clinical failure because the pain worsened the patient's QoL for at least 6 months post-PAE (from 5 to 6). An additional imaging finding in this case

was the appearance of linear hyperdensities with vascular distribution on follow-up CT (Figure 5).

Discussion

To the best of the authors' knowledge, this is the first systematic comparison between the smallest and the second smallest sizes of TGM (diameters 40–120 μ m and 100–300 μ m, respectively) as the main embolic agents for PAE. Compared with other TGM sizes, TGM 40–120 is characterized not only by a small-

er average diameter but also by narrower size calibration; consequently, this very small TGM is capable of deeper penetration into the prostatic microvasculature and of inducing more severe and irreversible tissue ischemia.⁸ These characteristics of TGM 40–120 are reflected in the significantly higher pPI (almost 60%) observed in group A in the first days post-PAE in this study. Notably, previous work using a similar PAE technique but microspheres $\geq 100 \mu$ m¹⁶ and the same method for measuring prostatic infarction, demonstrated

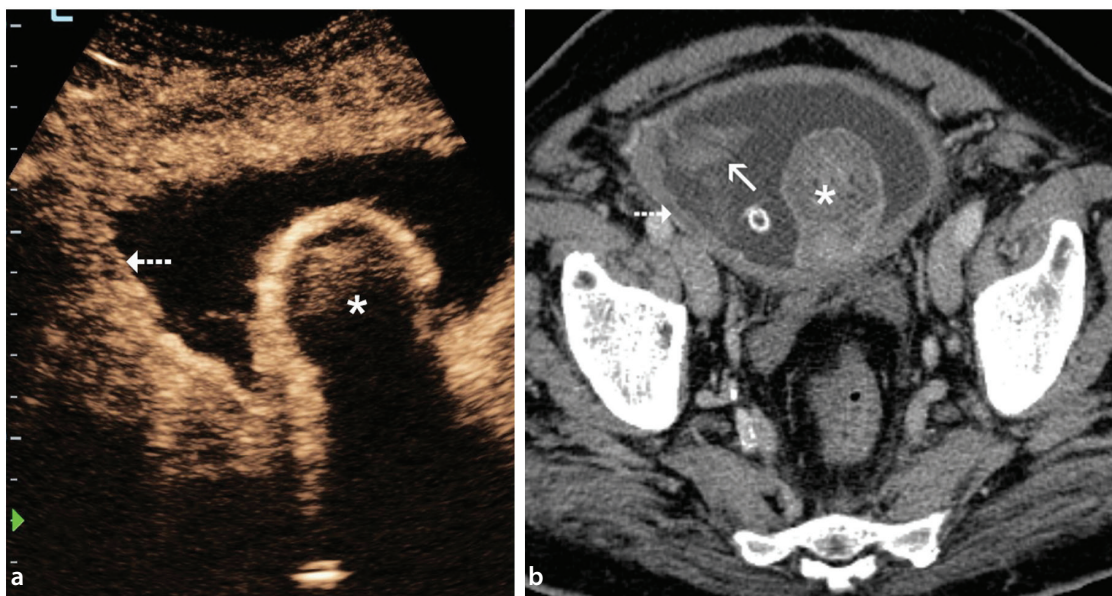


Figure 4. Representative images from a complicated case of prostatic artery embolization (PAE) with TGM 40–120. Three days post-PAE, axial contrast-enhanced ultrasonography image (a); shows extensive infarction (*) of a large intravesical prostatic protrusion. A small area of reduced enhancement is suspected in the right lateral bladder wall (dotted arrow). Axial contrast-enhanced computed tomography image (b); performed immediately thereafter, also shows prostatic infarction (*), along with hypoenhancement (dotted arrow) and irregularity of the right lateral bladder wall. An irregularly shaped intraluminal mass (arrow) corresponds to sloughed necrotic superficial layers of the bladder wall, as confirmed by cystoscopy (images not shown).

TGM, trisacryl gelatin microspheres.

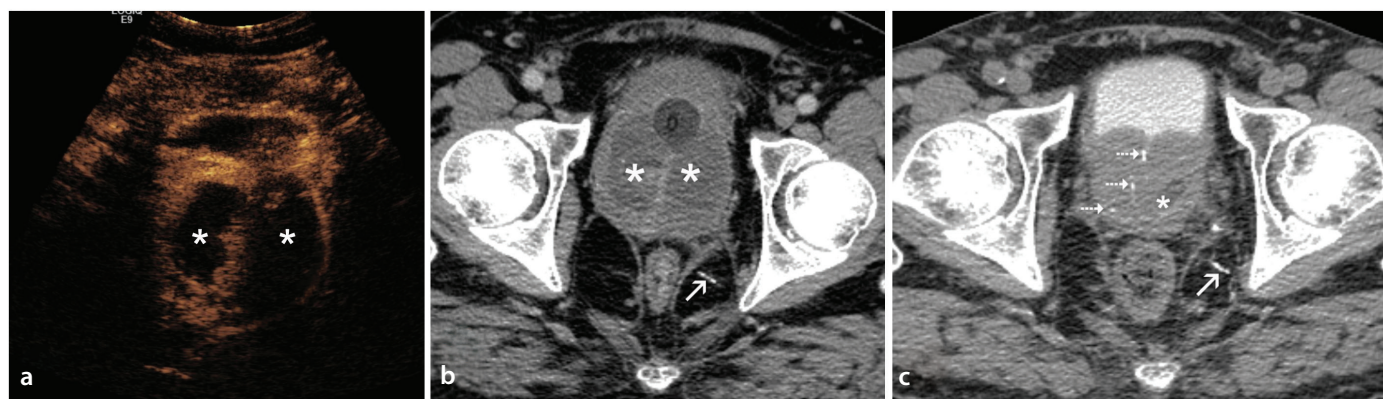


Figure 5. Representative images from a complicated case of prostatic artery embolization (PAE) with trisacryl gelatin microspheres 40–120 (severe persistent anal and perineal pain). Two days post-PAE, contrast-enhanced ultrasonography image (a) shows extensive infarction (*) of both prostatic lobes. Axial contrast-enhanced computed tomography (CT) image obtained the same day (b) confirms prostatic infarction (*) and additionally shows linear hyperdensity (indicating the presence of the embolic mixture) along the expected trajectory of the left middle rectal artery (arrow) in the ischio-rectal fossa. Four years post-PAE, axial contrast-enhanced CT image (c) shows significant prostate shrinkage with minimal residual prostatic infarction (*) and persistent linear hyperdensity in the left ischio-rectal fossa (arrow). Similar, smaller hyperdensities are noted at the expected locations of intraprostatic arterial branches (dotted arrows). These findings were not present prior to PAE (baseline images not shown).

significantly lower pPI (25.1% in the previous study vs. 48.3% with TGM 40–120 in this study). By contrast, two other PAE studies, one using small (50 and 100 μ m) PVA particles¹⁷ and the other using a combination of temperature-sensitive liquid embolic agent with PVA,¹⁸ reported even higher mean proportions of prostatic ischemia (70% and 62%, respectively). Conversely, larger spherical or non-spherical PVA particles appear to cause less pronounced prostatic ischemia, in the range of 10%–11%.¹⁹

Laboratory investigation has shown that the small size and narrow calibration of microspheres promote not only more distal penetration but also tighter intraluminal packing and more complete vascular occlusion.²⁰ This may explain the significantly larger volume of microsphere–contrast medium suspension required for PAE with the smaller microspheres (25.1 mL per procedure in group A vs. 20.0 mL per procedure in group B, despite comparable prostate size between groups). Of note, a previous study of TGM

100–300⁴ reported a mean suspension volume of 11.9 mL per procedure (mean PV of 63.8 mL), which is roughly equivalent to the 20.0 mL per procedure observed in the present study (mean PV of 107.6 mL in group B). No corresponding published data are available for TGM 40–120.

Prostatic infarction detected early post-PAE is considered a good predictor of subsequent prostate debulking.^{11,21} Therefore, it is not surprising that TGM 40–120 was associated not only with larger prostatic infarctions

but also with greater prostate shrinkage. Importantly, this difference became more pronounced and statistically significant at mid-term evaluation, suggesting a reduced risk of prostate regrowth/revascularization with the smaller microspheres. Prostate regrowth between 3 and 12 months post-PAE with TGM 100–300 has been reported previously, but in that study,⁴ the comparison was made with larger (300–500 μ m) microspheres of the same type. In contrast to the present findings, that study⁴ suggested that larger TGMs (not smaller ones) provide more durable PV reduction. The reasons for this discrepancy are unclear. A “rebound” increase in PV (147.9 mL, or a 3.1% increase compared with baseline) has also been reported²² after a mean of 25.7 months post-PAE with PVA 100 and 200 μ m. Compared with these studies,^{4,22} the sustained prostate shrinkage observed with TGM 40–120 in the present study appears superior; however, these findings were observed in a small subgroup of patients.

It is also unclear whether the additional PVA particles (used at the end of embolization of each PA in the present study) contributed to the prevention of prostate regrowth. The amount of injected PVA was minimal and was used in both groups. Although PVA 150–250 may not be the ideal embolic material for PAE, its use for complementary proximal embolization was dictated by availability and low cost. Larger PVA particles were selected rather than the smaller TGMs already used because reflux from this proximal position, particularly after occlusion of smaller peripheral branches, is high, and the risk of ischemic complications would be much higher if small TGMs (either 40–120 or 100–300) were to reflux.

Although TGM 40–120 appeared superior in terms of prostatic infarction and mid-term prostate shrinkage, no significant differences were observed in clinical efficacy (as reflected by IPSS and QoL improvement) between the two patient groups. This contrasts with the finding of the only other study with TGM 40–120,⁷ which (although limited to a small number of catheter-dependent patients), reported higher clinical success rates with the smaller microspheres (70.6% vs. 42.9%). Prostatic ischemia and its consequences are probably not the only therapeutic mechanisms of PAE. For example, symptomatic improvement has been observed post-PAE even in the absence of detectable prostatic infarction on contrast-enhanced magnetic resonance.^{21,23} The role of prostate shrinkage in symptomatic improvement post-PAE is also debated,

as some studies have failed to demonstrate a consistent relationship between PV reduction and IPSS improvement. Indeed, patients with minimal prostate shrinkage post-PAE may still experience substantial symptomatic relief.²⁴ Similarly, two patients in group B in the present study experienced sustained improvement in IPSS and QoL indices despite prostate regrowth. Conversely, none of the patients with persistent prostate shrinkage at mid-term follow-up in this study experienced symptomatic recurrence. This may represent a potential advantage of TGM 40–120, which requires validation with larger studies.

Regarding baseline characteristics, patients in group A were significantly older, and mild prostatic enlargement was more common in this group. However, evidence regarding the clinical predictive value of these features is contradictory,¹¹ and the authors believe their overall impact on clinical outcomes in this study was limited.

The adverse events and complications associated particularly with the smaller microspheres in this study merit attention. Severe intraprocedural pain during PAE with larger microspheres has rarely been reported;²⁵ in fact, the relatively painless and minimally invasive profile of PAE is considered a major advantage. Nevertheless, two patients in group A experienced severe pain almost immediately after injection of TGM 40–120, likely due to rapid and intense ischemia. In one of these patients, pain directly affected the technical outcome, causing early termination of the procedure and precluding bilateral embolization.

The most severe (grade 3) complications in this study also occurred exclusively in group A, in contrast to the previous small study with TGM 40–120,⁷ where no such events were reported. The two cases of bladder wall ischemia can be explained by the specific type of PA origin (common vesicoprostatic trunk) and reflux of the microspheres into the superior vesical artery. This type of PA origin is the most common in many populations.²⁶ Consistent with previous reports,^{27,28} recovery was complete, and no permanent sequelae occurred, although additional care and cystoscopic intervention were required.

The incident of severe anal and perineal pain post-PAE was associated with inadvertent occlusion of the middle rectal artery, as demonstrated on follow-up CT. This type of non-target embolization occurs occasionally during PAE, and its sequelae are usually

mild and self-limiting; however, prolonged perineal pain lasting 3 months post-PAE has been reported.²⁸ The severity and duration of pain in the present case may at least partially be attributed to the small size of the TGM used. It is unclear whether the patient's prior sigmoidectomy (and any associated arterial ligations) contributed to the severity of the complication by reducing collateral circulation. An interesting (not previously reported) imaging finding on follow-up CT in the same patient was the appearance of linear hyperdensities with vascular distribution. These likely represent calcifications, either dystrophic or due to calcification of the TGM themselves, with a mechanism similar to that proposed for uterine fibroids after successful embolization.²⁹ Finally, the failure of both CTA and DSA to identify the middle rectal artery in this patient highlights the potential role of CBCT in visualizing small but clinically relevant pelvic branches and preventing non-target embolization.

Several limitations of the present study should be acknowledged. The small sample size (particularly for group A and the subgroups with mid-term data) limits the generalizability of the conclusions of this study. Magnetic resonance imaging, which may be more accurate than CT and US/CEUS for PV and pPI measurement, was not systematically used. Uroflowmetry was not consistently performed, thus limiting the objective evaluation of functional outcomes.

In conclusion, compared with TGM 100–300, PAE with TGM 40–120 appears to provide comparable clinical benefit but significantly greater prostatic infarction and greater mid-term prostate shrinkage. The latter may indicate a more durable treatment effect of the smaller microspheres and merits further investigation. However, ischemic complications may be more frequent and more severe with the smaller microspheres.

Acknowledgments

The authors are grateful to Mr. Konstantinos Chouchos, MSc, and Mrs. Maria-Theodora Moschouri, medical student, for their valuable assistance in preparation of this manuscript.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

This work was financially and materially supported by Merit Medical. The authors certify that all sources of support have been clearly identified in the manuscript.

Supplementary Tables 1-3: <https://d2v96fxpocvxx.cloudfront.net/beb8919b-f013-4ea1-b1c8-40332e840fe1/content-images/b037e4c9-9cc6-4bd8-a188-8a86268af6b1.pdf>

References

1. Bamshad D, Sanghvi J, Galla N, et al. Early outcomes of prostatic artery embolization using n-butyl cyanoacrylate liquid embolic agent: a safety and feasibility study. *J Vasc Interv Radiol.* 2024;35(12):1855-1861. [\[Crossref\]](#)
2. Bilhim T, Vasco Costa N, Torres D, Akış S, Alves M, Papoila AL. Comparing embolic particles for prostatic artery embolization to treat lower urinary tract symptoms in patients with benign prostatic hyperplasia. *Eur Radiol.* 2025;35(2):1057-1066. [\[Crossref\]](#)
3. Antunes AA, Carnevale FC, da Motta Leal Filho JM, et al. Clinical, laboratorial, and urodynamic findings of prostatic artery embolization for the treatment of urinary retention related to benign prostatic hyperplasia. A prospective single-center pilot study. *Cardiovasc Intervent Radiol.* 2013;36(4):978-986. [\[Crossref\]](#)
4. Gonçalves OM, Carnevale FC, Moreira AM, Antunes AA, Rodrigues VC, Srougi M. Comparative study using 100-300 versus 300-500 μ m microspheres for symptomatic patients due to enlarged-BPH prostates. *Cardiovasc Intervent Radiol.* 2016;39(10):1372-1378. [\[Crossref\]](#)
5. Ayyagari R, Powell T, Staib L, et al. Prostatic Artery Embolization Using 100-300- μ m trisacryl gelatin microspheres to treat lower urinary tract symptoms attributable to benign prostatic hyperplasia: a single-center outcomes analysis with medium-term follow-up. *J Vasc Interv Radiol.* 2020;31(1):99-107. [\[Crossref\]](#)
6. Torres D, Costa NV, Pisco J, Pinheiro LC, Oliveira AG, Bilhim T. Prostatic artery embolization for benign prostatic hyperplasia: prospective randomized trial of 100-300 μ m versus 300-500 μ m versus 100- to 300- μ m + 300- to 500- μ m TGMs. *J Vasc Interv Radiol.* 2019;30(5):638-644. [\[Crossref\]](#)
7. Xhepa G, Sciacqua LV, Vanzulli A, et al. Prostate artery embolization (PAE) with small beads for the treatment of benign prostatic hyperplasia (BPH). *J Pers Med.* 2024;14(6):613. [\[Crossref\]](#)
8. Yamamoto A, Imai S, Kobatake M, Yamashita T, Tamada T, Umetani K. Evaluation of trisacryl gelatin microsphere embolization with monochromatic X rays: comparison with polyvinyl alcohol particles. *J Vasc Interv Radiol.* 2006;17(11 Pt 1):1797-1802. [\[Crossref\]](#)
9. Senturk C, Cakir V, Yorukoglu K, Yilmaz O, Goktay AY. Looking for the ideal particle: an experimental embolization study. *Cardiovasc Intervent Radiol.* 2010;33(2):336-345. [\[Crossref\]](#)
10. Garcia-Monaco RD, Garategui LG, Onorati MV, Rosasco NM, Peralta OA. Cadaveric specimen study of prostate microvasculature: implications for arterial embolization. *J Vasc Interv Radiol.* 2019;30(9):1471-1479. [\[Crossref\]](#)
11. Sun F, Lucas-Cava V, Sánchez-Margallo FM. Clinical predictive factors in prostatic artery embolization for symptomatic benign prostatic hyperplasia: a comprehensive review. *Transl Androl Urol.* 2020;9(4):1754-1768. [\[Crossref\]](#)
12. Cornelis FH, Bilhim T, Hacking N, Sapoval M, Tapping CR, Carnevale FC. CIRSE standards of practice on prostatic artery embolisation. *Cardiovasc Intervent Radiol.* 2020;43(2):176-185. [\[Crossref\]](#)
13. Carnevale FC, Moreira AM, Antunes AA. The "PerFecTED technique": proximal embolization first, then embolize distal for benign prostatic hyperplasia. *Cardiovasc Intervent Radiol.* 2014;37(6):1602-1605. [\[Crossref\]](#)
14. Filippiadis D, Pereira PL, Hausegger KA, Ryan AG, Binkert CA. CIRSE standards of practice for the classification of complications: the modified CIRSE classification system. *Cardiovasc Intervent Radiol.* 2026;49:2-6. [\[Crossref\]](#)
15. Svarc P, Taudorf M, Nielsen MB, Stroomberg HV, Røder MA, Lönn L. Postembolization syndrome after prostatic artery embolization: a systematic review. *Diagnostics (Basel).* 2020;10(9):659. [\[Crossref\]](#)
16. Moschouris H, Stamatiou K, Malagari K, et al. The value of contrast-enhanced ultrasonography in detection of prostatic infarction after prostatic artery embolization for the treatment of symptomatic benign prostatic hyperplasia. *Diagn Interv Radiol.* 2019;25(2):134-143. [\[Crossref\]](#)
17. Wang MQ, Zhang JL, Xin HN, et al. Comparison of clinical outcomes of prostatic artery embolization with 50- μ m Plus 100- μ m polyvinyl alcohol (PVA) particles versus 100- μ m PVA particles alone: a prospective randomized trial. *J Vasc Interv Radiol.* 2018;29(12):1694-1702. [\[Crossref\]](#)
18. Lin F, Chen Q, Gao M, et al. Retrospective observation of the early efficacy and safety of temperature-sensitive liquid embolic agent combined with polyvinyl alcohol microspheres for prostatic artery embolization in the treatment of lower urinary tract symptoms caused by benign prostatic hyperplasia. *Transl Androl Urol.* 2024;13(9):1847-1858. [\[Crossref\]](#)
19. Bilhim T, Pisco J, Pereira JA, et al. Predictors of clinical outcome after prostate artery embolization with spherical and nonspherical polyvinyl alcohol particles in patients with benign prostatic hyperplasia. *Radiology.* 2016;281(1):289-300. [\[Crossref\]](#)
20. Miller SR, Jernigan SR, Abraham RJ, Buckner GD. Comparing deposition characteristics of various embolic particles using an *in vitro* prostate microvasculature model. *J Vasc Interv Radiol.* 2024;35(10):1534-1543. [\[Crossref\]](#)
21. Lin YT, Amouyal G, Correias JM, Pereira H, et al. Can prostatic arterial embolisation (PAE) reduce the volume of the peripheral zone? MRI evaluation of zonal anatomy and infarction after PAE. *Eur Radiol.* 2016;26(10):3466-3473. [\[Crossref\]](#)
22. Thulasidasan N, Kok HK, Elhage O, Clovis S, Popert R, Sabharwal T. Prostate artery embolisation: an all-comers, single-operator experience in 159 patients with lower urinary tract symptoms, urinary retention, or haematuria with medium-term follow-up. *Clin Radiol.* 2019;74(7):569. [\[Crossref\]](#)
23. Frenk NE, Baroni RH, Carnevale FC, et al. MRI findings after prostatic artery embolization for treatment of benign hyperplasia. *AJR Am J Roentgenol.* 2014;203(4):813-821. [\[Crossref\]](#)
24. Majbar E, Swanberg J, Alamdari F, Peekers R, Chabok A. Early experience of prostate artery embolization: a retrospective single-centre study. *Scand J Urol.* 2025;60. [\[Crossref\]](#)
25. Pisco JM, Rio Tinto H, Campos Pinheiro L, et al. Embolisation of prostatic arteries as treatment of moderate to severe lower urinary symptoms (LUTS) secondary to benign hyperplasia: results of short- and mid-term follow-up. *Eur Radiol.* 2013;23(9):2561-2572. Erratum in: *Eur Radiol.* 2013;23(9):2573-2574. [\[Crossref\]](#)
26. Eldem FG, Atak F, Öcal O, Bozaci AC, Güdeoloğlu A, Peynircioğlu B. Angiographic prostatic arterial anatomy in a Turkish population with benign prostatic hyperplasia. *Turk J Med Sci.* 2021;51(2):518-522. [\[Crossref\]](#)
27. Moschouris H, Stamatiou K, Kornezos I, Kartsouni V, Malagari K. Favorable outcome of conservative management of extensive bladder ischemia complicating prostatic artery embolization. *Cardiovasc Intervent Radiol.* 2018;41(1):191-196. [\[Crossref\]](#)
28. Pisco JM, Bilhim T, Pinheiro LC, et al. Medium- and Long-Term outcome of prostate artery embolization for patients with benign prostatic hyperplasia: results in 630 Patients. *J Vasc Interv Radiol.* 2016;27(8):1115-1122. [\[Crossref\]](#)
29. Nicholson TA, Pelage JP, Ettles DF. Fibroid calcification after uterine artery embolization: ultrasonographic appearance and pathology. *J Vasc Interv Radiol.* 2001;12(4):443-446. [\[Crossref\]](#)



Post-cryoablation magnetic resonance imaging features of desmoid tumors: a pictorial essay

 Leonor G. Savarese¹

 Nicolas Papalexis²

 Marco Miceli²

 Michela Carta²

 Giancarlo Facchini²

¹University of Sao Paulo, Ribeirao Preto Medical School, Department of Medical Imaging, Hematology and Clinical Oncology, Ribeirão, Brazil

²IRCCS Istituto Ortopedico Rizzoli, Diagnostic and Interventional Radiology Unit, Bologna, Italy

ABSTRACT

Desmoid tumors (DTs) are rare, locally aggressive soft tissue neoplasms with highly variable clinical behavior. Although benign, their infiltrative nature can lead to considerable morbidity. DTs present a major challenge due to their unpredictable behavior and potential for misdiagnosis. Recently, there has been a large shift in the treatment strategy for DTs, and the number of cases being followed up with imaging has increased. Cryoablation has emerged as a minimally invasive treatment option, yet post-procedural imaging features remain poorly characterized. This study illustrates the magnetic resonance imaging (MRI) evolution of DTs following percutaneous cryoablation, emphasizing key patterns across pre-treatment, intra-procedural, and follow-up studies. The available MRI follow-up ranged from 6 to 30 months. Through a pictorial approach, we aimed to provide radiologists with practical insights to avoid misinterpretation of expected post-treatment changes as recurrence and to improve clinical management.

KEYWORDS

Desmoid tumor, magnetic resonance imaging, follow-up studies, soft tissue neoplasm, ablation technique, cryoablation, radiology, interventional

Desmoid-type fibromatosis, also known as desmoid tumor (DT) or aggressive fibromatosis, is a rare, locally invasive fibroblastic proliferation characterized by an unpredictable clinical course.^{1,2} These tumors account for < 3% of soft tissue neoplasms, with an incidence of approximately 3–5 cases per million annually. They typically affect individuals in their third to fourth decades of life and show a higher prevalence in women. Most cases are sporadic and arise in extra-abdominal locations, particularly the extremities and trunk.³

DTs pose a major diagnostic and management challenge due to their variable behavior, which can range from rapid growth to spontaneous regression. Recent guidelines emphasize an initial surveillance strategy to prevent overtreatment, reserving intervention for symptomatic or progressive cases. However, a decision to initiate therapy earlier may be considered when DTs are located near critical structures, as there may be a higher risk of morbidity before the disease stabilizes. When treatment is warranted, options include systemic therapies (such as chemotherapy or tyrosine kinase inhibitors) and local interventions.^{4,5}

Cryoablation has emerged as a promising local therapy and is now included in updated management guidelines. This minimally invasive technique employs cryoprobes to induce tumor cell death through extreme cold, achieved via the Joule–Thomson effect. Cellular destruction results from intracellular ice formation and vascular injury, leading to ischemia. Image-guided cryoablation, typically performed under computed tomography guidance, allows precise visualization of a low-attenuation ice ball in soft tissues, providing visual confirmation that the tumor is included in the cryoablated volume.⁶ Recent studies have shown that image-guided percutaneous cryoablation appears to be safe and effective for local control in

Handling editor: Bahri Üstünsöz

Corresponding author: Leonor G. Savarese

E-mail: lsavarese@hcrp.usp.br

Received 10 June 2025; revision requested 21 July 2025;
last revision received 05 October 2025; accepted 24
October 2025.



Epub: 08.12.2025

Publication date: 01.09.2026

DOI: 10.4274/dir.2025.253499

You may cite this article as: Savarese LG, Papalexis N, Miceli M, Carta M, Facchini G. Post-cryoablation magnetic resonance imaging features of desmoid tumors: a pictorial essay. *Diagn Interv Radiol.* 2026;32(5):709-714.

patients with extra-abdominal DTs.⁷⁻¹¹ In this pictorial essay, we illustrate six representative cases of extra-abdominal DTs treated with percutaneous cryoablation, including three men and three women (age range 33–54 years). Tumor sites comprise the gluteal region (n = 2), lumbar wall (n = 1), abdominal wall (n = 1), scapular region (n = 1), and calf (n = 1) (Table 1).

Role of magnetic resonance imaging

Magnetic resonance imaging (MRI) plays a critical role in the diagnosis, treatment planning, and follow-up of DTs. Imaging findings correlate with histopathologic features; hyperintense areas on T2-weighted and enhancing sequences reflect high cellularity, whereas hypointense, non-enhancing regions correspond to fibrous, low-cellularity components.^{12,13} An increase in collagenization within the tumor is associated with a decrease in cellularity, suggesting a reduction in tumor activity. This can be seen on MRI as a decrease in enhancement and T2 hyperintensity, which are imaging features indicating a positive response to therapy.¹⁴

Post-cryoablation imaging findings

Post-cryoablation MRI findings evolve. In early follow-up, considerable edema may be noted in the cryoablation area and adjacent soft tissues (Figure 1). Early imaging signs after the procedure include a non-enhancing ablation cavity with a high T2 signal, corresponding to coagulative necrosis. A thin inner rim of low T2 signal, indicative of hemosiderin rim (hemorrhagic congestion), can be better visualized on T2 gradient echo sequences. A thicker outer rim of intermediate/low T2 signal with post-contrast enhancement corresponds to granulation tissue/vascular fibrosis (Figure 2).

In subsequent follow-ups, the non-enhancing ablation cavity shows a size reduc-

tion and may show a marked decrease in T2 signal intensity, indicative of dense fibrosis. Additionally, a progressive reduction in post-contrast enhancement of the outer rim of granulation tissue/vascular fibrosis can be noted (Figure 3). Finally, a reduction in lesion volume is expected in successful cases (Figure 4), with a progressive replacement by adipose tissue observed in subcutaneous cases (Figure 5). If tumor recurrence occurs, it can be observed on post-contrast sequences as a nodular enhancing area (Figure 2). Imaging findings are summarized in Table 2.

The success rate of cryoablation is directly related to the ice ball encompassing the entire lesion during the procedure. However, cryoablation may induce the “abscopal effect,” leading to the reduction or disappearance of the tumor in non-frozen sites, possibly due to immune modulation (Figure 6). This phenomenon is of high importance in tumor therapy, as it effectively reduces the likelihood of tumor recurrence.¹⁵

Diffusion-weighted imaging (DWI) correlates with the content of the treated tissue, showing higher apparent diffusion coefficient (ADC) values in cases of low cellularity and fibrous content. It can also be used to detect changes in the treated tissue post-cryoablation. A recent article¹⁶ reported low ADC values (mean $0.90 \times 10^3 \text{ mm}^2/\text{s}$) in early follow-up of treated tissue after cryotherapy, possibly due to tumor necrosis. This should not be confused with recurrence, especially when no enhancement is present. Further studies are needed to confirm DWI patterns

in treated tissue post-cryoablation. However, this pattern was also observed in some cases in this study, as shown in Figure 7.

Discussion

The management of DTs has become increasingly individualized, tailored around patients’ needs and lesions’ features. Aggressive surgical approaches are generally avoided in favor of more conservative, minimally invasive strategies.¹⁷ The current consensus has shifted towards a “watch-and-wait” first-line approach, and intervention is performed for progressive or symptomatic cases.^{11,17} In this context, image-guided percutaneous cryoablation has emerged as a valuable and minimally invasive alternative to surgery and systemic therapy.¹⁸

Recent studies have shown encouraging outcomes in terms of local control and symptom relief, especially when complete tumor coverage is achieved during ablation.¹⁹ This, combined with minimal side effects, has led to its inclusion in recent international guidelines.^{11,17-19} Furthermore, the potential for cryoablation to induce systemic antitumor effects—the so-called abscopal effect—suggests opportunities for synergy with systemic therapies, including immunotherapy.²⁰

Despite these advancements, standardized radiologic criteria for treatment response remain lacking. MRI features, such as transient T2 hyperintensity or DWI changes, may mimic recurrence, especially in early follow-up, if not properly contextualized.²¹

Table 1. Patients’ demographics and characteristics

Case	Age (years)	Sex	Tumor location	Magnetic resonance imaging follow-up duration (months)
1	35	Male	Right gluteal region	18
2	54	Female	Right lumbar subcutaneous region	30
3	48	Female	Left abdominal wall	8
4	33	Male	Left gluteal subcutaneous region	25
5	50	Female	Left scapular region	18
6	33	Female	Calf	6

Table 2. Key magnetic resonance imaging findings after cryoablation of desmoid tumors

Imaging feature	Definition/description
Non-enhancing ablation cavity	Central zone without enhancement, corresponding to coagulative necrosis
Hemosiderin rim	Thin inner rim of low T2 signal due to hemorrhagic congestion
Granulation tissue/vascular fibrosis	Outer rim with intermediate/high T2 signal and post-contrast enhancement
Progressive fibrosis	Reduction in T2 signal and enhancement over time
Nodular enhancement	New or growing enhancing nodule at cavity margin, suspicious for recurrence

Main points

- Magnetic resonance imaging (MRI) is fundamental in the post-cryoablation follow-up of desmoid tumors, enabling precise differentiation between expected treatment-related changes and tumor recurrence.
- Key imaging findings after cryoablation include a central non-enhancing cavity, peripheral granulation tissue with contrast enhancement, and progressive fibrosis reflected by decreasing T2 signal intensity.
- Recognizing the temporal evolution of these MRI features is essential to avoid misinterpretation and to guide appropriate clinical decision-making.

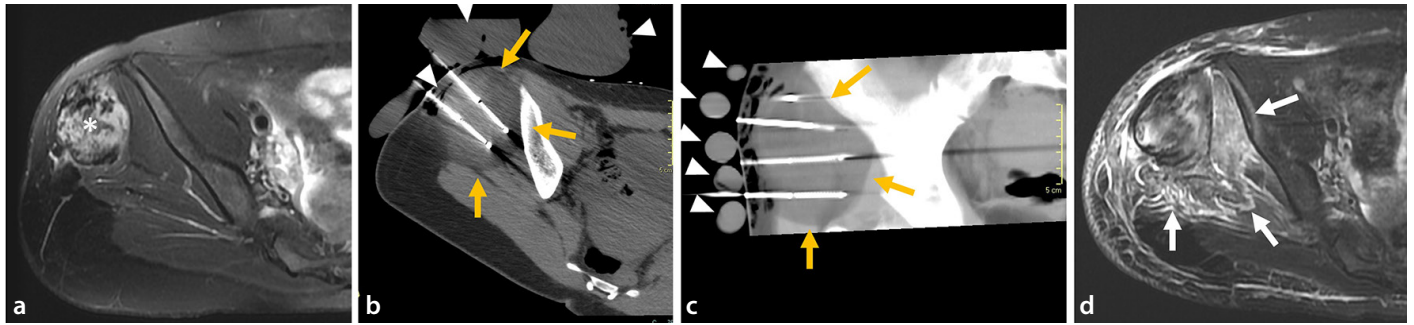


Figure 1. A 35-year-old man. Axial T2 fat-sat imaging (a) reveals a desmoid tumor in the right gluteal region (asterisk). A computed tomography scan taken during the procedure, axial (b) and reconstructed coronal (c) views, shows cryoprobes placed within the lesion, with the ice ball (yellow arrows) covering the entire tumor as well as a safety margin of healthy tissue. In the early follow-up, 1 week post-procedure, considerable edema was noted in the cryoablation area and adjacent soft tissues. A well-demarcated zone of edema (white arrows in d) that extends beyond the lesion's margins, encompassing the gluteal muscles and the iliac bone, corresponds to the area covered by the ice ball. Note the thermoprotection of the skin, achieved by using sterile gloves filled with warmed saline and a subcutaneous injection of air (arrowheads in b, c).

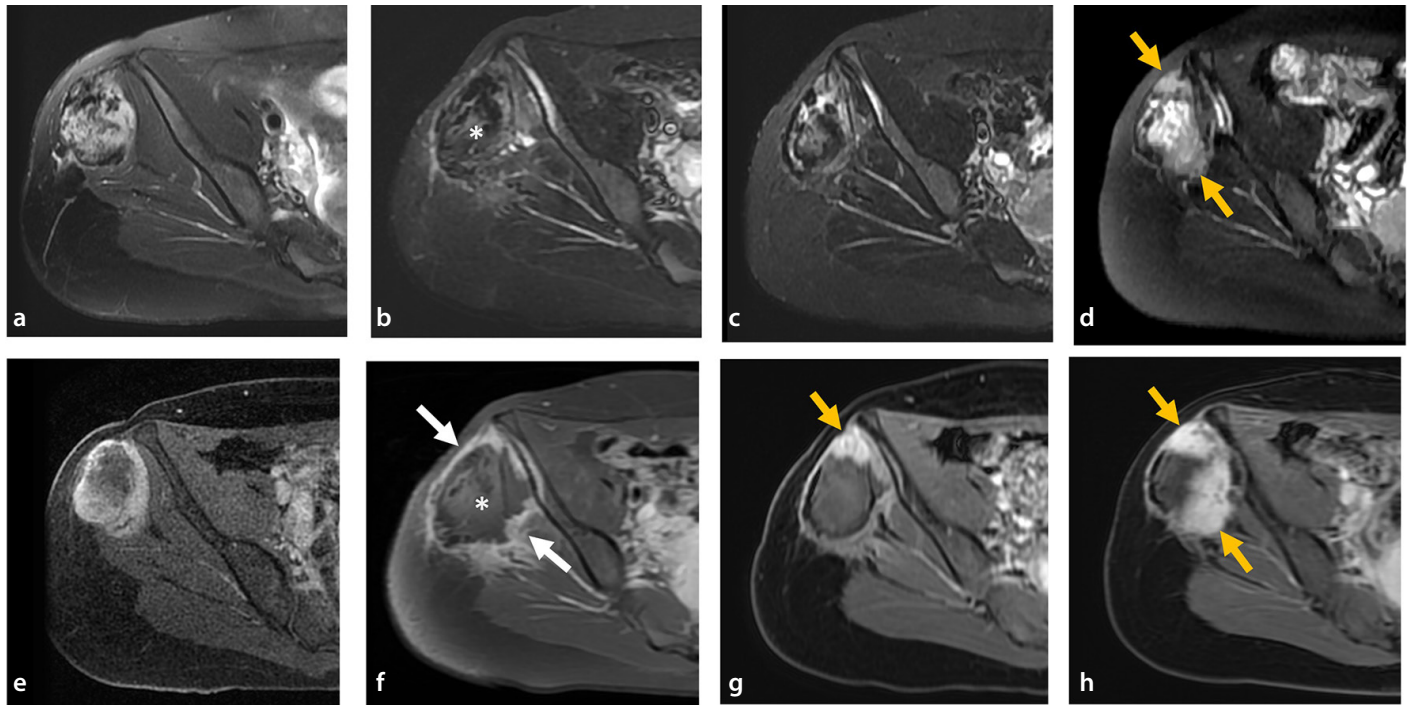


Figure 2. A 35-year-old man. Axial T2 fat-sat/short tau inversion recovery (STIR) (a-d) and post-contrast T1 fat-sat images (e-h) of a desmoid tumor in the right gluteal region, pre- and post-cryoablation, across different time frames (a, e: pre-treatment; b, f: 3 months; c, g: 9 months; d, h: 18 months). At the 3-month follow-up, a non-enhancing ablation cavity corresponding to the ablation cavity was observed (asterisk), displaying a markedly low signal on T2 fat-sat/STIR, suggestive of dense fibrosis, and surrounded by granulation tissue/vascular fibrosis with post-contrast enhancement (white arrows). The 9-month follow-up revealed a potential recurrence at the anterior margin of the ablation cavity, more clearly delineated on the post-contrast T1 fat-sat sequence as a nodular enhancing area (yellow arrow). By the 18-month follow-up, this enhancing nodular lesion had increased in size, extending to the medial and posterior margins, confirming the recurrence (yellow arrows).

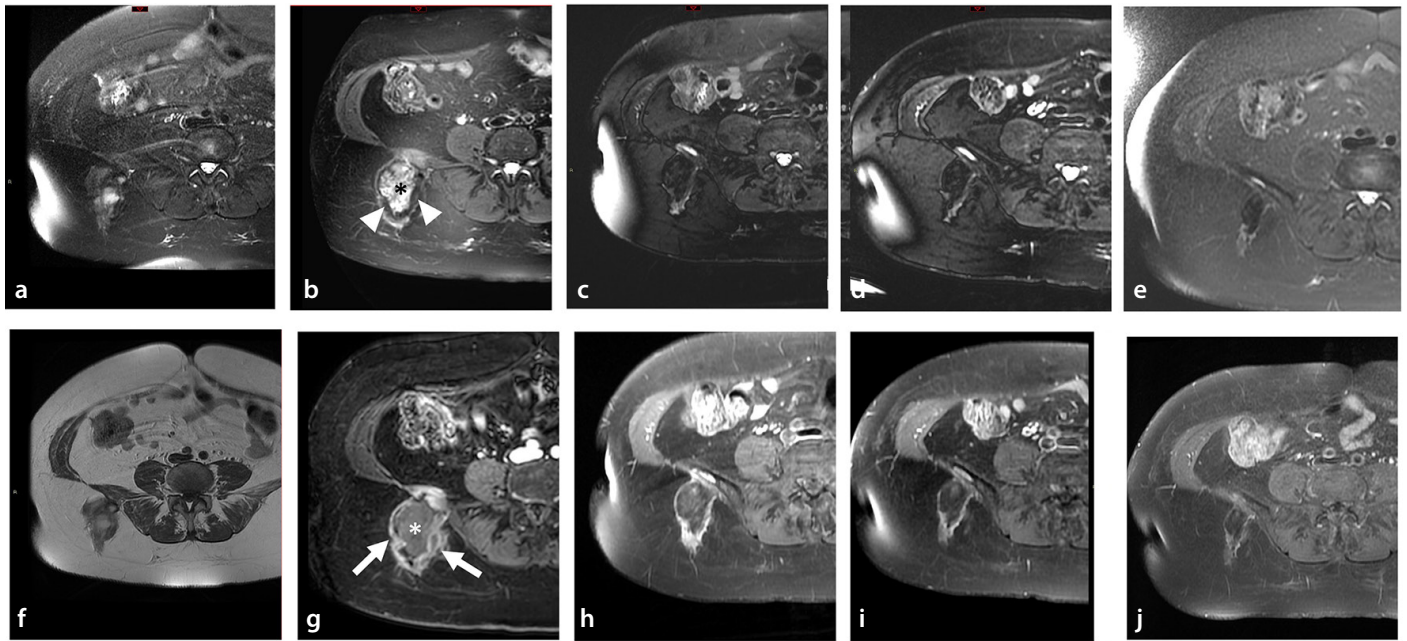


Figure 3. A 54-year-old woman. Axial T2 fat-sat (a-e) and post-contrast T1 (f) and T1 fat-sat (g-j) imaging of a subcutaneous desmoid tumor in the right lumbar area, pre- and post-cryoablation, at different time frames (a, f: pre-treatment; b, g: 2 months; c, h: 7 months; d, i: 13 months; e, j: 30 months). In the 2 months post-procedure images, note a non-enhancing ablation cavity with high T2 signal at the lesion's center, indicative of coagulative necrosis (asterisk). It is surrounded by an inner thin rim of low T2 signal (arrowheads) corresponding to a hemosiderin rim (hemorrhagic congestion) and an outer rim of high T2 signal and post-contrast enhancement (arrows) corresponding to granulation tissue/vascular fibrosis. In subsequent follow-ups, the non-enhancing ablation cavity exhibits a reduction in size and a marked decrease in T2 signal intensity, indicative of dense fibrosis. A progressive reduction in the post-contrast enhancement of the outer rim of granulation tissue/vascular fibrosis can also be seen.

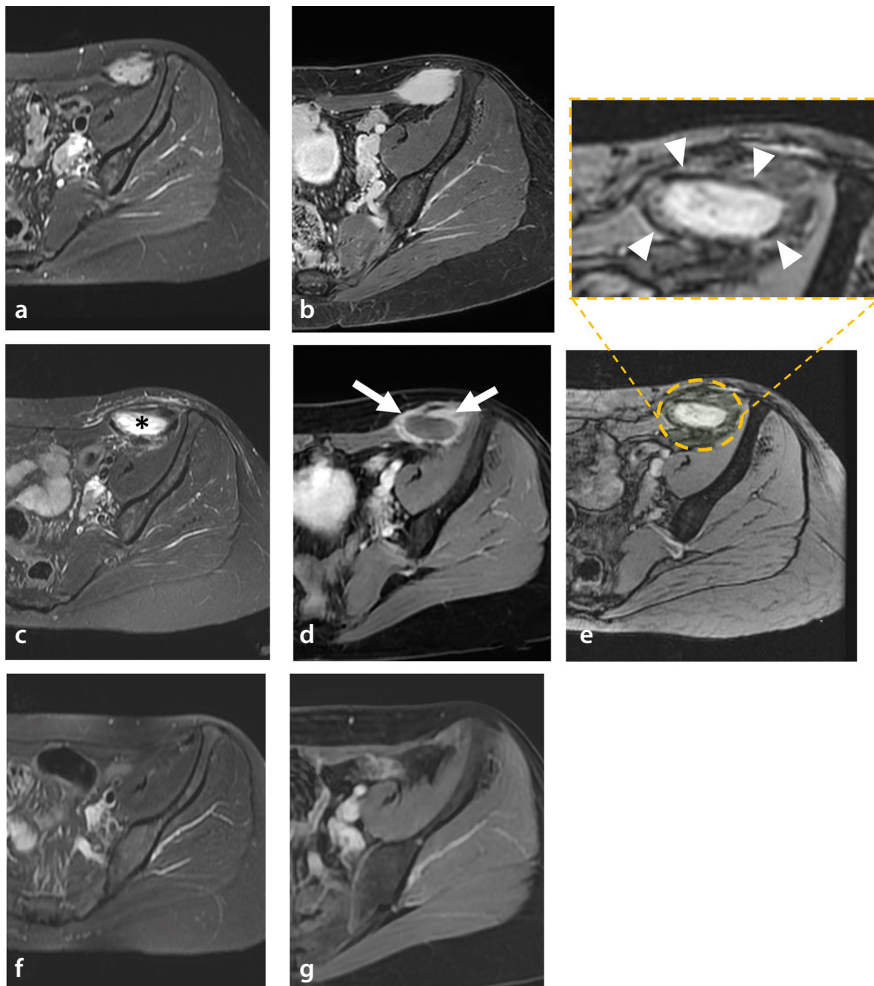


Figure 4. A 48-year-old woman. Axial T2 with fat-sat (a-c), axial T1 fat-sat post-contrast (d-f), and T2 gradient echo (g) of a left abdominal wall desmoid tumor, pre- and post-cryoablation, at different time frames (a, d: pre-treatment; b, e, g: 1 month; c, f: 8 months). In the images taken 1 month post-procedure, note a non-enhancing ablation cavity with high T2 signal corresponding to coagulative necrosis (asterisk), showing the same pattern of ablation cavity and surrounding rims as previously described, with progressive decrease in enhancement over time. A thin inner rim of low T2 signal, corresponding to hemosiderin rim (hemorrhagic congestion), can be better identified on the T2 gradient echo sequence (arrowheads—zoomed image). A thick outer rim of intermediate/low T2 signal and post-contrast enhancement (arrows) corresponds to granulation tissue/vascular fibrosis. In subsequent follow-ups, complete resolution of the lesion was observed, with only minor residual fibro-cicatricial changes.

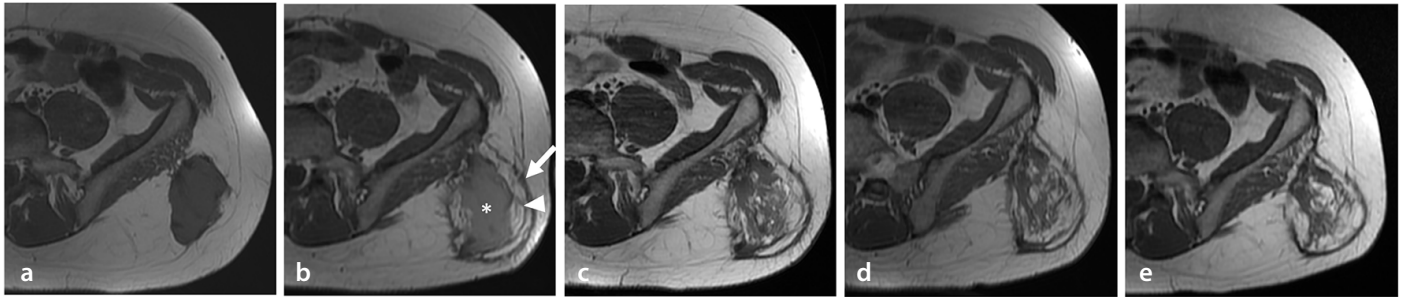


Figure 5. A 33-year-old man. Axial T1 imaging of a subcutaneous desmoid tumor in the left gluteal region, at different time frames (a: pre-treatment; b: 6 months; c: 11 months; d: 17 months; e: 25 months). In the 6-month post-procedure images, note the ablation cavity encompassing the entire lesion, with a high T1 signal indicative of coagulative necrosis (asterisk). In this case, it is surrounded by an outer rim of intermediate T1 signal (arrow), corresponding to granulation tissue/vascular fibrosis, and an inner rim of fat tissue (arrowhead). In subsequent follow-ups, a better definition of the outer rim of granulation tissue/vascular fibrosis and a reduction in the dimensions of the ablation cavity are evident; these are progressively replaced by adipose tissue.

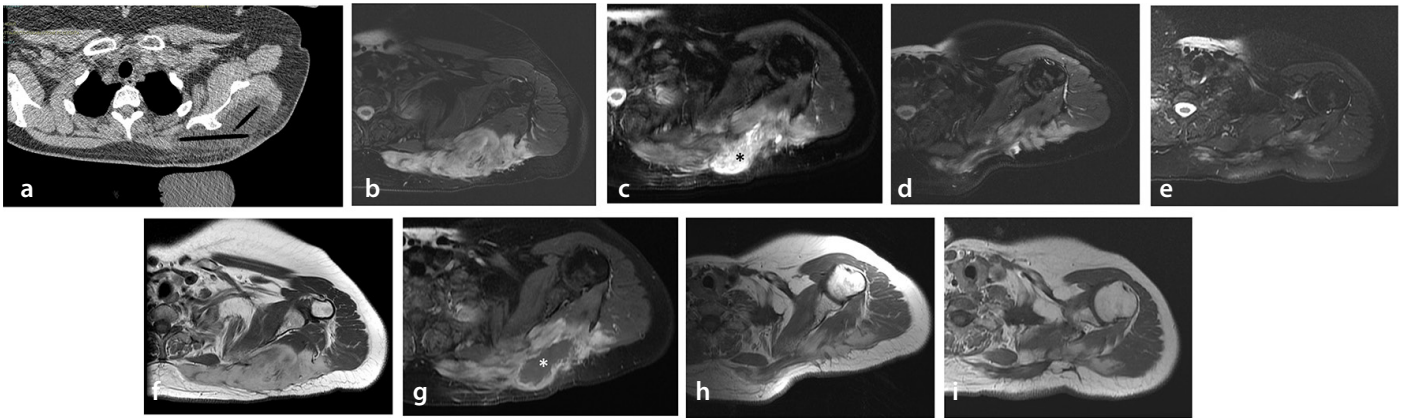


Figure 6. A 50-year-old woman. Axial T2 fat-sat (b, c, d, e) and post-contrast axial T1 with and without fat-sat imaging (f-i) have revealed a large, infiltrative desmoid tumor affecting the musculature and subcutaneous tissue of the left scapular region, noted for its progressive growth in control studies over the past 2 years. A computed tomography scan, performed immediately after the removal of the cryoprobes (a), displays the ice ball covering portions of the tumor. Images (b) and (f) show the pre-treatment images. At the 3-month post-procedure follow-up (c, g), a non-enhancing ablation cavity (asterisk) with high T2 signal was noted, indicative of coagulative necrosis, and a significant volume of tumor not included in the ablation zone. By the 10- (d, h) and 18-month (e, i) follow-ups, there was a notable reduction in the overall size of the lesion, including areas that were not directly ablated, possibly due to an abscopal effect of the cryoablation.

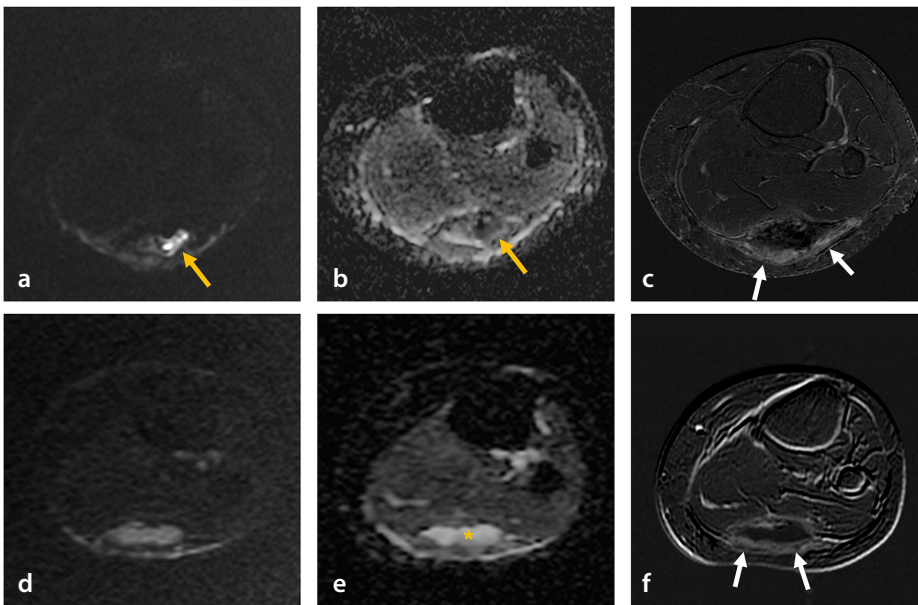


Figure 7. A 33-year-old woman. Axial diffusion-weighted imaging (a, d), apparent diffusion coefficient (ADC) maps (b, e), and post-contrast subtraction images (c, f) of a desmoid tumor in the calf in different time frames post-cryoablation (a-c: 3 months; d-f: 6 months). In early follow-up after cryotherapy, there may be areas of low ADC values, possibly due to tumor necrosis (yellow arrows in a and b; mean $0.8 \times 10^{-3} \text{ mm}^2/\text{s}$), which should not be confused with recurrence, especially when no enhancement is present. The observed peripheral enhancement (white arrows in c and f) corresponds to granulation tissue/vascular fibrosis. In the follow-up control, an increase in ADC values is observed (asterisk in e; mean $2.8 \times 10^{-3} \text{ mm}^2/\text{s}$).

A deeper understanding of the temporal evolution of these features is essential to avoid misinterpretation and guide clinical decision-making.

Emerging technologies, including artificial intelligence and radiomics, may offer tools to differentiate residual tumor from post-ablation changes. Recent studies have shown that deep learning applied to baseline MRI can predict clinical progression of DTs with high accuracy (up to 93%), suggesting a potential role for artificial intelligence-based tools in risk stratification and treatment planning.²² Finally, growing interest in cryoablation's immunomodulatory potential may support future therapeutic combinations that enhance outcomes in patients with DTs.

In conclusion, our understanding of DTs is evolving rapidly. MRI plays a central role in evaluating post-cryoablation changes and guiding patient management. DTs treated with percutaneous cryoablation exhibit a predictable sequence of imaging changes over time. Key post-treatment MRI findings include a central non-enhancing cavity of coagulative necrosis, peripheral granulation tissue with enhancement, and progressive fibrosis characterized by decreasing T2 signal and enhancement. Accurate recognition of these patterns is critical to distinguish treatment response from recurrence. Familiarity with expected imaging patterns and their timeline can help avoid misdiagnosis of recurrence and optimize treatment strategies. Future research integrating imaging biomarkers, artificial intelligence, and immunologic response will likely further refine our approach to this complex entity.

Footnotes

Conflict of interest disclosure

The authors declared that they have no conflict of interest.

References

1. Sbaraglia M, Bellan E, Dei Tos AP. The 2020 WHO Classification of Soft Tissue Tumours: news and perspectives. *Pathologica*. 2021;113(2):70-84. [\[Crossref\]](#)
2. WHO Classification of Tumours Editorial Board. WHO Classification of Tumours: Soft Tissue and Bone Tumours. 5th ed. Vol 3 [Internet]. Lyon: International Agency for Research on Cancer; 2020 [cited 2025 Nov 6]. [\[Crossref\]](#)
3. Kasper B, Baumgarten C, Garcia J, et al. An update on the management of sporadic desmoid-type fibromatosis: a European Consensus Initiative between Sarcoma Patients EuroNet (SPAEN) and European Organization for Research and Treatment of Cancer (EORTC)/Soft Tissue and Bone Sarcoma Group (STBSG). *Ann Oncol*. 2017;28(10):2399-2408. [\[Crossref\]](#)
4. Gronchi A, Colombo C, Le Péchoux C, et al. Sporadic desmoid-type fibromatosis: a stepwise approach to a non-metastasising neoplasm—a position paper from the Italian and the French Sarcoma Group. *Ann Oncol*. 2014;25(3):578-583. [\[Crossref\]](#)
5. Desmoid Tumor Working Group. The management of desmoid tumours: a joint global consensus-based guideline approach for adult and paediatric patients. *Eur J Cancer*. 2020;127:96-107. [\[Crossref\]](#)
6. Papalexis N, Savarese LG, Peta G, et al. The new ice age of musculoskeletal intervention: role of percutaneous cryoablation in bone and soft tissue tumors. *Curr Oncol*. 2023;30(7):6744-6770. [\[Crossref\]](#)
7. Auloge P, Garnon J, Robinson JM, et al. Percutaneous cryoablation for advanced and refractory extra-abdominal desmoid tumors. *Int J Clin Oncol*. 2021;26(6):1147-1158. [\[Crossref\]](#)
8. Yan YY, Walsh JP, Munk PL, et al. A single-center 10-year retrospective analysis of cryoablation for the management of desmoid tumors. *J Vasc Interv Radiol*. 2021;32(9):1277-1287. [\[Crossref\]](#)
9. Efrima B, Ovadia J, Drukman I, et al. Cryosurgery for symptomatic extra-abdominal desmoids. A proof of concept study. *J Surg Oncol*. 2021;124(4):627-634. [\[Crossref\]](#)
10. Saltiel S, Bize PE, Goetti P, et al. Cryoablation of extra-abdominal desmoid tumors: a single-center experience with literature review. *Diagnostics (Basel)*. 2020;10(8):556. [\[Crossref\]](#)
11. Kurtz JE, Buy X, Deschamps F, et al. CRYODESMO-O1: a prospective, open phase II study of cryoablation in desmoid tumour patients progressing after medical treatment. *Eur J Cancer*. 2021;143:78-87. [\[Crossref\]](#)
12. Bouhamama A, Lame F, Mastier C, et al. Local control and analgesic efficacy of percutaneous cryoablation for desmoid tumors. *Cardiovasc Intervent Radiol*. 2020;43(1):110-119. [\[Crossref\]](#)
13. Shinagare AB, Ramaiya NH, Jagannathan JP, et al. A to Z of desmoid tumors. *AJR Am J Roentgenol*. 2011;197(6):W1008-1014. [\[Crossref\]](#)
14. Ganeshan D, Amini B, Nikolaidis P, Assing M, Vikram R. Current update on desmoid fibromatosis. *J Comput Assist Tomogr*. 2019;43(1):29-38. [\[Crossref\]](#)
15. Okuda M, Yoshida K, Kobayashi S, Gabata T. Desmoid-type fibromatosis: imaging features and course. *Skeletal Radiol*. 2023;52(7):1293-1303. [\[Crossref\]](#)
16. Chen J, Qian W, Mu F, Niu L, Du D, Xu K. The future of cryoablation: an abscopal effect. *Cryobiology*. 2020;97:1-4. [\[Crossref\]](#)
17. Garnon J, Cazzato RL, Autrusseau PA, et al. Desmoid fibromatosis: interventional radiology (sometimes) to the rescue for an atypical disease. *Br J Radiol*. 2025;98(1170):840-850. [\[Crossref\]](#)
18. Cazzato RL, Gantzer J, de Marini P, et al. Sporadic desmoid tumours: systematic review with reflection on the role of cryoablation. *Cardiovasc Intervent Radiol*. 2022;45(5):613-621. [\[Crossref\]](#)
19. Kasper B, Baldini EH, Bonvalot S, et al. Current management of desmoid tumors: a review. *JAMA Oncol*. 2024;10(8):1121-1128. [\[Crossref\]](#)
20. Yang X, Gao X, Xu C, et al. Cryoablation synergizes with anti-PD-1 immunotherapy induces an effective abscopal effect in murine model of cervical cancer. *Transl Oncol*. 2025;51:102175. [\[Crossref\]](#)
21. Dux DM, Chodakiewitz Y, Bitton R, et al. MRI features that predict progression of residual disease after ablation of extra-abdominal desmoid fibromatosis. *Eur Radiol*. 2025;35(7):4161-4170. [\[Crossref\]](#)
22. Fares R, Atlan LD, Druckmann I, et al. Imaging-based deep learning for predicting desmoid tumor progression. *J Imaging*. 2024;10(5):122. [\[Crossref\]](#)



How may complementary magnetic resonance imaging findings facilitate the diagnosis of inflammatory hand arthritis involving the distal interphalangeal joint? A prospective cohort study

Yasin Yaraşır¹
Gizem Ayan²
Hanife Avcı³
Levent Kılıç²
Üstün Aydingöz¹
Umut Kalyoncu²
Adalet Elçin Yıldız¹

¹Hacettepe University Faculty of Medicine,
Department of Radiology, Ankara, Türkiye

²Hacettepe University Faculty of Medicine,
Department of Rheumatology, Ankara, Türkiye

³Hacettepe University Faculty of Medicine,
Department of Biostatistics, Ankara, Türkiye

The findings of the study were partially presented at the annual European Congress of Radiology 2025 (ECR 2025) in Vienna, Austria.

Handling editor: Zeynep Maraş Özdemir

Corresponding author: Adalet Elçin Yıldız

E-mail: aelcindr@gmail.com

Received 10 June 2025; revision requested 12 July 2025;
accepted 01 September 2025.



Epub: 06.10.2025

Publication date: 01.09.2026

DOI: 10.4274/dir.2025.253502

PURPOSE

In this study, we aimed to characterize the role of magnetic resonance imaging (MRI) in making a specific diagnosis of inflammatory hand arthritis (IHA), particularly in early stages or ambiguous cases.

METHODS

Patients aged ≥ 18 years with suspicious IHA in at least one joint were enrolled in this single-center prospective study. Three Tesla MRI (3T-MRI) with a fine-tuned protocol was utilized, whereby differential diagnoses were made by radiologists according to the predominant involvement of synovium or synovioentheseal complex (SEC) and/or specific degenerative findings. Physical examination, laboratory findings, treatment response, and already-established classification criteria were used to reach the final diagnostic groups: psoriatic arthritis (PsA), rheumatoid arthritis (RA), erosive osteoarthritis or calcium pyrophosphate dihydrate deposition disease (EOA/CPPD), and arthritis with distal interphalangeal joint (DIPJ) involvement (ADIPI) not otherwise classified into any group. Statistical analyses mainly included pairwise comparisons of MRI findings across diagnostic groups.

RESULTS

Of 80 patients enrolled, 57 [42 women; mean age, 54 years (range, 28–79 years)] constituted the final group with eventual clinical diagnoses of 11 PsA, 14 RA, 11 EOA/CPPD, and 21 ADIPI. MRI revealed no difference between the PsA and ADIPI groups, except for nailbed enthesitis ($P = 0.048$, effect size: 0.416). A comparison between PsA and RA revealed that enthesitis, excluding pulley enthesitis, was more frequently observed in PsA ($P = 0.033$, effect size: 0.497). Periarticular soft tissue edema was also more common in PsA than RA ($P = 0.042$, effect size: 0.461). When the ADIPI and PsA groups were combined, enthesitis and periarticular soft tissue edema were more common than in other groups ($P < 0.001$).

CONCLUSION

SEC inflammation and periarticular edema on MRI strongly predict PsA, especially in patients with DIPJ arthritis who do not meet rheumatological classification criteria.

CLINICAL SIGNIFICANCE

3T-MRI with a fine-tuned protocol enables a more accurate differential diagnosis of hand inflammatory arthritis, potentially guiding earlier and more targeted interventions.

KEYWORDS

Arthritis, psoriatic, arthritis, rheumatoid, magnetic resonance imaging, enthesitis, hand

Inflammatory hand arthritis (IHA) is a diagnostic challenge, particularly in the early stages, where clinical and radiographic findings may be non-specific. Although rheumatoid arthritis (RA) and psoriatic arthritis (PsA) share overlapping features, their distinct treatment approaches necessitate accurate differentiation.¹ However, limitations in diagnostic criteria lead to diagnostic uncertainty, resulting in delays that can worsen outcomes.²

Despite advances in imaging, the literature remains limited in identifying reliable imaging biomarkers to distinguish inflammatory arthritis subtypes at early presentation.³ Traditional approaches rely on radiographic assessment of joint distribution, specific deformities, and bony changes, such as periostitis or erosions, which typically become apparent only in later stages.^{4,5} However, in earlier phases, the predominant site of inflammation, whether at the synovium or the enthesis, becomes critical for accurate diagnosis. Imaging microanatomical regions, such as the synovioentheseal complex (SEC), offers a potential alternative, particularly in diagnostically ambiguous cases.^{1,6}

Notably, in 10%–15% of PsA cases, musculoskeletal symptoms may precede overt skin manifestations—a form referred to as “PsA *sine psoriase*.”⁷ Among the distinctive patterns of involvement observed in these patients, distal interphalangeal joint (DIPJ) arthritis stands out as a key feature.^{8,9} To date, few studies have been conducted on the role of imaging in arthritis cases involving the DIPJ without skin psoriasis or not satisfying the classification criteria, including the Classification Criteria for Psoriatic Arthritis (CASPAR). In our study, we identify a final diagnostic subgroup of patients who exhibited clinical features suggestive of PsA with DIPJ involvement (ADIPI) but did not meet CASPAR. These patients were categorized as

having arthritis with ADIPI, a term we use to describe cases that may represent “PsA *sine psoriase*” or remain unclassifiable despite clinical evidence of inflammatory DIPJ arthritis.

This study aims to assess the diagnostic utility of a tailored 3 Tesla magnetic resonance imaging (3T-MRI) protocol for IHA, with particular emphasis on SEC inflammation and related features. We hypothesized that 3T-MRI can reliably differentiate PsA from RA, erosive osteoarthritis (EOA), and crystal arthropathies, even in diagnostically challenging scenarios, by highlighting distinct inflammatory patterns. To test this, we conducted a prospective study using a fine-tuned MRI protocol on a cohort of 57 patients with untreated IHA.

Methods

This study was approved by the Hacettepe University Clinical Studies Ethics Committee (registration number: KA-21113, approval date: December 29, 2021) and conducted per the Declaration of Helsinki and Good Clinical Practice guidelines. Written and verbal informed consent were obtained from all participants.

Study design and patient enrollment

This study was conducted at Hacettepe University Hospitals between September 1, 2021, and June 1, 2024. Patients presenting to the rheumatology outpatient clinic with untreated IHA were considered for recruitment. Although the aim was to include a broad spectrum of inflammatory arthritis cases, practical limitations and study-specific objectives resulted in a partial representation of certain disease subtypes.

The inclusion criteria were adults ≥ 18 years of age with clinical, laboratory, or ultrasonographic findings suggestive of inflammatory arthritis in at least one hand joint. Patients with absolute contraindications for MRI, such as the presence of metallic implants or devices incompatible with MRI or severe claustrophobia, were excluded. Similarly, patients with a known allergy to gadolinium-based contrast agents or impaired renal function contraindicating gadolinium use (e.g., estimated glomerular filtration rate < 30 mL/min/1.73 m²) were also excluded. Given the study's focus on diagnostic utility in differential diagnosis, patients diagnosed with other diseases outside the study's scope (e.g., infectious arthritis) and those without evidence of inflammation on MRI were excluded to main-

tain the study's reliability and focus (Figure 1).

Notably, although the cohort was consecutive in terms of cases with ADIPI, the inclusion of other disease types, such as RA and PsA, was influenced by the study's specific objectives and the availability of MRI requisitions. This situation highlights how patients with ADIPI constituted the majority of MRI referrals in this study, despite the higher prevalence of RA and PsA in general clinical practice.

Imaging (hand radiography and magnetic resonance imaging) protocol and procedure

Hand radiographs were obtained in an anterior–posterior projection. MRI of the symptomatic hand was performed after the patient was positioned supine with their arms by their sides. T1-weighted axial and coronal spin echo (SE) images were obtained to evaluate the hand anatomy and joint structures. T2-weighted fat-saturated (FS) SE and coronal short-tau inversion recovery sequences were acquired to show edematous areas and fluid in the joint space. Proton density-weighted coronal FS SE sequences were obtained to evaluate the ligaments, especially the collateral ligaments. Axial and coronal FS T1-weighted SE images were obtained 5–10 minutes after intravenous contrast media injection to show periarticular soft tissue inflammation and osteitis. A three-dimensional (3D) zero echo time sequence was included in the protocol to create computed tomography-like images, confirming erosions and evaluating osteophytes or new bone formation more clearly. The post-contrast 3D fast spoiled gradient recalled echo sequence, which provides higher contrast resolution, was also used. The total duration of the examination was 50 minutes (Table 1). A gadolinium-based contrast agent (Gadovist® 0.1 mmol/kg, Bayer Pharmaceuticals) was used in all patients.

Evaluation strategy for index test and reference standard

Two radiologists (Y.Y., a 5th-year radiology resident, and A.E.Y., a radiology consultant with 6 years of dedicated musculoskeletal radiology experience) performed the MRI assessment and reached a consensus using the hospital's picture archiving and communication system. Radiologists were aware of the symptomatic hand and that the patient was being evaluated for inflammatory arthritis; however, they were blinded to the final

Main points

- Differentiating inflammatory arthritides in early stages or with atypical presentations remains challenging, especially in cases that do not fulfill the classification criteria.
- Synovioentheseal complex involvement and periarticular edema can be used as a diagnostic imaging marker for psoriatic arthritis in ambiguous cases.
- Three Tesla magnetic resonance imaging with a fine-tuned protocol enables a more accurate differential diagnosis of inflammatory arthritis of the hand, potentially guiding earlier and more targeted interventions.

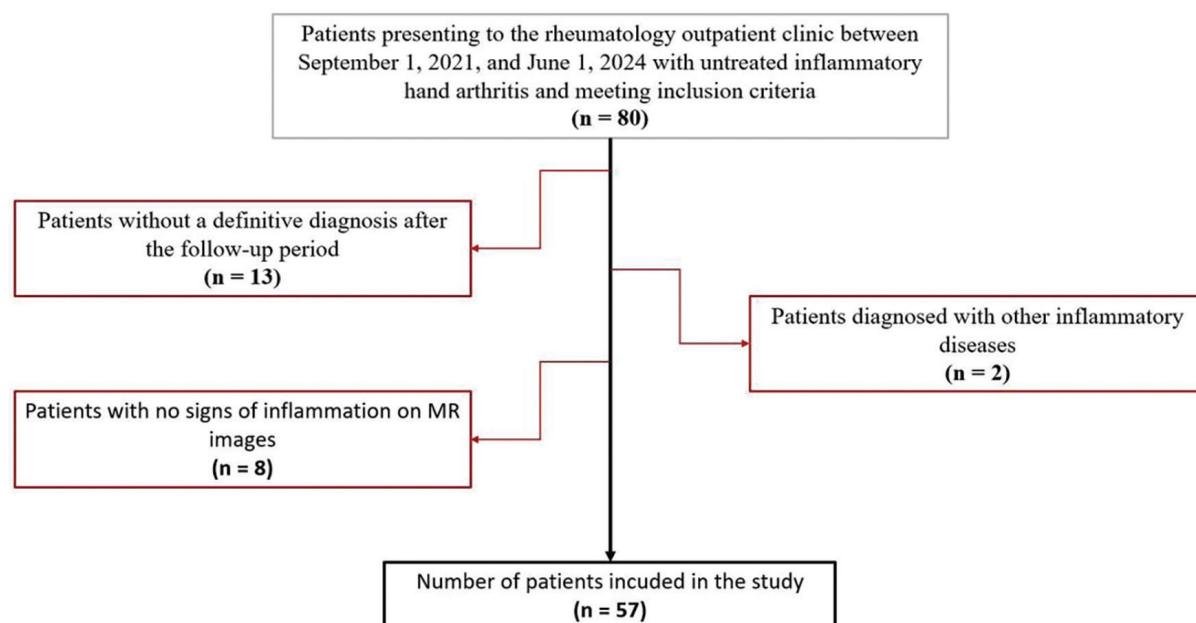


Figure 1. Flowchart of the study. MR, magnetic resonance.

clinical diagnosis and were not provided with laboratory results, treatment history, or rheumatological classification. Radiographs were labeled as normal or abnormal in the presence of any findings, including joint space narrowing, osteophytes, new bone formation, and erosions. The MRI findings were categorized in a dichotomous fashion as “present” or “absent” for enthesitis [both classical (collateral ligament enthesitis and nailbed enthesitis) and functional (extensor peritendinitis, pulley enthesitis)], synovitis, tenosynovitis, osteitis, erosion, periarticular soft tissue edema, osteophyte, periarticular bone proliferation, and joint space narrowing. The distribution of synovitis was classified into nine groups based on joint involvement, and some other findings were further categorized into subgroups (Table 2).

Based on the MRI findings, patients with predominant inflammation of the SEC were assigned a presumptive diagnosis of PsA (Figure 2), and patients with predominant synovial inflammation were assigned a presumptive diagnosis of RA (Figure 3). Additionally, patients presenting with inflammatory arthritis characterized by prominent degenerative joint findings without evidence of SEC involvement were given a presumptive diagnosis of EOA or calcium pyrophosphate dihydrate deposition disease (EOA/CPD) (Figure 4).

After clinical follow-up, patients were categorized into PsA, seropositive RA (SpRA), seronegative RA (SnRA), and EOA/CPD defi-

nite diagnostic groups based on evaluation by the rheumatologist (U.K.) with the aid of diagnostic and classification criteria: the 2010 American College of Rheumatology/European Alliance of Associations for Rheumatology classification criteria for RA and CASPAR for PsA.^{10,11} Skin psoriasis was required for patients to be classified as PsA. A separate group, ADIPI, was formed for patients who could not be classified into the aforementioned diagnostic categories but exhibited clinical and examination findings suggestive of PsA *sine psoriase*. Although rheumatologist had access to MRI reports, he independently established final diagnoses by self-blinding to imaging findings and presumptive radiological interpretations during the diagnostic decision-making process. Thirteen patients without a definitive diagnosis during the follow-up period and two patients diagnosed with other inflammatory arthritides were excluded from the analysis.

Statistical analysis

The analyses were conducted using the free and open-source software R (version 4.3.2, <https://cran.r-project.org>) and the SPSS for Windows Version 23.0 statistical package (Chicago, IL, USA), with the assistance of an academic biostatistician. Descriptive statistics were expressed as medians, means/standard deviations, and percentages. As appropriate, group differences were evaluated using the Student’s t-test, Mann-Whitney U test, one-way ANOVA, or Kruskal-Wallis test. Categorical data were compared using the

Pearson chi-squared test, the Fisher’s exact test, or the Yates-corrected chi-squared test. Diagnostic concordance was measured using weighted kappa coefficients. Logistic regression and receiver operating characteristic analysis were used to assess the predictive model. To identify risk factors for PsA, univariate analysis followed by multiple binary logistic regression was conducted. Variables with a *P* value of < 0.20 in univariate analysis were included as candidates in the multivariate model. A forward elimination method based on likelihood ratio statistics was used for variable selection. Clinically relevant potential confounders, such as age, were deliberately included to control for their effects. Model calibration (goodness-of-fit) was assessed using the Hosmer-Lemeshow test, with a *P* value of > 0.05 indicating an adequate model fit. Classification accuracy was evaluated using the classification table, and discriminative ability was assessed with the area under the curve (AUC). The pROC package was used for the AUC and its plot for the final model. A *P* value of < 5% was considered statistically significant. The sample size was calculated to require 81 participants. The statistical guarantor (H.A.) was responsible for ensuring the transparency and accuracy of all analyses and addressing any related queries.

Results

Study cohort

Hand MRI was performed on 80 patients, and 57 were included after excluding those

Table 1. Magnetic resonance imaging protocol optimized for inflammatory hand arthritis

Parameter/sequence	Ax T1 FSE	Cor T1	Cor STIR	Cor PD	Ax T2 (FS)	Sag T2 (FS)	Ax T1 FSE (FS)	C+ ax T1 FSE (FS)	C+ cor T1	C+ 3D cor FSPGR (FS)	3D cor ZTE
FOV (cm)	12	20	20	20	12	18	12	12	20	20	20
TR (ms)	793	684	3,048	1,950	2,794	4,068	793	793	684	min	2
TE (ms)	min	min	42	42	75	70	min	min	min	min	~0*
ETL	3	4	14	10	20	22	3	3	4	-	-
Flip angle (°)	111	111	120	111	160	111	111	111	111	10	2
Band width (Hz/pixel)	83.33	83.33	83.33	83.33	62.5	62.5	83.33	83.33	83.33	31.25	83.33
Matrix size (Fr × Phase)	512 × 360	640 × 460	440 × 380	640 × 460	352 × 352	640 × 460	512 × 360	512 × 361	640 × 460	400 × 350	320 × 320
NEX	1	1	1.5	1	1	1	1	1	1	1	4
Slice thickness (mm)	3	2	2	2	3	2.5	3	3	2	0.4	0.6
Interslice gap (mm)	1	0.5	0.5	0.5	1	1	1	1	0.5	-	-
Scan duration (min)	3-4	1	2-3	1-2	3-4	1-2	3-4	3-4	1	3	4-5

*TE time is a very small value of approximately 8 μs; Ax, axial; Cor, coronal; Sag, sagittal; FS, fat saturated; FSE, fast spin echo; 3D, three dimensional; FSPGR, fast spoiled gradient echo; TE, time to echo; TR, time to repetition; ZTE, zero time to echo; FOV, field of view; ETL, echo train length; NEX, number of excitations; Hz, hertz; Fr, frequency; C+, contrast enhanced.

who met the exclusion criteria. The final cohort comprised 42 women and 15 men, with a median age of 54 years (range 28–79 years). Among the hands imaged, 37 were right hands, and 20 were left hands. The median duration of symptoms was 12 months (interquartile range: 1–84) (Table 3). An alluvial diagram illustrating the study flow is presented in Figure 5.

Subgroup analysis

The PsA and ADIPI groups demonstrated comparable demographics, including gender distribution, age, and symptom duration. Nailbed enthesitis was more prevalent in patients with ADIPI than in those with PsA ($P = 0.048$, effect size: 0.416). Radiographic abnormalities were also more common in the ADIPI group than in the PsA group

($P = 0.035$, effect size: 0.467). Other MRI findings showed no significant differences between these two groups.

The ADIPI group showed no significant difference in overall enthesitis rates compared with the SnRA group when all four subtypes of enthesitis were considered ($P = 0.170$). However, when pulley enthesitis was excluded from the analysis, the ADIPI group demonstrated significantly higher enthesitis rates than the SnRA group (90.5% vs. 37.5%; $P = 0.008$, effect size: 0.553). Periarticular edema was also more frequent in ADIPI cases than in SnRA ($P = 0.028$, effect size: 0.472).

The ADIPI and EOA/CPPD patient groups differed according to many variables. The patients with EOA/CPPD [mean age: 63 (60–65) years] were significantly older than patients with ADIPI

Table 2. Magnetic resonance imaging features assessed and their modifiers

Synovitis
Only wrist
Wrist + metacarpophalangeal joints
Wrist + interphalangeal joints
Wrist + metacarpophalangeal + interphalangeal joints
Only metacarpophalangeal joints
Metacarpophalangeal + interphalangeal joints
Only proximal interphalangeal joints
Only distal interphalangeal joints
Proximal interphalangeal joints + distal interphalangeal joints
Enthesitis
Classical (collateral ligament enthesitis and nail-bed enthesitis)
Functional (extensor peritendinitis and pulley enthesitis)
Tenosynovitis
With pulley enthesitis
Without pulley enthesitis
Osteitis
Subchondral
Diaphyseal
Erosions
Marginal
Central
Periarticular soft tissue edema
Osteophytes
Marginal
Hook
Periarticular bone proliferation
Joint space narrowing
Uniform
Non-uniform

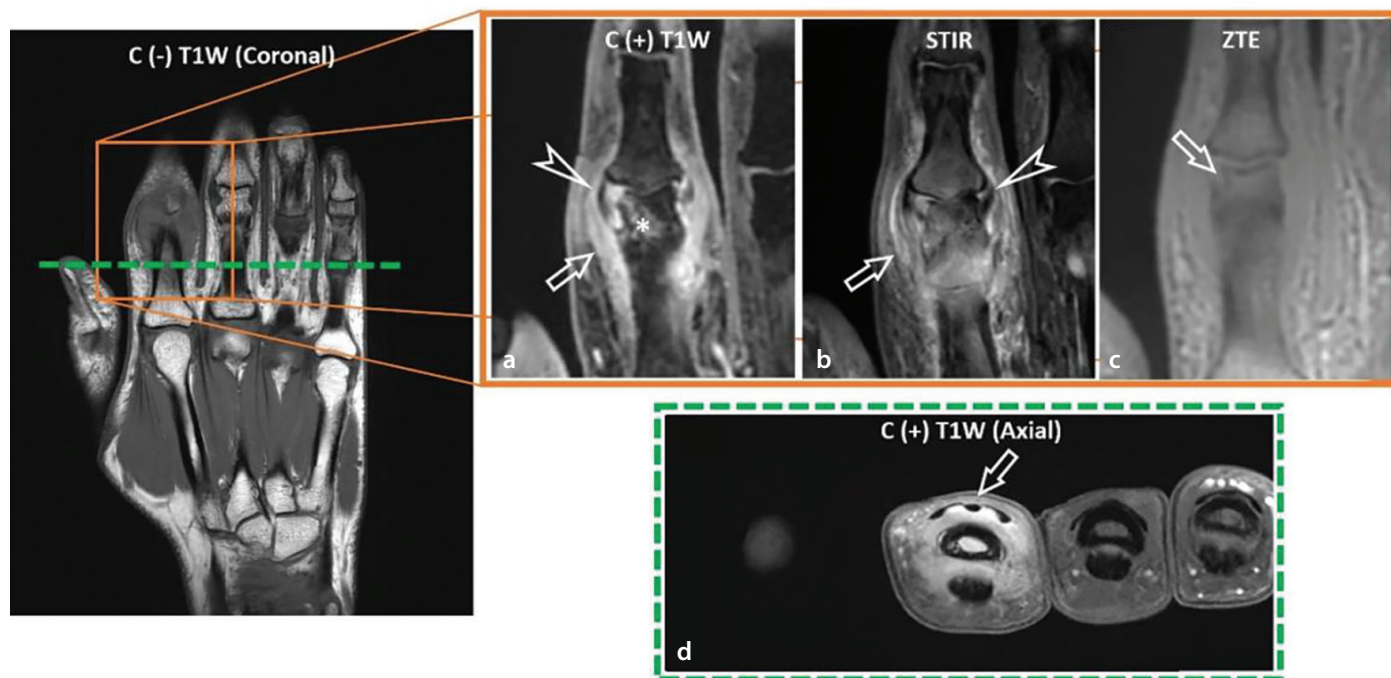


Figure 2. A patient with predominantly synovioentheseal complex inflammation. A 36-year-old male patient with a 4-month history of swelling and pain in the right second finger. Magnetic resonance imaging (MRI) demonstrated widespread periarticular soft tissue inflammation at the level of the second finger proximal interphalangeal joint (a and b, arrows). Enthesitis of the collateral ligament was observed (a and b, arrowheads), accompanied by a millimetric erosion at the insertion site of the collateral ligament (c, arrow). Additionally, periostitis of the proximal phalanx and extensive osteitis extending to the diaphysis (a, asterisk) were noted. Axial images revealed extensor peritendinitis (d, arrow). The MRI findings were consistent with inflammatory arthritis, and the final diagnosis was psoriatic arthritis. ZTE, zero time to echo; C+, contrast enhanced; C-, contrast non-enhanced; STIR, short tau inversion recovery; T1W, T1-weighted.

Table 3. Descriptive data of the study group*	
Variables	n (%)
Total number of patients	57 (100)
Gender	
Female	42 (73.6)
Male	15 (26.4)
Age (years)	
Mean $\pm$ SD	53.95 $\pm$ 12.30
Median [min–max]	54 [28–79]
Magnetic resonance-imaged hand	
Right	37 (64.9)
Left	20 (35.1)
Duration of symptoms (months)	
Mean $\pm$ SD	14.12 $\pm$ 16.73
Median [min–max]	12 [1–84]
Distribution of final diagnoses	
PsA	11 (19.2)
SpRA	6 (10.5)
SnRA	8 (14.0)
EOA/CPPD	11 (19.2)
ADIIPI	21 (36.8)

*Patients excluded from the study are not included in this table; n, number of patients; SD, standard deviation; PsA, psoriatic arthritis; SpRA, seropositive rheumatoid arthritis; SnRA, seronegative rheumatoid arthritis; EOA/CPPD, erosive osteoarthritis/calcium pyrophosphate dihydrate deposition disease; ADIIPI: arthritis involving distal interphalangeal joint.

[mean age: 46 (42–55) years] ($P < 0.001$, effect size = 0.678). Regarding the MRI findings, enthesitis was much more common in the ADIIPI group (90.4%) than in the EOA/CPPD group (9.0%) ($P < 0.001$, effect size = 0.798). The presence and type of osteophytes differed significantly between the groups. While 14.2% of patients in the ADIIPI group had marginal osteophytes, all patients in the EOA/CPPD group had marginal and/or hook osteophytes ($P < 0.001$, effect size: 0.835). Periarticular edema was observed in 76.1% of the ADIIPI group but in no cases in the EOA/CPPD group ($P < 0.001$, effect size: 0.724). The erosion patterns of these two groups were also different; although central erosions were more common in the EOA/CPPD group, marginal erosions were more common in the ADIIPI group ($P = 0.010$, effect size: 0.573).

The SpRA and SnRA groups displayed no significant differences regarding either demographic or MRI findings. Although subchondral osteitis was more frequent in the SpRA group than in the SnRA group, the difference was not statistically significant ($P = 0.103$).

The SpRA and SnRA patient groups were combined under the title of RA, since the two groups did not show significant differences in any variable. Compared with RA, patients

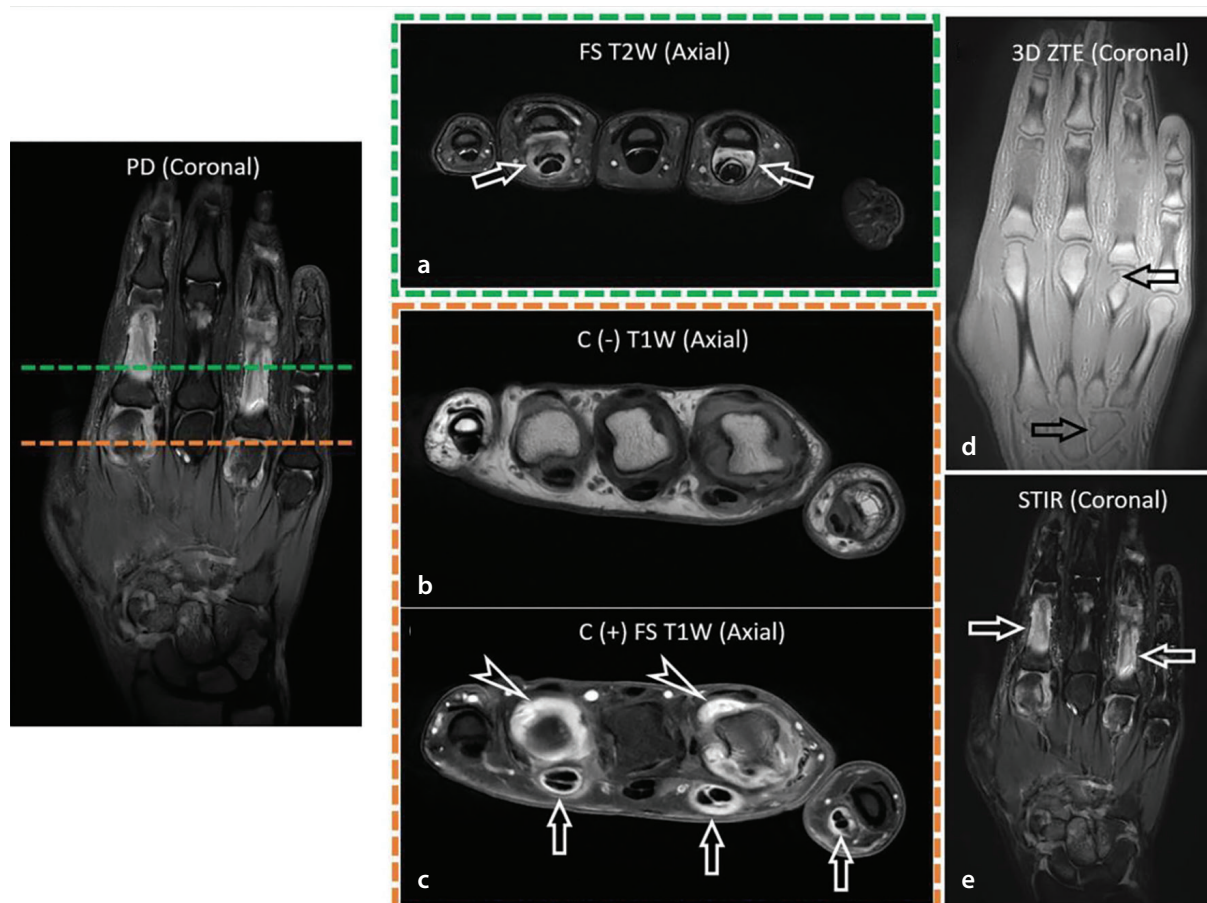


Figure 3. A patient with predominantly synovial inflammation. A 22-year-old female patient had been experiencing diffuse pain and swelling in her hands, more pronounced in the left hand, for the past 6 months. Magnetic resonance imaging (MRI) showed flexor tenosynovitis in the first, second, and fourth fingers (a, c, and e, arrows). There was also prominent synovitis and effusion in the metacarpophalangeal (MCP) joints that are easily seen as enlargement of joint spaces on pre-contrast T1W image (b) with intense contrast enhancement on post-contrast images (c, arrowheads). Additionally, synovitis was accompanied by erosions in the wrist and MCP heads, most notably on the radial side of the fourth metacarpal [(d) arrows]. These MRI findings were consistent with an inflammatory arthritis characterized by intense synovial inflammation, leading to a diagnosis of seropositive rheumatoid arthritis. 3D, three dimensional; ZTE, zero time to echo; C+, contrast enhanced; C-, contrast non-enhanced; FS, fat saturated; T1W, T1-weighted; T2W, T2-weighted; PD, proton density-weighted.

with PsA showed no significant difference in overall enthesitis rates (90.9% vs. 64.3%; $P = 0.180$). However, when pulley enthesitis was excluded from the analysis, the PsA group demonstrated significantly higher enthesitis rates than the RA group (90.9% vs. 42.9%; $P = 0.033$, effect size: 0.497). Additionally, the PsA group exhibited higher rates of periarticular edema than the RA group (81.8% vs. 35.7%; $P = 0.042$, effect size: 0.461). There was no significant difference between the groups in terms of other findings, including the distribution of synovitis.

The ADIPI group was combined with the PsA group before further analysis because there was no significant difference between the PsA and ADIPI groups; on the contrary, nailbed enthesitis, which is well known to be associated with PsA, was more common in the ADIPI group. Enthesitis, apart from pulley enthesitis, was significantly more prevalent in the PsA + ADIPI group than in the RA

group (93.7% vs. 42.9%; $P < 0.001$, effect size: 0.515). Periarticular edema was also seen more in the PsA + ADIPI group than in the other groups ($P < 0.008$, effect size: 0.410).

The distribution of MRI findings by final diagnosis groups is presented in Table 4.

Risk factors for psoriatic arthritis

Logistic regression identified enthesitis as a significant predictor for PsA + ADIPI, with an odds ratio of 24.27 [95% confidence interval (CI): 2.624–63.309] across all groups. This model's AUC was 0.81, with a correct classification rate of 83.7%.

Diagnostic agreement

The weighted kappa coefficient indicated excellent agreement between MRI-based presumptive diagnoses and definite clinical diagnoses ($\kappa = 0.90$; 95% CI: 0.812–0.961).

Discussion

Our study demonstrated that MRI-based differential diagnoses showed high concordance with definitive clinical diagnoses. A fine-tuned hand MRI protocol successfully identified subtle yet critical findings that aided differentiation. Prior studies have similarly highlighted MRI's role in early arthritis diagnosis.^{12,13} Nevertheless, considering its limitations, imaging should be interpreted in conjunction with clinical and laboratory data.^{14,15}

The results from patients with PsA and those with ADIPI were similar except for radiographic findings and nailbed enthesitis. The higher prevalence of radiographic abnormalities in ADIPI may indicate more aggressive disease or rapid progression, which brings early clinical presentation with diagnostic challenges. Montilla et al.¹⁶ similarly reported early radiographic changes in almost half of

	PsA	SpRA	SnRA	EOA/CPPD	ADIPI	Total (n [%] ^b)
Synovitis (n [%] ^a)	11 [19.3]	6 [10.5]	8 [14.0]	11 [19.3]	21 [36.9]	57 (100)
Enthesitis* (n [%])	10 [25.0]	3 [7.5]	6 [15.0]	1 [2.5]	20 [50.0]	40 (70.2)
Ligamentous enthesitis (n [%])	7 [25.9]	3 [11.1]	3 [11.1]	1 [3.7]	13 [48.2]	27 (47.3)
Extensor peritendinitis (n [%])	5 [31.3]	2 [12.5]	-	1 [6.2]	8 [50.0]	16 (28.0)
Nail-bed enthesitis (n [%])	2 [11.8]	1 [5.9]	-	1 [5.9]	13 [76.4]	17 (29.8)
Pulley enthesitis (n [%])	6 [25.0]	1 [4.2]	5 [20.8]	1 [4.2]	11 [45.8]	24 (42.1)
Tenosynovitis (n [%])	6 [23.1]	2 [7.7]	5 [19.2]	2 [7.7]	11 [42.3]	26 (45.6)
Without functional enthesitis (n [%])	1	1	-	1	-	3
With functional enthesitis (n [%])	5	1	5	1	11	23
Osteitis (n [%])	5 [20.0]	5 [20.0]	2 [8.0]	4 [16.0]	9 [36.0]	25 (43.8)
Subcortical (n [%])	3	5	2	2	4	16
Diaphyseal (n [%])	2	-	-	2	5	9
Erosion (n [%])	4 [12.1]	5 [15.2]	3 [9.0]	9 [27.3]	12 [36.4]	33 (57.8)
Marginal (n [%])	4	4	3	3	11	25
Central (n [%])	-	1	-	6	1	8
New bone formation (n [%])	1 [12.5]	-	1 [12.5]	-	6 [75.0]	8 (14.0)
Periarticular soft tissue edema (n [%])	9 [30.0]	3 [10.0]	2 [6.7]	-	16 [53.3]	30 (52.6)
Osteophyte (n [%])	1 [5.9]	1 [5.9]	1 [5.9]	11 [64.7]	3 [17.6]	17 (29.8)
Marginal (n [%])	1	1	1	8	3	14
Hook (n [%])	-	-	-	3	-	3
Joint space narrowing (n [%])	3 [8.9]	4 [11.8]	2 [5.9]	11 [32.3]	14 [41.1]	34 (59.6)
Uniform (n [%])	2	3	2	4	10	21 (36.8)
Non-uniform (n [%])	1	1	0	7	4	13 (22.8)

*Patients with classical or functional enthesitis in at least one region were considered positive for enthesitis; n, number of observations; ^a, percentage within the respective finding; ^b, percentage among all included patients; PsA, psoriatic arthritis; SpRA, seropositive rheumatoid arthritis; SnRA, seronegative rheumatoid arthritis; EOA/CPPD, erosive osteoarthritis/calcium pyrophosphate dihydrate deposition disease; ADIPI, arthritis with distal interphalangeal joint involvement.

patients with PsA with ADIPI within 1 year, suggesting rapid early progression.

All patients with PsA met the CASPAR criteria and had skin psoriasis, whereas patients with ADIPI lacked both, except for four borderline cases. Scarpa et al.⁹ reported that ADIPI may occur regardless of skin psoriasis, reinforcing the view that current classification criteria may overlook early PsA.² Our findings support the utility of MRI, particularly SEC inflammation, in enhancing diagnostic sensitivity alongside CASPAR.

Discussion of magnetic resonance imaging findings

Synovitis prevalence and distribution did not significantly differ between groups, consistent with Narváez et al.,¹⁷ who found synovitis to be non-specific in distinguishing PsA from RA. Although previous studies suggested more severe synovitis in RA,^{18,19} we did not grade severity due to concerns over objectivity.

We found that patients in the ADIPI group did not differ from those with PsA,

and enthesitis, apart from pulley enthesitis, emerged as distinguishing features that set these two groups apart from other types of arthritis. Moreover, nailbed enthesitis, well-documented in PsA,^{20,21} was even more frequent in ADIPI, supporting their clinical similarity.

In a high-resolution MRI study conducted by Tan et al.,²² it was reported that PsA was differentiated from osteoarthritis (OA) by collateral ligament enthesitis, extensor peritendinitis, and nailbed involvement with predominant SEC inflammation. In our study, the main MRI finding of the ADIPI group was also SEC inflammation. In contrast, the Multimodal Imaging of the Distal Interphalangeal-Joint Synovio-Entheseal Complex in PsA study questioned the specificity of SEC inflammation in DIPJ, noting similar findings in OA.²³ McGonagle et al.¹⁴ also noted that enthesitis may occur in OA, particularly with age-related degeneration.

Pulley enthesitis did not differ significantly across groups, likely due to confounding factors such as mechanical stress or diabe-

tes.²⁴ Agten et al.²⁵ previously stated that flexor tenosynovitis, closely associated with pulley function, was found to be present in almost all healthy subjects at specific degrees, further limiting its diagnostic value. To address this, we conducted a different analysis excluding pulley findings.

Although no significant group differences were found in osteitis, trends were informative. Osteitis was more frequent in SpRA (83.3%) than in SnRA (25.0%), in line with reports linking it to seropositivity.²⁶ Additionally, diaphyseal osteitis was common in the PsA + ADIPI group but absent in patients with RA, supporting literature that PsA leads to more extensive osteitis.^{19,27}

Consistent with the findings of this study, Addimanda et al.²⁸ and Grainger et al.²⁹ emphasized central erosions as a key feature of EOA, though marginal erosions are also common. However, the literature comparing erosions in PsA and RA is inconsistent; whereas some studies, including ours, report no difference,^{17,19,30} another study found erosions more common in PsA.³¹ These contrasting results highlight the heterogeneity in the

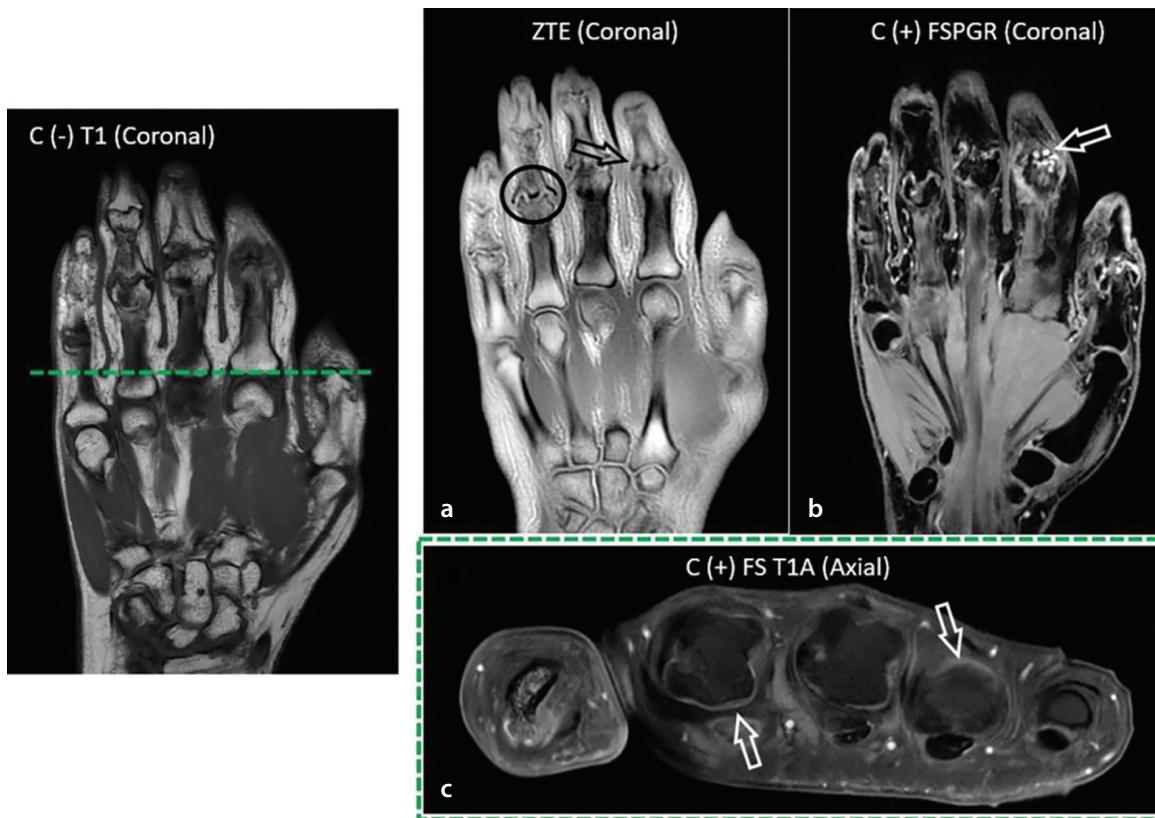


Figure 4. A case of inflammatory arthritis with predominantly degenerative findings. A 62-year-old female patient with a 12-month history of diffuse pain, more pronounced in the left hand. Magnetic resonance imaging (MRI) revealed marked joint space narrowing, prominent osteophytes (a, arrow), and subchondral cysts (b, arrow), most evident in the proximal interphalangeal joints. Gull-wing appearance (a, circle), secondary to central erosions, was also observed. An axial section at the metacarpophalangeal joint level showed mild synovitis (c, arrows). Based on these MRI findings, suggesting degenerative coexisting with inflammatory features, the patient was diagnosed with erosive osteoarthritis. ZTE, zero time to echo; C+, contrast enhanced; C-, contrast non-enhanced; FS, fat saturated; FSPGR, fast spoiled gradient recalled echo.

literature regarding the diagnostic utility of erosions in distinguishing PsA from RA.

Periarticular soft tissue edema was observed very frequently in the PsA (81.8%) and ADIPI groups (76.2%), less frequently in the RA group (35.7%), and never in the EOA/CPPD group. This supports previous US studies highlighting periarticular edema as a distinguishing feature of PsA.^{18,31} Notably, Abrar et al.³¹ incorporated periarticular soft tissue edema into a highly accurate PsA and RA differentiation model.

Strengths, limitations, and future recommendations

The use of a fine-tuned MRI protocol in this study allowed detailed evaluation of joint and periarticular structures, producing findings consistent with prior studies and demonstrating the technique's clinical utility. The prospective design minimized selection bias and provided standardized data.

Although the target sample size was calculated as 81 participants, only 57 patients were ultimately included due to stringent inclusion criteria and unforeseen exclusions.

This reduction limited the statistical power of subgroup analyses and constrained the generalizability of our findings. Therefore, although our results indicate the meaningful diagnostic utility of SEC inflammation and periarticular edema in distinguishing PsA and related patterns, these conclusions should be interpreted with caution. Additionally, the single-center design may introduce institutional bias and restrict the validity of our findings. Future multicenter studies with larger, more diverse cohorts are necessary to validate our findings and establish broader clinical applicability. Future studies are also encouraged to explore prognostic imaging markers and treatment-related changes, incorporating standardized classifications and artificial intelligence tools to enhance objectivity.

Although the radiologists interpreted the MRI without knowledge of the final clinical diagnosis, and the rheumatologist had access to MRI reports, the latter conducted his diagnostic assessments by self-blinding to imaging findings. Nonetheless, the potential for residual incorporation bias cannot be fully excluded and may still influence the observed high concordance ($\kappa = 0.90$).

The MRI was performed with arms at the side to improve patient comfort and reduce motion artifacts, though this slightly reduced signal strength. Future studies could investigate positions that strike a balance between comfort and optimal signal quality.

Finally, our findings have important implications for clinical practice. In ambiguous cases, MRI findings could support accurate diagnosis and earlier initiation of disease-modifying antirheumatic drugs, potentially preventing irreversible joint damage. Conversely, distinguishing PsA from degenerative mimics, such as EOA/CPPD, can help avoid overtreatment. Integrating specific MRI features into rheumatology workflows, particularly for seronegative or atypical presentations, could enhance diagnostic precision and personalize treatment strategies. Future research should explore prospective validation of imaging-based algorithms and assess their impact on clinical outcomes.

In conclusion, this study provides new insights by systematically evaluating IHA using high-resolution 3T-MRI. Through a prospective design involving early undifferentiated

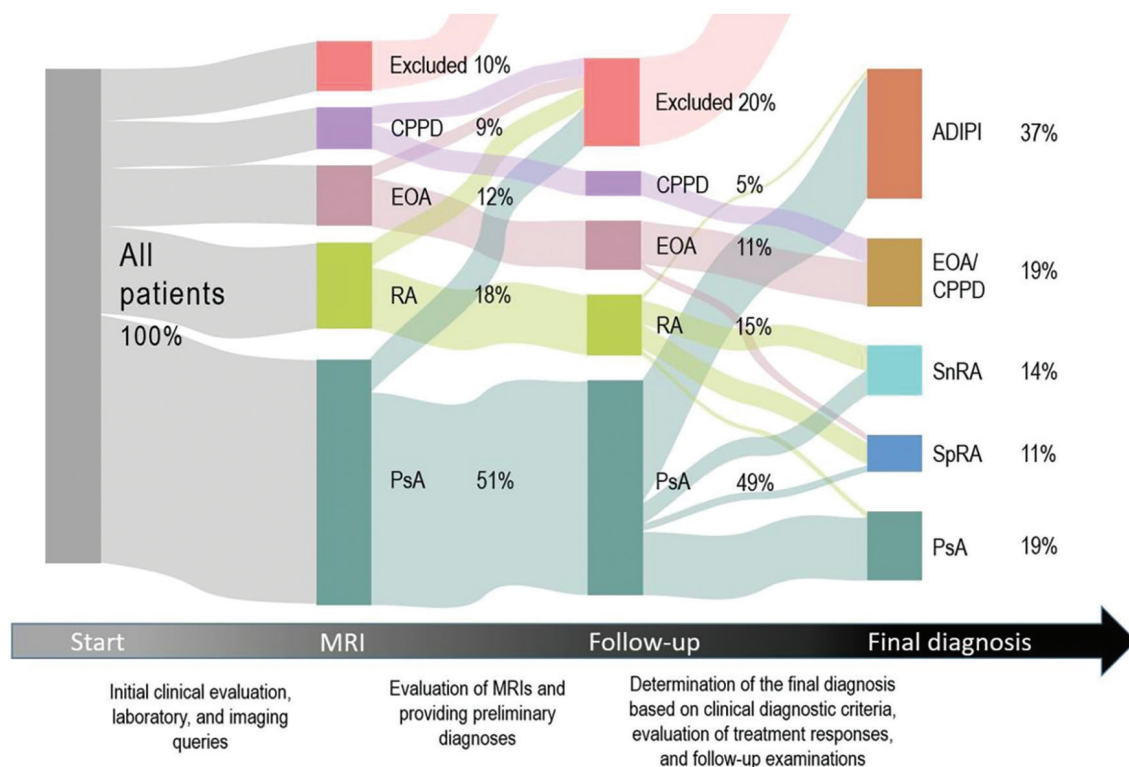


Figure 5. Alluvial diagram illustrating the study flow. **Start:** Eighty patients aged ≥ 18 years with inflammatory arthritis of the hand in at least one joint based on physical examination, laboratory, and ultrasonography findings and who were not receiving any treatment other than non-steroidal anti-inflammatory drugs for their symptoms were enrolled. Patients who were receiving steroid and disease-modifying agent treatment, who were absolutely contraindicated for magnetic resonance imaging (MRI), or who were allergic to gadolinium contrast agents were excluded. **Magnetic resonance imaging:** Among 80 patients referred to MRI by a rheumatologist, 41 patients were presumptively diagnosed with psoriatic arthritis (PsA), 14 with rheumatoid arthritis (RA), 10 with erosive OA (EOA), and 7 with calcium pyrophosphate dihydrate crystal deposition disease (CPPD). Six patients were excluded because they had no inflammatory findings on MRI, and two patients were excluded because their images were not diagnostic at this stage. **Follow-up:** A further 13 patients were excluded because they did not fall into any of the final diagnostic groups during the follow-up period, and two patients were excluded because they were diagnosed with a different condition (thyroid acropachy and infection). **Final diagnosis:** After clinical follow-up, 57 of 80 patients were included in the final diagnostic groups, with their diagnoses accurately confirmed; 11 of 57 patients were diagnosed with PsA, 6 with SpRA, 8 with SnRA, 11 with EOA/CPPD, and 21 with ADIPI.

arthritis, we demonstrate that imaging markers such as SEC inflammation and periarticular edema-especially when present in combination-may serve as early indicators of PsA. Importantly, the study offers a balanced interpretation by addressing diagnostic overlaps and proposing a potential composite MRI pattern that may support more accurate subtype classification in clinical practice. The strong diagnostic concordance between MRI and clinical assessments underscores the importance of multidisciplinary collaboration between radiologists and rheumatologists in managing these complex conditions.

Footnotes

Conflict of interest disclosure

Adalet Elçin Yıldız was not serving in any editorial role at the time of the manuscript's submission, evaluation, acceptance, or ahead of print publication. At the time of inclusion of the article in the regular issue, she is serving as Section Editor of the journal. She

was not involved in the editorial evaluation, peer-review process, or publication decision of this manuscript. The other authors declare no conflict of interest.

Funding

The authors certify that they have disclosed below all direct or indirect affiliation or financial interests in connection with the content of this paper from Grant/Research Support: Hacettepe University Scientific Research Projects.

References

- Shiraishi M, Fukuda T, Igarashi T, Tokashiki T, Kayama R, Ojiri H. Differentiating rheumatoid and psoriatic arthritis of the hand: multimodality imaging characteristics. *Radiographics*. 2020;40(5):1339-1354. [\[Crossref\]](#)
- López-Medina C, McGonagle D, Gossec L. Subclinical psoriatic arthritis and disease interception-where are we in 2024? *Rheumatology (Oxford)*. 2025;64:56-64. [\[Crossref\]](#)
- Narváez J, Sirvent E, Narváez JA, et al. Usefulness of magnetic resonance imaging of the hand versus anticyclic citrullinated peptide antibody testing to confirm the diagnosis of clinically suspected early rheumatoid arthritis in the absence of rheumatoid factor and radiographic erosions. *Semin Arthritis Rheum*. 2008;38(3):101-109. [\[Crossref\]](#)
- Kluckman ML, Bernard S, Bui-Mansfield LT. A systematic approach to radiographic evaluation of arthritis of the hand and wrist. *Contemp Diagn Radiol*. 2021;44(11):1-7. [\[Crossref\]](#)
- McQueen FM. Imaging in early rheumatoid arthritis. *Best Pract Res Clin Rheumatol*. 2013;27(4):499-522. [\[Crossref\]](#)
- Yıldız AE, Aydingöz Ü. Pivotal role of the synovioentheseal complex in the imaging of arthritis and rheumatic diseases. *Diagn Interv Radiol*. 2024;30:409-418. [\[Crossref\]](#)
- Olivieri I, Padula A, D'Angelo S, Cutro MS. Psoriatic arthritis sine psoriasis. *J Rheumatol Suppl*. 2009;83:28-29. [\[Crossref\]](#)
- Wang Q, Zhang H, Dai SM. Differentiating psoriatic arthritis sine psoriasis from

- seronegative rheumatoid arthritis-Experiences from five patients. *Int J Rheum Dis*. 2022;25(9):1088-1092. [\[Crossref\]](#)
9. Scarpa R, Cosentini E, Manguso F, et al. Clinical and genetic aspects of psoriatic arthritis sine psoriasis. *J Rheumatol*. 2003;30(12):2638-2640. [\[Crossref\]](#)
10. Taylor W, Gladman D, Helliwell P, et al. Classification criteria for psoriatic arthritis: development of new criteria from a large international study. *Arthritis Rheum*. 2006;54(8):2665-2673. [\[Crossref\]](#)
11. Aletaha D, Neogi T, Silman AJ, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League against rheumatism collaborative initiative. *Arthritis Rheum*. 2010;62(9):2569-2581. [\[Crossref\]](#)
12. Yue J, Wu D, Tam LS. The role of imaging in early diagnosis and prevention of joint damage in inflammatory arthritis. *Expert Rev Clin Immunol*. 2018;14(6):499-511. [\[Crossref\]](#)
13. Boeren AMP, Oei EHG, van der Helm-van Mil AHM. The value of MRI for detecting subclinical joint inflammation in clinically suspect arthralgia. *RMD Open*. 2022;8(2):e002128. [\[Crossref\]](#)
14. McGonagle D, Hermann KG, Tan AL. Differentiation between osteoarthritis and psoriatic arthritis: implications for pathogenesis and treatment in the biologic therapy era. *Rheumatology (Oxford)*. 2015;54(1):29-38. [\[Crossref\]](#)
15. Braum LS, McGonagle D, Bruns A, et al. Characterisation of hand small joints arthropathy using high-resolution MRI-limited discrimination between osteoarthritis and psoriatic arthritis. *Eur Radiol*. 2013;23(6):1686-1693. [\[Crossref\]](#)
16. Montilla CA, González CM, López-Longo F, et al. SAT0025 evolution of psoriatic arthritis with distal interphalangeal joints involvement at onset. *Ann Rheum Dis*. 2001;60(Suppl 1):A148. [\[Crossref\]](#)
17. Narváez J, Narváez JA, de Albert M, Gómez-Vaquero C, Nolla JM. Can magnetic resonance imaging of the hand and wrist differentiate between rheumatoid arthritis and psoriatic arthritis in the early stages of the disease. *Semin Arthritis Rheum*. 2012;42(3):234-245. [\[Crossref\]](#)
18. Zabotti A, Salvin S, Quartuccio L, De Vita S. Differentiation between early rheumatoid and early psoriatic arthritis by the ultrasonographic study of the synovio-enthesal complex of the small joints of the hands. *Clin Exp Rheumatol*. 2016;34(3):459-465. [\[Crossref\]](#)
19. Marzo-Ortega H, Tanner SF, Rhodes LA, et al. Magnetic resonance imaging in the assessment of metacarpophalangeal joint disease in early psoriatic and rheumatoid arthritis. *Scand J Rheumatol*. 2009;38(2):79-83. [\[Crossref\]](#) Erratum in: *Scand J Rheumatol*. 2010;39(6):527.
20. McGonagle D, Kavanaugh A, McInnes IB, et al. Association of the clinical components in the distal interphalangeal joint synovio-enthesal complex and subsequent response to ixekizumab or adalimumab in psoriatic arthritis. *Rheumatology (Oxford)*. 2024;63(11):3115-3123. [\[Crossref\]](#)
21. Tan A, Benjamin M, Toumi H, et al. The relationship between the extensor tendon entheses and the nail in distal interphalangeal joint disease in psoriatic arthritis—a high-resolution MRI and histological study. *Rheumatology (Oxford)*. 2007;46(2):253-256. [\[Crossref\]](#)
22. Tan AL, Grainger AJ, Tanner SF, Emery P, McGonagle D. A high-resolution magnetic resonance imaging study of distal interphalangeal joint arthropathy in psoriatic arthritis and osteoarthritis: are they the same? *Arthritis Rheum*. 2006;54(4):1328-1333. [\[Crossref\]](#)
23. Guldberg-Møller J, Mogensen M, Ellegaard K, et al. Multimodal imaging of the distal interphalangeal-joint synovio-enthesal complex in psoriatic arthritis (MIDAS): a cross-sectional study on the diagnostic accuracy of different imaging modalities comparing psoriatic arthritis to psoriasis and osteoarthritis. *RMD Open*. 2022;8(1):e002109. [\[Crossref\]](#)
24. Furlan A, Stramare R. The thickening of flexor tendons pulleys: a useful ultrasonographical sign in the diagnosis of psoriatic arthritis. *J Ultrasound*. 2018;21(4):309-314. [\[Crossref\]](#)
25. Agten CA, Roskopf AB, Jonczy M, et al. Frequency of inflammatory-like MR imaging findings in asymptomatic fingers of healthy volunteers. *Skeletal Radiol*. 2018;47(2):279-287. [\[Crossref\]](#)
26. Boeters DM, Nieuwenhuis WP, Verheul MK, et al. MRI-detected osteitis is not associated with the presence or level of ACPA alone, but with the combined presence of ACPA and RF. *Arthritis Res Ther*. 2016;18:179. [\[Crossref\]](#)
27. McQueen FM, Dalbeth N, Doyle A. MRI in psoriatic arthritis: insights into pathogenesis and treatment response. *Curr Rheumatol Rep*. 2008;10(4):303-310. [\[Crossref\]](#)
28. Addimanda O, Mancarella L, Dolzani P, et al. Clinical and radiographic distribution of structural damage in erosive and nonerosive hand osteoarthritis. *Arthritis Care Res (Hoboken)*. 2012;64(7):1046-1053. [\[Crossref\]](#)
29. Grainger AJ, Farrant JM, O'Connor PJ, et al. MR imaging of erosions in interphalangeal joint osteoarthritis: is all osteoarthritis erosive? *Skeletal Radiol*. 2007;36(8):737-745. [\[Crossref\]](#)
30. Abrar DB, Schleich C, Brinks R, et al. Differentiating rheumatoid and psoriatic arthritis: a systematic analysis of high-resolution magnetic resonance imaging features-preliminary findings. *Skeletal Radiol*. 2021;50(3):531-541. [\[Crossref\]](#)
31. Abrar DB, Schleich C, Nebelung S, et al. High-resolution MRI of flexor tendon pulleys using a 16-channel hand coil: disease detection and differentiation of psoriatic and rheumatoid arthritis. *Arthritis Res Ther*. 2020;22(1):40. [\[Crossref\]](#)



The hematoma measurement accuracy of dual-energy computed tomography angiography in spontaneous intracerebral hemorrhage

Baoteng Zhang
 Qingjing Huang
 Feijian Wu
 Suying Wu
 Jianfang Huang
 Zhihua Lu

The First Hospital of Putian City, Department of
Radiology, Putian, China

PURPOSE

To investigate the accuracy of hematoma volume (V), surface area (S), and surface regularity (SR) measured on dual-energy computed tomography angiography (DECTA), using non-contrast computed tomography (NCCT) as the reference standard.

METHODS

A total of 129 patients with spontaneous intracerebral hemorrhage (sICH) who underwent both NCCT and DECTA scans were retrospectively studied. Patients were stratified by V: Group 1 (< 30 mL), Group 2 (30–60 mL), and Group 3 (> 60 mL); and by morphology: regular, irregular, and lobular. DECTA data were post-processed to generate conventional computed tomography angiography (CTA) and 60 keV virtual monoenergetic imaging (VMI). These, along with NCCT images, were imported into 3D Slicer software to obtain V and S; SR was subsequently calculated. Deviation percentages (ΔV , ΔS , ΔSR) of conventional CTA and 60 keV VMI relative to NCCT were calculated. We assessed correlations using simple linear regression, compared different modalities with unpaired t-tests or Mann–Whitney U tests, evaluated agreement via Bland–Altman analysis, and compared deviation percentages across groups using one-way analysis of variance or Kruskal–Wallis H tests. Inter-reader consistency was assessed using the intraclass correlation coefficient (ICC).

RESULTS

The average V was 28.86 ± 24.15 mL. All ICCs were excellent (0.991–0.999). Strong correlations were found for V and S between conventional CTA/60 keV VMI and NCCT (r: 0.987–0.996). Bland–Altman analysis for V showed mean biases of 0.21 mL (conventional CTA) and –0.04 mL (60 keV VMI) against NCCT, with 95% limits of agreement of –4.72 to 5.14 mL and –4.51 to 4.43 mL, respectively. However, SR values from both conventional CTA and 60 keV VMI were significantly lower than those from NCCT (all $P < 0.001$). In volume-based stratification, no significant differences in ΔS (conventional CTA) or ΔV , ΔS , and ΔSR (60 keV VMI) were found among groups (all $P \geq 0.410$). For conventional CTA, ΔV was significantly smaller in Group 3 (> 60 mL) than in Group 1 (3.07% vs. 5.57%, adjusted $P = 0.047$), whereas ΔSR was larger (13.26% vs. 9.88%, adjusted $P = 0.040$). Morphology-based stratification revealed no significant differences in ΔV , ΔS , or ΔSR across groups for either modality (all P values ≥ 0.085).

CONCLUSION

Hematoma volume measurements from DECTA show good agreement with NCCT measurements, suggesting potential utility for follow-up assessment in certain clinical scenarios. However, DECTA-derived SR measurements are significantly lower than those from NCCT. Volume measurement accuracy of conventional CTA was higher for large hematomas (> 60 mL).

CLINICAL SIGNIFICANCE

Accurate hematoma measurement is crucial for prognosis and management in sICH. This study indicates that V measurements from DECTA are comparable to those from NCCT in certain clinical scenarios, offering potential added value when CTA is clinically indicated.

KEYWORDS

Dual-energy, computed tomography angiography, virtual monoenergetic imaging, intracerebral hemorrhage, hematoma volume

Handling editor: Mehmet Ruhi Onur

Corresponding author: Zhihua Lu

E-mail: luzhuren0438@163.com

Received 09 January 2026; revision requested 24
February 2026; accepted 08 April 2026.



Epub: 01.06.2026

Publication date: 01.09.2026

DOI: 10.4274/dir.2026.263859

Spontaneous intracerebral hemorrhage (sICH) is a neurological emergency characterized by non-traumatic rupture of cerebral vessels and subsequent hematoma formation within the parenchyma, associated with high mortality and morbidity.¹ Hematoma location, hematoma volume (V), hematoma expansion (HE), and perihematomal edema are independent predictors of functional outcomes and mortality.²⁻⁴ Thus, accurate hematoma assessment is vital for treatment planning and prognosis prediction.

Neuroimaging plays a pivotal role in the diagnosis of intracerebral hemorrhage (ICH) and monitoring its progression. Non-contrast computed tomography (NCCT), favored for its speed and diagnostic accuracy, is the first-line modality. Various NCCT signs based on hematoma density and shape are used to predict HE.^{5,6} Hematoma morphology is a recognized prognostic factor for HE, with visual scales categorizing hematomas by morphological characteristics.⁷⁻⁹ Irregular margins have been linked to HE and worse outcomes. To reduce subjective bias, Oge et al.¹⁰ investigated surface regularity (SR), calculated from volume and surface area (S), and its relationship to HE, finding that hematomas evolve into more irregular shapes with significantly reduced SR during follow-up.

Computed tomography (CT) angiography (CTA) is used to identify underlying etiologies (e.g., aneurysms or vascular malformations). The “spot sign” on CTA also predicts HE.¹¹ Dual-energy CT has advanced rapidly and is widely applied in clinical neuroradiology. Dual-energy CTA (DECTA) provides not only conventional CTA data but also additional parameters such as iodine maps, virtual non-contrast (VNC) images, and virtual monoenergetic imaging (VMI), potentially enhancing ICH detection and HE predic-

tion.¹² In practice, neuroradiologists need to measure V on CTA to assess progression, yet the accuracy of such cross-modal measurements remains unclear.

Hematoma volume is a key indicator of early mortality and 90-day neurological recovery.^{2,13} three dimensional (3D) Slicer, a free, open-source image analysis platform, facilitates V measurement via semi-automated segmentation and 3D reconstruction based on radiodensity, offering accurate and reproducible assessment independent of hematoma shape or location.^{14,15}

This study used 3D Slicer software to retrospectively evaluate the accuracy of V, S, and SR measurements on DECTA relative to NCCT.

Methods

Study population and grouping

This retrospective study recruited patients diagnosed with sICH at the First Hospital of Putian City between January 2020 and June 2024. Inclusion criteria: (1) CT scans performed within 6 hours of symptom onset, with NCCT and DECTA performed sequentially within 30 minutes; (2) a single parenchymal hematoma with volume ≥ 5 mL; (3) no pregnancy, iodine contrast allergy, or severe cardiopulmonary/renal dysfunction. Exclusion criteria: (1) hematoma outline obscured by intratumoral hemorrhage or hemorrhagic transformation of ischemic infarction, and (2) poor image quality or incomplete clinical data. Figure 1 shows the patient selection flowchart. The Ethics Committee of the First Hospital of Putian City approved the study (protocol number: 2022-037, date:

28.12.2022); written informed consent was waived due to the retrospective and anonymized design of this study.

Based on NCCT-measured volume using 3D Slicer, patients were stratified as Group 1 (< 30 mL), Group 2 (30-60 mL), and Group 3 (> 60 mL). Based on the hematoma's maximal slice, morphology was categorized as regular (round/ellipsoid, smooth margins), irregular (frayed margins), or lobular.¹⁶

Image acquisition and reconstruction

All scans were performed on a third-generation dual-source CT scanner (SOMATOM Force; Siemens Healthineers, Erlangen, Germany); NCCT was followed by DECTA. Scanning range: skull base to vertex; NCCT parameters: tube voltage: 120 kV, pitch: 0.6 mm; DECTA parameters: tube voltages: 80 kV and 150 kV, pitch: 0.7 mm. Reconstruction: matrix: 512×512 , slice thickness: 1.0 mm, spacing: 0.7 mm, and convolution kernel: Qr40. Automated tube current modulation (CareDose4D) was used.

Iopromide (370 mgI/mL; Bayer, Berlin, Germany) was injected via the antecubital vein at 1.0 mL/kg (4.5 mL/s), followed by a 40 mL saline flush. Bolus tracking in the ascending aorta [100 Hounsfield unit (HU) threshold + 4s delay] triggered scanning.

The 80 kV, Sn150 kV, and blended 115 kV (50/50 mix, approximating conventional 120 kV) images were processed on a Siemens workstation (syngo.via VB10B). Based on literature supporting improved head/neck CTA quality at 60 keV,^{17,18} 60 keV VMI images were reconstructed; NCCT, conventional CTA, and

Main points

- Hematoma volume and surface area measured on dual-energy computed tomography angiography (DECTA)-derived images from conventional computed tomography angiography (CTA) and 60 keV virtual monoenergetic imaging show excellent correlation with non-contrast computed tomography measurements.
- The DECTA approach has potential value for the follow-up assessment of hematoma volume in certain clinical scenarios.
- Conventional CTA is more accurate for volume measurements of large hematomas (> 60 mL) than for small- and medium-sized hematomas.

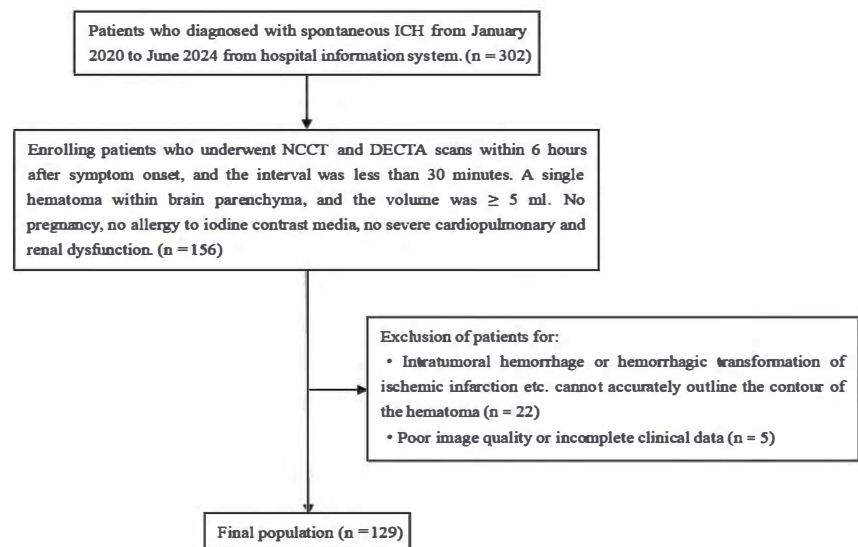


Figure 1. The flowchart of the inclusion and exclusion criteria for the study. ICH, intracerebral hemorrhage; NCCT, non-contrast computed tomography; DECTA, dual-energy computed tomography angiography.

60 keV VMI sets were used for subsequent analysis.

Hematoma measurement

DICOM images were analyzed using 3D Slicer (v5.7.0; Brigham and Women's Hospital, Boston, MA, USA). Two well-trained radiologists, blinded to clinical data, semi-automatically delineated hematoma outlines (threshold: 40–100 HU) with manual correction. The software calculated V and S; the average of the two radiologists' measurements was used (Figure 2); SR was calculated as follows:

$SR = 6\sqrt{\pi}(\frac{V}{S^2})$, ranging from 0 (irregular) to 1 (spherical) (10). Deviation percentages (ΔV , ΔS , and ΔSR) for conventional CTA and 60 keV VMI relative to NCCT were computed using the following formula:

$$\Delta V = \frac{|V_{test} - V_{control}|}{V_{control}} \times 100\%$$

Statistical analysis

Analyses were performed using SPSS Statistics (v25.0; IBM, Armonk, NY, USA) and GraphPad Prism (v8.0.2; GraphPad Software, San Diego, CA, USA). Continuous variables are expressed as mean $\pm$ SD or median (interquartile range) based on normality (Kolmogorov–Smirnov test). Categorical variables are presented as n (%). Simple linear regression assessed correlations. Inter-modality comparisons used unpaired t-tests or Mann–Whitney U tests. Agreement was evaluated via Bland–Altman analysis. Deviation percentages across groups were compared using one-way analysis of variance or the Kruskal–Wallis H test, with post-hoc Dunn–Bonferroni tests. Box plots were used to visualize the distribution of deviation percentages for conventional CTA and 60 keV VMI. The intraclass correlation coefficient (ICC) assessed inter-reader consistency: 0.41–0.60 (moderate), 0.61–0.80 (strong), 0.81–1.00 (excellent). A *P* value < 0.05 was considered statistically significant.

Results

The study included 129 patients, including 86 men (66.7%) and 43 women (33.3%), with a mean age of 61.47 ± 11.06 years. The interval from symptom onset to scan was 3.12 ± 1.80 hours. Volume groups: Group 1: 80 (62.0%), Group 2: 33 (25.6%), Group 3: 16 (12.4%). Morphology: regular: 29 (22.5%), irregular: 63 (48.8%), lobular: 37 (28.7%). Location: deep (basal ganglia/thalamus): 74 (57.4%), lobar: 53 (41.1%), infratentorial: 2 (1.6%). Inter-reader ICCs for V and S across all modalities were excellent (0.991–0.999) (Table 1).

Strong linear correlations existed for V and S between conventional CTA/60 keV VMI and NCCT (all *P* < 0.001, *r*: 0.987–0.996) (Figures 3a1–b2), with no significant difference between correlation coefficients (all *P* $\geq$ 0.359). However, SR values from both conventional CTA and 60 keV VMI were significantly lower than those from NCCT (all *P* < 0.001). Bland–Altman analysis (Figure 4) showed for conventional CTA vs. NCCT: mean bias (V: 0.21 mL, S: 5.22 mm², SR: –0.06) with 95% limits of agreement (LoA) of –4.72, 5.14 mL, –10.70, 21.14 mm², –0.19, 0.08. For 60 keV VMI vs. NCCT: mean bias (V: –0.04 mL, S: 3.21 mm², SR: –0.05) with 95% LoA of –4.51, 4.43 mL, –8.45, 14.88 mm², –0.19, 0.10.

The results of volume-stratified analysis were as follows: (1) ΔV (Conventional CTA): Group 1: 5.57%, Group 2: 5.21%, Group 3: 3.07%; ΔV in Group 3 was significantly low-

er than in Group 1 (*z* = –2.419, adjusted *P* = 0.047, Figure 5a); (2) ΔV (60 keV VMI): no significant difference among groups (5.23%, 4.66%, 3.40%; *H* = 1.785, *P* = 0.410, Figure 5b); (3) ΔS (Conventional CTA and 60 keV VMI): no significant differences among volume groups (*H* = 1.665/0.381, *P* = 0.435/0.827) (Figure 6a, b); (4) ΔSR (Conventional CTA): Group 1: 9.88%, Group 2: 7.47%, Group 3: 13.26%; ΔSR in Group 3 was significantly higher than in Group 2 (*z* = 2.475, adjusted *P* = 0.040) (Figure 7a); (5) ΔSR (60 keV VMI): no significant difference among groups (7.35%, 4.88%, 8.75%; *H* = 0.657, *P* = 0.720) (Figure 7b).

The results of the morphology-stratified analysis indicated no significant differences in ΔV , ΔS , or ΔSR among the regular, irregular, and lobular groups for either conventional CTA or 60 keV VMI (all *H* $\geq$ 1.243, all *P* $\geq$ 0.085) (Figures 8–10).

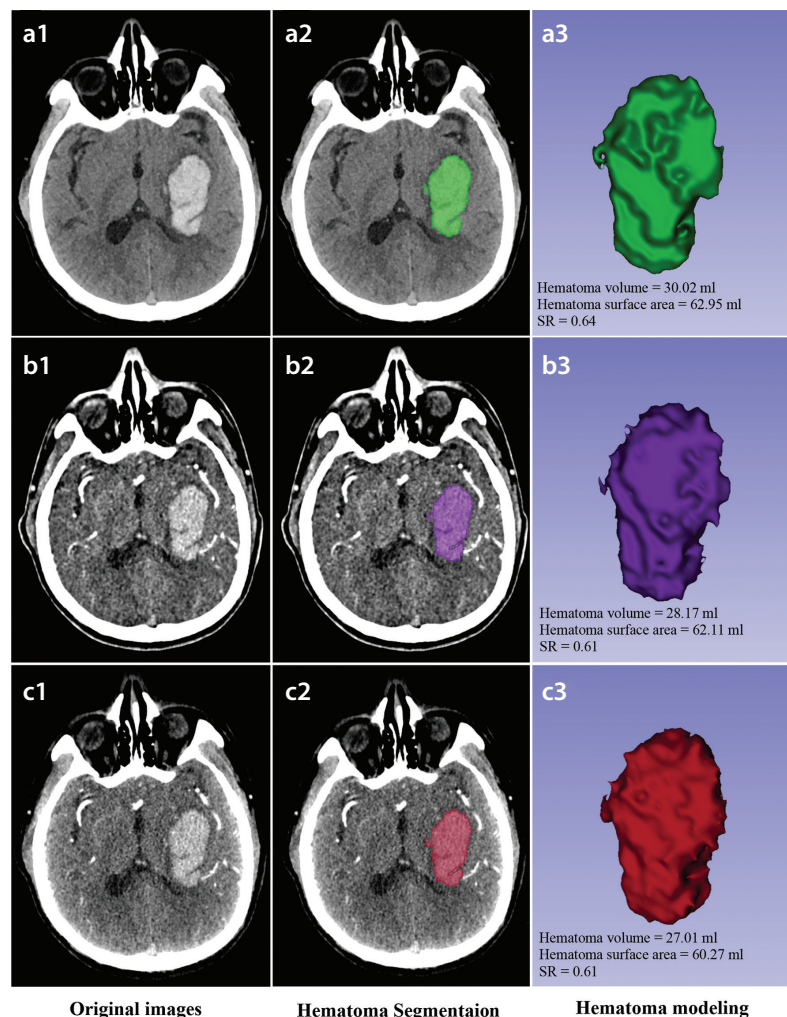


Figure 2. Hematoma segmentation and modeling of a patient with hemorrhage in the left basal ganglia region. Figures a1–a3 illustrate the hematoma segmentation and modeling results on NCCT using 3D Slicer software; Figures b1–b3 and c1–c3 depict the segmentation and modeling results for the same patient on conventional CTA and 60 keV VMI, respectively. Hematoma volume and surface area are displayed below the 3D models, and SR values were calculated accordingly. NCCT, non-contrast computed tomography; CTA, computed tomography angiography; SR, surface regularity; 3D, three dimensional; VMI, virtual monoenergetic imaging.

Table 1. The inter-reader ICC values of hematoma volume and surface area measured on NCCT, conventional CTA, and 60 keV VMI				
Modalities	Parameters	NCCT	Conventional CTA	60 keV VMI
	Volume	0.999	0.999	0.991
	Surface area	0.998	0.998	0.998
	Please note that these values were inter-reader ICC values. NCCT, non-contrast computed tomography; CTA, computed tomography angiography; VMI, virtual monoenergetic imaging; ICC, intraclass correlation coefficient.			

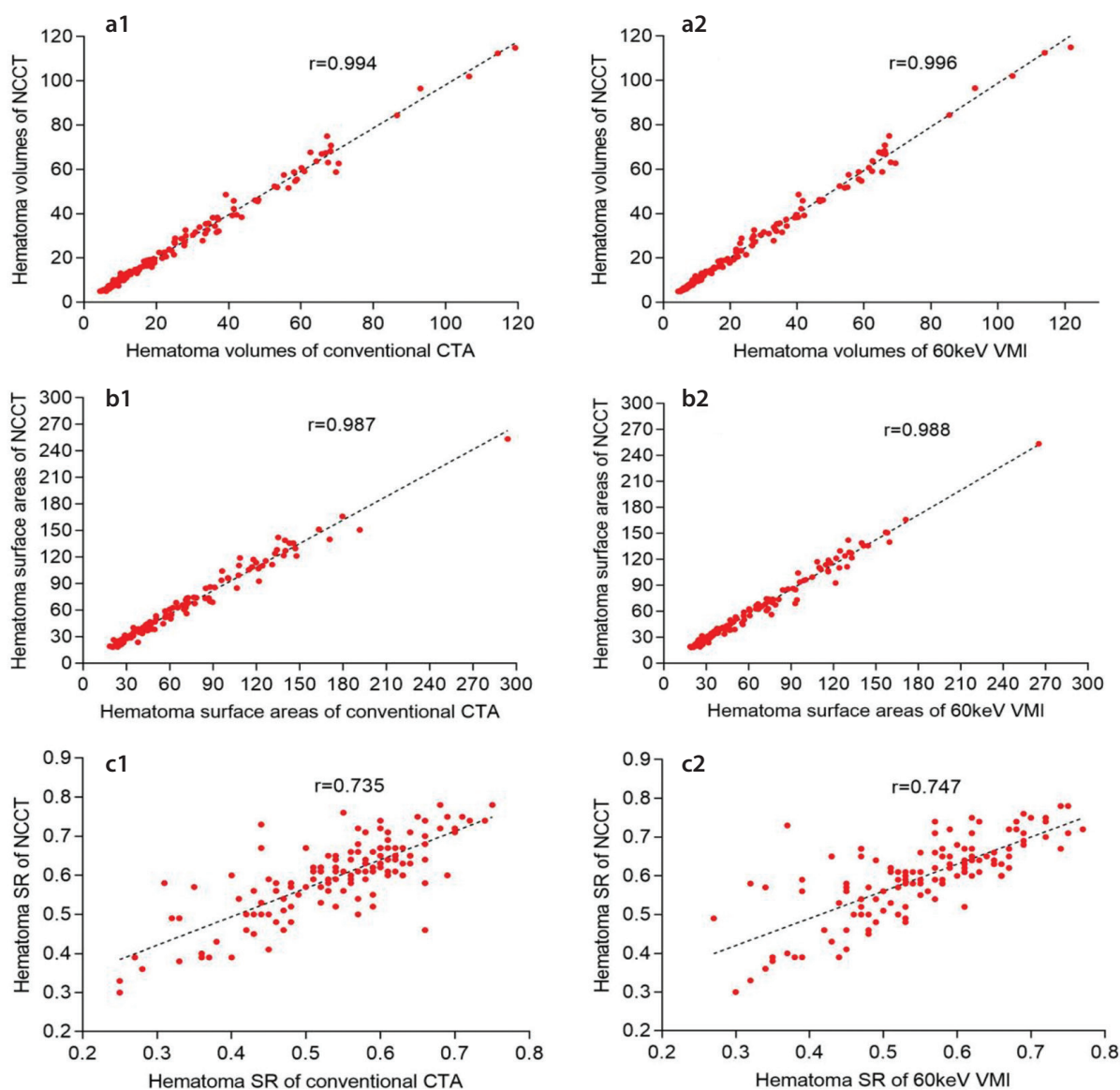


Figure 3. Simple linear correlation analysis between the volume, surface area, and SR of conventional CTA, 60 keV VMI, and those of NCCT. Figures a1–a2 and b1–b2 demonstrate that the volume and surface area of conventional CTA and 60 keV VMI are linearly correlated with those of NCCT, respectively (r approaching 1); Figures c1–c2 show a moderate linear correlation between the SR of conventional CTA, 60 keV VMI, and that of NCCT. NCCT, non-contrast computed tomography; CTA, computed tomography angiography; SR, surface regularity; VMI, virtual monoenergetic imaging.

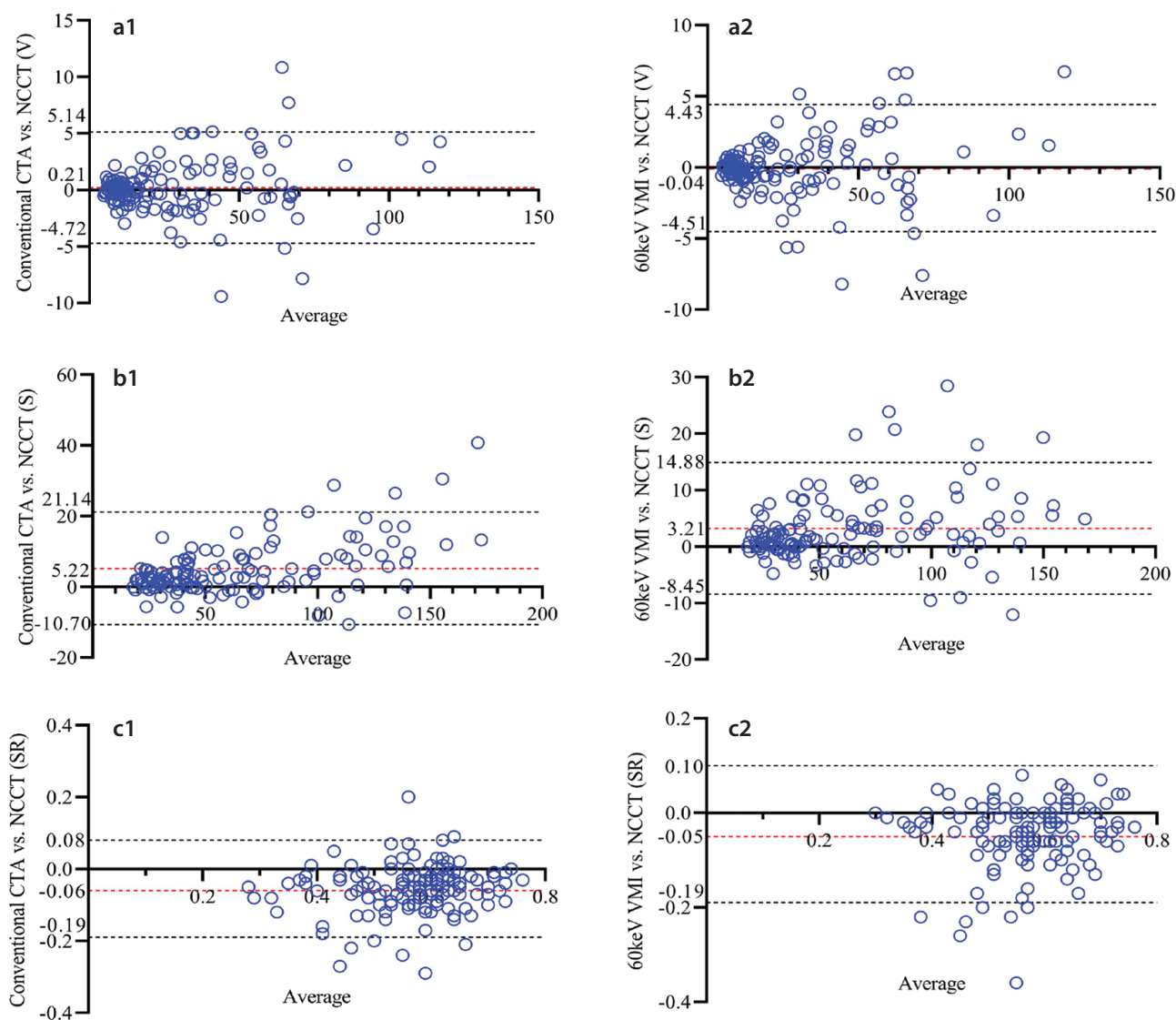


Figure 4. Bland–Altman scatter plots for V, S, and SR of conventional CTA/60 keV VMI vs. NCCT. Dashed red lines represent the mean bias of measurement deviations. Dashed black lines represent the 95% limits of agreement. CTA, computed tomography angiography; NCCT, non-contrast computed tomography; V, hematoma volume; S, surface area; SR, surface regularity; VMI, virtual monoenergetic imaging.

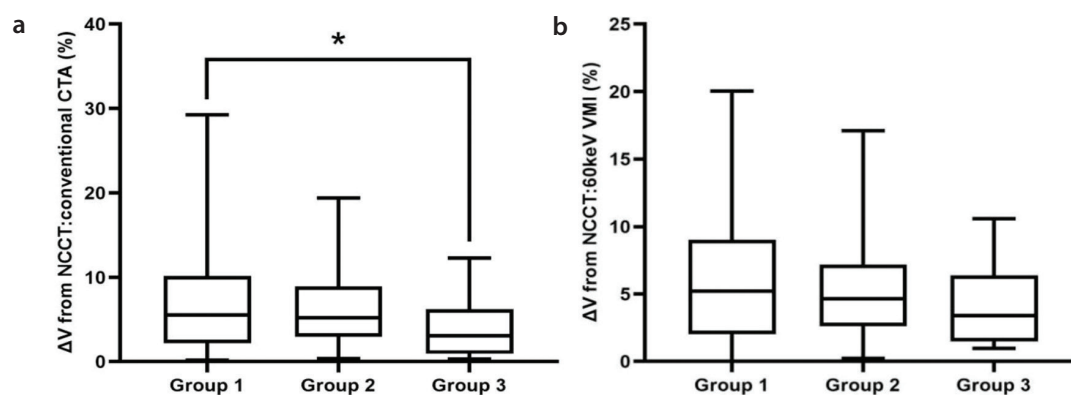


Figure 5. Comparison of ΔV among different hematoma volume groups in conventional CTA (a) and 60 keV VMI (b). The box plots display the upper quartiles, lower quartiles, and medians. Group 1: < 30 mL; Group 2: 30–60 mL; Group 3: > 60 mL. *indicates statistically significant difference. CTA, computed tomography angiography; VMI, virtual monoenergetic imaging; ΔV , deviation percentage in hematoma volume.

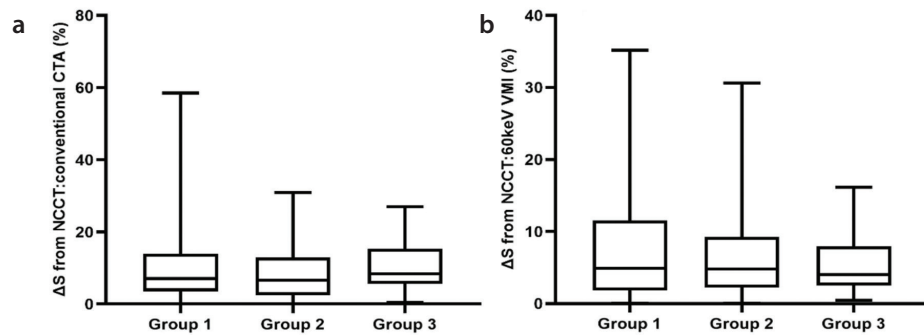


Figure 6. Comparison of ΔS among different hematoma volume groups in conventional CTA (a) and 60 keV VMI (b). The box plots display the upper quartiles, lower quartiles, and medians. Group 1: < 30 mL; Group 2: 30–60 mL; Group 3: > 60 mL. CTA, computed tomography angiography; VMI, virtual monoenergetic imaging; ΔS , deviation percentage in surface area.

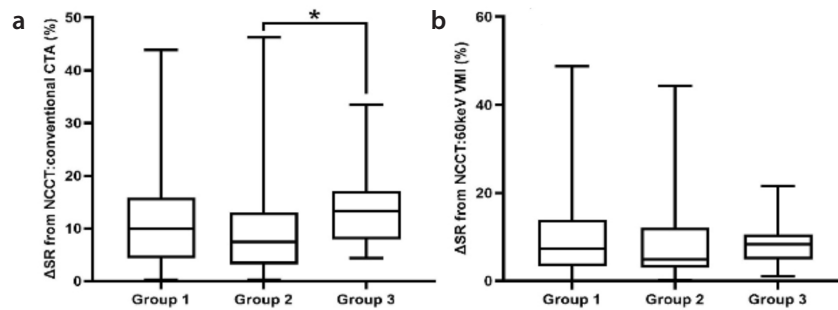


Figure 7. Comparison of ΔSR among different hematoma volume groups in conventional CTA (a) and 60 keV VMI (b). The box plots display the upper quartiles, lower quartiles, and medians. Group 1: < 30 mL; Group 2: 30–60 mL; Group 3: > 60 mL. *indicates statistically significant difference. CTA, computed tomography angiography; VMI, virtual monoenergetic imaging; ΔSR , deviation percentage in surface regularity.

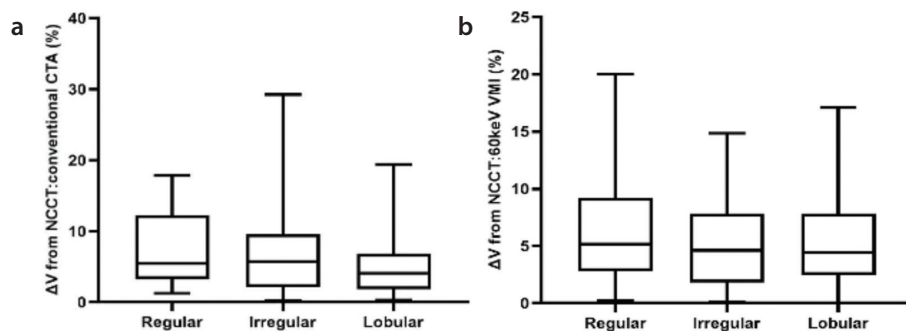


Figure 8. Comparison of ΔV among different hematoma morphology groups in conventional CTA (a) and 60 keV VMI (b). The box plots display the upper quartiles, lower quartiles, and medians. CTA, computed tomography angiography; VMI, virtual monoenergetic imaging; ΔV , deviation percentage in hematoma volume.

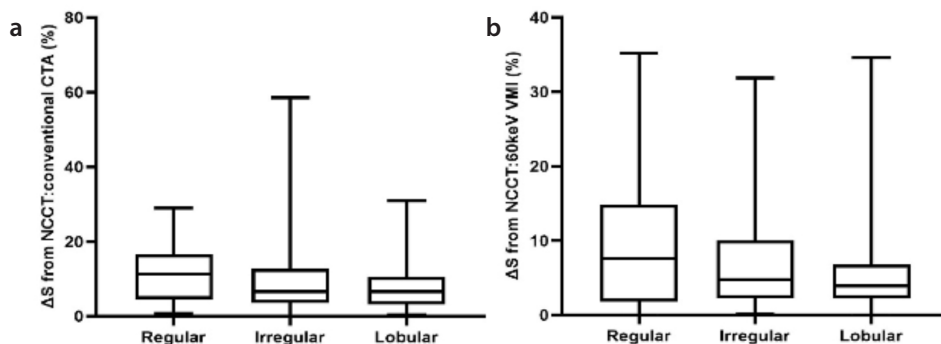


Figure 9. Comparison of ΔS among different hematoma morphology groups in conventional CTA (a) and 60 keV VMI (b). The box plots display the upper quartiles, lower quartiles, and medians. CTA, computed tomography angiography; VMI, virtual monoenergetic imaging; ΔS , deviation percentage in surface area.

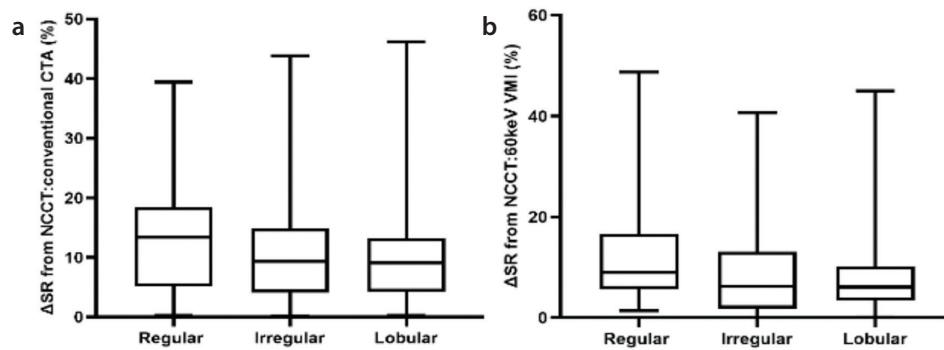


Figure 10. Comparison of Δ SR among different hematoma morphology groups in conventional CTA (a) and 60 keV VMI (b). The box plots display the upper quartiles, lower quartiles, and medians. CTA, computed tomography angiography; VMI, virtual monoenergetic imaging; Δ SR, deviation percentage in surface regularity.

Discussion

This study evaluated the accuracy of hematoma measurements on DECTA. To our knowledge, it is among the first to comprehensively analyze the accuracy of hematoma V, S, and SR measurements on conventional CTA and 60 keV VMI from DECTA compared with NCCT. We observed excellent linear correlations for V and S between conventional CTA/60 keV VMI and NCCT, regardless of V and shape. Bland–Altman analysis confirmed good agreement for V, with minimal mean bias and 95% LoA ranges narrower than established thresholds for HE (6 mL)⁵ and surgical decision-making (30/60 mL).¹³ However, SR was consistently underestimated on DECTA. In volume-based analysis, Δ S for conventional CTA and Δ V/ Δ S/ Δ SR for 60 keV VMI showed no significant differences across groups, suggesting error magnitudes may scale similarly with volume. Notably, for conventional CTA, the estimation error of V was smaller for large hematomas (> 60 mL), whereas the estimation error of SR was larger. No significant differences were found across morphology groups for any parameter.

Accurate measurement of V is crucial for assessing prognosis and planning treatment.² Although NCCT is the standard, clinicians may need to assess V on other modalities, and its accuracy remains unclear. Schlunk et al.¹⁹ reported that CTA and magnetic resonance imaging could significantly over- or underestimate volume compared with post-contrast CT, considered the most accurate modality. In contrast, our study found no significant systematic bias in volume measurements between DECTA and NCCT. This discrepancy may be due to Schlunk et al.'s¹⁹ smaller sample size (CTA: n = 20) or methodological differences, including the use of DECTA with advanced post-processing.

CTA provides additional diagnostic value (e.g., detecting etiologies and the spot sign). Our study demonstrated that the narrow 95% LoA range for volume measurement on DECTA (approximately $\pm$ 5 mL) was lower than the 6 mL threshold for HE diagnosis and the 30/60 mL thresholds for surgical decision-making. Therefore, V measured on DECTA may be comparable to that measured on NCCT in certain clinical scenarios, thereby adding clinical value for patients undergoing CTA.

Although hematoma surface properties contain information on growth dynamics,⁵ they are less studied. Fisher's "domino" model suggests vessel shearing causes irregular growth.⁵ As a quantitative 3D morphology metric (adapted from tumor imaging),²⁰ SR has been linked to HE dynamics.¹⁰ Our finding of significantly lower SR on DECTA indicates that SR reproducibility across these modalities is inferior to size-based metrics (V, S), possibly due to contrast enhancement affecting margin delineation. Therefore, the value of SR measured on contrast-enhanced images requires further investigation.

Our study has several limitations. First, the retrospective design and moderate sample size may limit generalizability. Second, we focused on conventional CTA and 60 keV VMI; VNC images, which may more closely approximate NCCT, were not evaluated and warrant future study. Finally, intraventricular hemorrhage cases were excluded, which merits separate investigation.

In conclusion, V measurements from DECTA show good agreement with those of NCCT, indicating potential utility for follow-up assessment in specific clinical scenarios. However, DECTA tends to underestimate SR compared with NCCT. Furthermore, conventional CTA derived from DECTA demon-

strated higher volumetric accuracy for large hematomas (> 60 mL) compared with smaller ones.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

This work was supported by the Fujian Provincial Science and Technology Department, Fujian, China [grant number 2023J011727].

References

1. Mazzoleni V, Padovani A, Morotti A. Emergency management of intracerebral hemorrhage. *J Crit Care.* 2023;74:154232. [\[Crossref\]](#)
2. Teo KC, Fong SM, Leung WCY, et al. Location-specific hematoma volume cutoff and clinical outcomes in intracerebral hemorrhage. *Stroke.* 2023;54(6):1548-1557. [\[Crossref\]](#)
3. Huang YW, Huang HL, Li ZP, Yin XS. Research advances in imaging markers for predicting hematoma expansion in intracerebral hemorrhage: a narrative review. *Front Neurol.* 2023;14:1176390. [\[Crossref\]](#)
4. Cliteur MP, Sondag L, Cunningham L, et al. The association between perihematoma oedema and functional outcome after spontaneous intracerebral haemorrhage: a systematic review and meta-analysis. *Eur Stroke J.* 2023;8(2):423-433. [\[Crossref\]](#)
5. Lv XN, Deng L, Yang WS, Wei X, Li Q. Computed tomography imaging predictors of intracerebral hemorrhage expansion. *Curr Neurol Neurosci Rep.* 2021;21(5):22. [\[Crossref\]](#)
6. Song L, Qiu X, Zhang C, et al. Combining non-contrast CT signs with onset-to-imaging time to predict the evolution of intracerebral hemorrhage. *Korean J Radiol.* 2024;25(2):166-178. [\[Crossref\]](#)

7. Wang CW, Liu YJ, Lee YH, et al. Hematoma shape, hematoma size, Glasgow coma scale score and ICH score: which predicts the 30-day mortality better for intracerebral hematoma? *PLoS One*. 2014;9(7):e102326. [\[Crossref\]](#)
8. Blacquiere D, Demchuk AM, Al-Hazzaa M, et al. Intracerebral hematoma morphologic appearance on noncontrast computed tomography predicts significant hematoma expansion. *Stroke*. 2015;46(11):3111-3116. [\[Crossref\]](#)
9. Delcourt C, Zhang S, Arima H, et al. Significance of hematoma shape and density in intracerebral hemorrhage: the intensive blood pressure reduction in acute intracerebral hemorrhage trial study. *Stroke*. 2016;47(5):1227-1232. [\[Crossref\]](#)
10. Oge DD, Topcuoglu MA, Gocmen R, Arsava EM. The dynamics of hematoma surface regularity and hematoma expansion in acute intracerebral hemorrhage. *J Clin Neurosci*. 2020;74:160-163. [\[Crossref\]](#)
11. Hiramatsu M, Haruma J, Hishikawa T, Sugi K, Date I. Intracerebral hemorrhage. *No Shinkei Geka*. 2021;49(2):284-292. [\[Crossref\]](#)
12. Tran NA, Sodickson AD, Gupta R, Potter CA. Clinical applications of dual-energy computed tomography in neuroradiology. *Semin Ultrasound CT MR*. 2022;43(4):280-292. [\[Crossref\]](#)
13. Hemphill JC 3rd, Greenberg SM, Anderson CS, et al. Guidelines for the management of spontaneous intracerebral hemorrhage: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2015;46(7):2032-2060. [\[Crossref\]](#)
14. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*. 2012;30(9):1323-1341. [\[Crossref\]](#)
15. Xu X, Chen X, Zhang J, et al. Comparison of the Tada formula with software slicer: precise and low-cost method for volume assessment of intracerebral hematoma. *Stroke*. 2014;45(11):3433-3435. [\[Crossref\]](#)
16. Chen M, Li Z, Ding J, Lu X, Cheng Y, Lin J. Comparison of common methods for precision volume measurement of hematoma. *Comput Math Methods Med*. 2020;2020:6930836. [\[Crossref\]](#)
17. Schneider D, Apfalter P, Sudarski S, et al. Optimization of kiloelectron volt settings in cerebral and cervical dual-energy CT angiography determined with virtual monoenergetic imaging. *Acad Radiol*. 2014;21(4):431-436. [\[Crossref\]](#)
18. Wichmann JL, Nöske EM, Kraft J, et al. Virtual monoenergetic dual-energy computed tomography: optimization of kiloelectron volt settings in head and neck cancer. *Invest Radiol*. 2014;49(11):735-741. [\[Crossref\]](#)
19. Schlunk F, Kuthe J, Harmel P, et al. Volumetric accuracy of different imaging modalities in acute intracerebral hemorrhage. *BMC Med Imaging*. 2022;22(1):9. [\[Crossref\]](#)
20. Pérez-Beteta J, Molina-García D, Ortiz-Alhambra JA, et al. Tumor surface regularity at MR imaging predicts survival and response to surgery in patients with glioblastoma. *Radiology*. 2018;288(1):218-225. [\[Crossref\]](#)