EISSN 1305-3612

DIR

Diagnostic and Interventional Radiology

TURKISH SOCIETY OF RADIOLOGY

dirjournal.org

VOLUME 31
ISSUE 2
March 2025

Official Journal of the Turkish Society of Radiology

E-ISSN: 1305-3612

www.dirjournal.org

DIR
Diagnostic and Interventional
Radiology

Official Journal of the Turkish Society of Radiology

E-ISSN: 1305-3612
www.dirjournal.org

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 (Index Medicus® abbreviation: Tani Girisim Radyol), the current title's exact Turkish translation.

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 ULAKBIM TR Index, DOAJ, HINARI, EMBASE, CINAHL, Scopus, Gale and CNKI.**

The journal is published online.

Owner: Can Çevikol on behalf of Turkish Society of Radiology

Responsible Manager: Mehmet Ruhi Onur

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
galenos
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: March 2025
EISSN 1305-3612
International scientific journal published bimonthly.

Turkish Society of Radiology is one of the foremost medical specialty organizations in Türkiye. It was formed by the merger of the two main radiology societies of Türkiye, one of which was founded in 1924. The Society is based in Ankara, Türkiye.

TSR 1924
TURKISH SOCIETY OF RADIOLOGY

Diagnostic and Interventional Radiology

Volume 31 • Issue 2

March 2025

www.dirjournal.org

Contents

Full text of these articles can be accessed online at www.dirjournal.org or through PubMed (https://www.ncbi.nlm.nih.gov/pmc/journals/2754/).

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242827

ABDOMINAL IMAGING

INVITED REVIEW

Unusual liver tumors: spectrum of imaging findings with pathologic correlation

Nir Stanietzky[1]
Ahmed Ebada Salem[2,3]
Khaled M. Elsayes[1]
Maryam Rezvani[2]
Sarah Palmquist[1]
Imran Ahmed[1]
Ahmed Marey[3]
Silvana Faria[1]
Ayman H. Gaballah[1]
Christine O. Menias[4]
Akram M. Shaaban[2]

[1]University of Texas MD Anderson Cancer Center, Department of Radiology, Texas, USA

[2]The University of Utah, Department of Radiology, Utah, USA

[3]Alexandria University Faculty of Medicine, Department of Radiodiagnosis and Intervention, Alexandria, Egypt

[4]Mayo Clinic, Department of Radiology, Arizona, USA

ABSTRACT

The liver is a common location for both primary and secondary cancers of the abdomen. Radiologists become familiar with the typical imaging features of common benign and malignant liver tumors; however, many types of liver tumors are encountered infrequently. Due to the rarity of these lesions, their typical imaging patterns may not be easily recognized, meaning their underlying pathologic features may not be discovered or suggested until an invasive biopsy is performed. In this review article, we discuss multiple hepatic neoplasms that are both unusual and rare. Some have typical imaging patterns, whereas others are non-specific and can only be included in the differential diagnosis. The clinical history and serologic findings are often critical in suggesting these entities; therefore, these are also discussed to familiarize the radiologist with the appropriate clinical setting of each. The article includes an image-rich description of each entity with accompanying figures describing the ultrasonography, computed tomography, and magnetic resonance imaging features of each disease process. Novel therapies and prognosis of several of the diseases are also included in the discussion.

KEYWORDS

Clinical context, differential diagnosis, hepatic neoplasms, imaging features, pathologic correlation

Radiologists who are unfamiliar with the many etiologies of unusual hepatic tumors may misinterpret these lesions. Some present with unique imaging features, whereas others present in a similar fashion to common neoplasms. This article will serve as a useful reference for both general and subspecialized radiologists when encountering such lesions.

Primary hepatic neuroendocrine tumors

Intrabdominal neuroendocrine tumors (NET) typically originate from the gastrointestinal tract, specifically the appendix, ileum, and rectum. The liver is a common site for NET metastases; however, primary hepatic neuroendocrine tumors (PHNETs) are extremely rare and are believed to arise either from ectopic pancreatic cells or stem cells in the liver. As PHNETs are usually slow growing, they are typically discovered incidentally.[1] The most common ages of presentation are 40–50 years, and the tumor tends to be hormonally inactive, with non-specific clinical symptoms, ranging from asymptomatic to abdominal pain.[2] If hormonal symptoms occur, the patient typically demonstrates carcinoid syndrome or Cushing syndrome.

On imaging, PHNET presents as a large mixed cystic and solid lesion with surrounding satellite nodules. The solid component often demonstrates a hypervascular enhancement on the arterial phase, more so in the periphery, with delayed enhancement centrally (Figure 1). On magnetic resonance imaging (MRI), there is hyperintense T2 weighted signal and marked restriction on diffusion-weighted imaging (DWI).[3] The tumor can produce tumor thrombus,[1] and can be confused with other arterially enhancing lesions, such as hepatocellular carcinomas (HCCs). However, PHNETs do not tend to occur in patients with cirrhosis or chronic liver disease.

Corresponding author: Khaled M. Elsayes

E-mail: kmelsayes@mdanderson.org

Received 25 April 2024; revision requested 12 May 2024; accepted 18 May 2024.

Epub: 10.06.2024

Publication date: 03.03.2025

DOI: 10.4274/dir.2024.242827

You may cite this article as: Stanietzky N, Salem AE, Elsayes KM, et al. Unusual liver tumors: spectrum of imaging findings with pathologic correlation. *Diagn Interv Radiol.* 2025;31(2):58-67.

Depending on the tumor grade differentiation and Ki-67 proliferation index, these lesions may demonstrate uptake on fluorodeoxyglucose (^{18}F-FDG)-positron emission tomography (PET)/computed tomography (CT). Low-grade tumors are typically hypometabolic, whereas grade 2 tumors can be hypo- or hypermetabolic. In contrast, grade 3, poorly differentiated neuroendocrine neoplasms are typically ^{18}F-FDG-PET/CT avid. Gallium-68-DOTA-somatostatin analogue-PET/CT and Indium-111 octreotide scanning, which are specific receptor imaging techniques, demonstrate a higher positive imaging rate for grade 1 and grade 2 tumors.[4]

No global consensus on the treatment of these lesions exists. Surgical resection is the treatment of choice, with a reported 10-year survival rate of 68%.[1,4] For patients demonstrating recurrence or who are not eligible for surgery, transcatheter chemoembolization can be used, with a 5-year survival rate of 74%–78%.[1] Other treatment options include yttrium-90- and lutetium-177-labelled peptides. There is limited data on the effect of chemotherapy on the treatment and prognosis of PHNETs.[1]

Extrapulmonary small cell carcinoma

The lung is the most common site of small cell carcinoma (SCC). Extrapulmonary SCC (EPSCC) usually occurs in the gastrointestinal tract and accounts for only 2.5%–5.0% of SCC,[2] with around 1,000 cases diagnosed in the United States per year. Both EPSCC and small cell lung cancer (SCLC) share some histopathologic features with NETs; EPSCC demonstrates a slight male predominance and presents at a mean age of 64, approximately 5–10 years earlier than SCC of the lung. The proportion of patients with EPSCC who smoke is lower than in SCC of the lung.[5] These extrapulmonary tumors typically appear as a large, heterogeneous mass with non-specific imaging findings, and are indistinguishable from other common hepatic neoplasms (Figure 2).[6] Once hepatic EPSCC is diagnosed through biopsy, an extensive diagnostic workup including chest CT, PET/CT, and bronchoscopy is critical to exclude an extrahepatic primary site. As the liver is also the most common site of metastatic disease in other forms of EPSCC, determining the site of primary disease can be challenging when more than one organ is involved.[5]

The management of EPSCC is extrapolated from the treatment of SCLC due to the similar histologic features. However, this approach has limited evidence-based data.[5] As with other NETs, the Ki-67 proliferation index is used to determine the grade. Unlike many other neuroendocrine neoplasms, EPSCC does not show a direct correlation between grade and aggressiveness; in fact, one study showed a higher number of metastases in tumors with a lower Ki-67 index.[5] The response rate to chemotherapy is higher than that of SCLC. Of all the types of EPSCC, those originating in the gastrointestinal tract have the poorest 3-year survival rate (7% vs. an overall rate of 28%).[6]

Undifferentiated embryonal sarcoma

Undifferentiated embryonal sarcoma (UES) is a rare, highly aggressive malignant tumor of mesenchymal origin most commonly affecting children aged 6–10, with a slight male predominance.[7,8] Although this tumor is rare, it is the third most common primary hepatic tumor in children after hepatoblastoma and HCC. This tumor is typically asymptomatic in children and can present with abdominal pain and fever in adults. Rarely, patients may present with an acute abdomen due to tumor rupture. In contrast to other pediatric liver tumors, such as hepatoblastoma and HCC, UES usually presents with normal alpha-fetoprotein levels, whereas hepatoblastoma presents with elevated alpha-fetoprotein levels in 95% of cases. The most common sites of UES metastasis are the lung, pleura, and peritoneum.[8]

On imaging, the UES tumor has a predilection for the right hepatic lobe, is large (approx. 10–29 cm), and is predominantly cystic in appearance due to the high water content of its myxoid stroma. Post-contrast imaging shows progressive delayed enhancement of

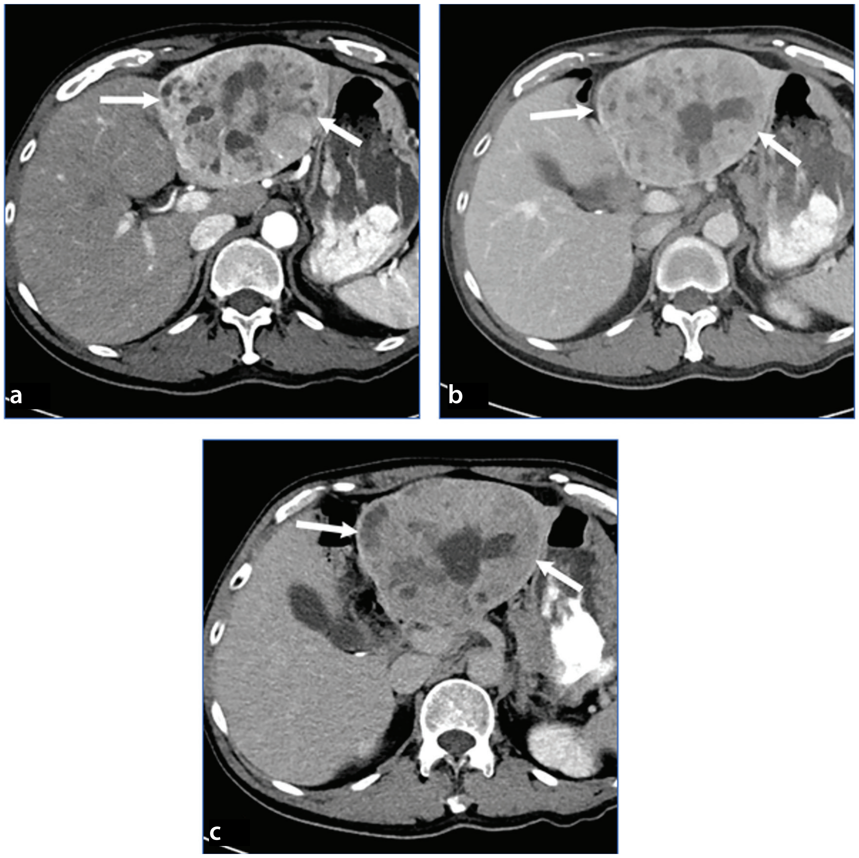


Figure 1. Primary hepatic neuroendocrine tumor. Axial contrast-enhanced computed tomography images of the liver during late arterial **(a)**, portal venous **(b)**, and delayed **(c)** phases of contrast enhancement show a round, heterogeneously enhancing primary hepatic neuroendocrine tumor replacing the lateral segments of the left lobe of the liver (arrows). The mass demonstrates increased enhancement during the arterial phase (particularly peripherally), washout of contrast material during the portal phase, and increased enhancement on the delayed phase as compared with the surrounding liver parenchyma.

Main points

- Unusual hepatic tumors are infrequently seen and it is therefore important for radiologists to be familiar with their imaging findings.
- While the imaging findings of many of these unusual tumors are non-specific, familiarity with these disease entities allows for their inclusion in the differential diagnosis.
- The clinical features of these entities are also described to aid in the differential diagnosis.

a thick peripheral rim, which corresponds to a fibrous pseudo capsule.[8] CT demonstrates a fluid attenuating mass with thick peripheral rim of soft tissue. Calcifications are not typically present. Obtaining a delayed phase can aid in making an accurate diagnosis since delayed enhancement would not be seen in a simple hepatic cyst.[7]

Moreover, MRI shows a predominantly cystic-appearing mass with similar signal intensity to cerebrospinal fluid and a thick rim with low signal and delayed enhancement on both T1- and T2-weighted imaging, corresponding to the fibrous pseudocapsule (Figure 3). The tumor may contain focal areas of hyperintense signal on T1-weighted images, correlating to areas of intratumoral hemorrhage.[8]

Ultrasonography typically shows a solid isoechoic to hyperechoic mass relative to the background liver with varying degrees of anechoic regions, which correspond to internal necrosis and cystic degeneration.[8] A cystic-appearing mass on CT and MRI that appears solid on ultrasonography favors the diagnosis of UES.

The differential diagnosis includes mesenchymal hamartoma of the liver, which can be difficult to distinguish from UES on pathology and imaging. The age of presentation can help guide the diagnosis, as UES is rare in children under 5 years, whereas mesenchymal hamartoma of the liver typically presents by 2 years. Due to its predominantly cystic appearance on cross-sectional imaging, UES can easily be misdiagnosed as a hydatid cyst or abscess. Thorough clinical workup to look for peripheral eosinophilia seen with hydatid cysts and signs of infection seen with abscesses can aid in proper diagnosis.[8] Treatment consists of multiagent chemotherapy followed by surgery in cases amenable to resection.[7]

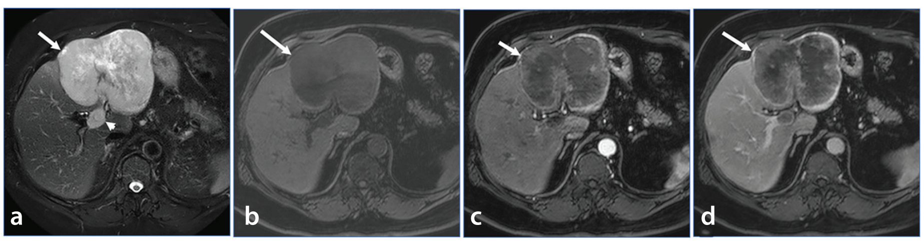


Figure 2. Extrapulmonary small cell carcinoma. Axial T2-weighted magnetic resonance imaging (MRI) with fat suppression **(a)** and pre-contrast **(b)** and dynamic post-gadolinium T1-weighted MRI with fat suppression in the arterial **(c)** and portal venous **(d)** phases show an extrapulmonary small cell carcinoma of the left hepatic lobe (white arrows) demonstrating high T2 signal intensity, low T1 signal intensity, and intense peripheral enhancement and poor central enhancement, with invasion of the left portal vein (short white arrow).

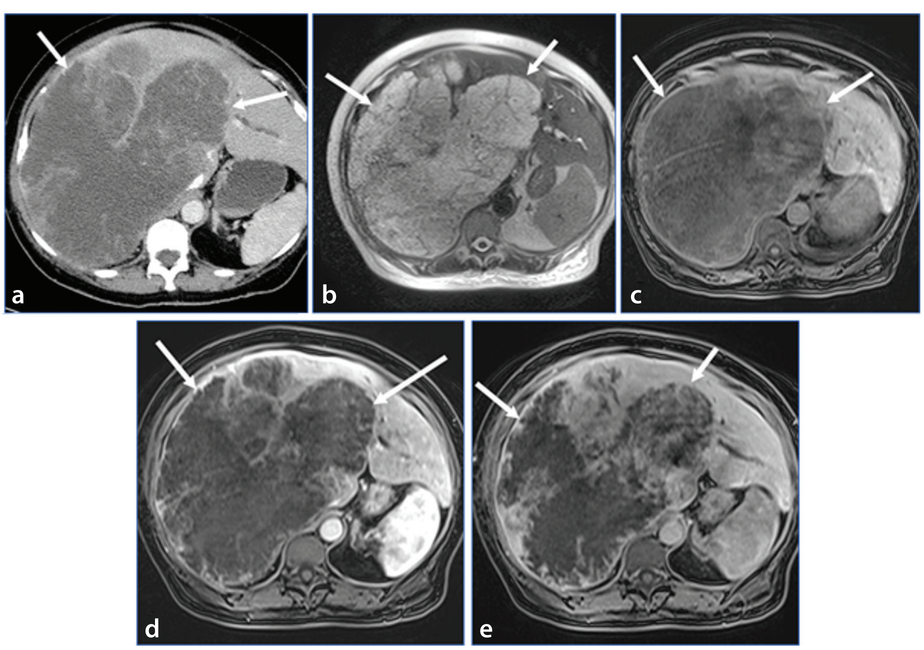


Figure 3. Undifferentiated embryonal sarcoma. Axial contrast-enhanced computed tomography **(a)** demonstrates a large predominantly cystic mass involving most of the right and part of the left hepatic lobe. Axial T2-weighted imaging **(b)** shows high signal intensity of the tumor, giving a cystic appearance. Axial dynamic gadolinium-enhanced T1-weighted imaging **(c-e)** show gradual contrast accumulation, revealing the solid nature of the tumor. This was pathologically proven to be embryonal sarcoma with possible cartilaginous differentiation.

Angiomyolipoma

Hepatic angiomyolipomas (AMLs) are rare, benign, mesenchymal tumors that consist of blood vessels, smooth muscle, and fat elements, and are more frequent in women and non-cirrhotic livers; AMLs more commonly occur in the kidneys and rarely involve the liver. An AML is associated with tuberous sclerosis in 20% of renal cases but only 6% of hepatic cases.[9] In most cases, patients are asymptomatic, and their hepatic AML is discovered incidentally. The imaging appearance varies depending on the degree of fat composition. The fat content is variable, ranging from 90% to barely detectable.[10] Other hepatic lesions can also contain fat, such as hepatic adenoma, HCC and, rarely, focal nodular hyperplasia.[10] Definitive diagnosis is based on pathologic evaluation of the smooth muscle component and positive staining for homatropine methyl bromide-45 and smooth muscle markers.[11]

On ultrasonography, hepatic AML appears highly echogenic and is indistinguishable from hemangioma (Figure 4). For lipid-rich AML, MRI evaluation demonstrates hyperintense signal on T1-weighted images with signal loss on fat suppression sequences, consistent with macroscopic fat. Distinguishing hepatic AML from HCC through imaging can be challenging. Some helpful AML features include isointensity on the portal venous phase, early draining veins, and intratumoral vessels. In addition, HCC frequently demonstrates restricted diffusion and a tumor capsule.[10] A small percentage (4%) of the epithelioid subtype of hepatic AML can demonstrate malignant behavior with local invasion, recurrence after resection, and metastasis.[11]

Angiosarcoma

Hepatic angiosarcoma is a malignant tumor that is extremely rare overall but is the most common hepatic mesenchymal tumor and has an extremely poor prognosis. It is more commonly seen in elderly men, and approximately one-fourth of cases are associated with exposure to thorium dioxide (Thorotrast) and vinyl chloride.

Clinically, patients typically present with hepatomegaly and other non-specific symptoms, such as abdominal pain, weight loss, and fatigue. The median survival is poor at just 6 months.[12] Large angiosarcomas can cause hematologic abnormalities, such as disseminated intravascular coagulation, thrombocytopenia, and microangiopathic hemolytic anemia. Metastasis is common at initial diagnosis, most commonly involving the spleen and lungs. Approximately 15%–27% of patients may present with acute abdominal pain and anemia due to tumor rupture and hemoperitoneum.[8] It is critical to be aware of potential massive hemorrhage as a complication of biopsy.

The tumor morphology of hepatic angiosarcoma can vary in appearance on imaging, showing multiple nodules/masses, a large dominant mass, or a diffuse infiltrative pattern (Figure 5). Intratumoral hemorrhage and necrosis are often present. On non-contrast CT, the tumor is hypoattenuating compared with normal background liver, with internal foci of hyperattenuation corresponding to hemorrhage. Contrast-enhanced CT shows intense peripheral nodular enhancement and can resemble a cavernous hemangioma but will not follow the blood pool on all phases and will generally not feature the true peripheral nodular discontinuous enhancement that is common in a cavernous hemangioma. More frequently, the tumor will appear hypodense on both arterial and portal venous phases with foci of early heterogenous enhancement, occasionally with a central or ring pattern, but to a lesser degree than the aorta. On delayed phases, the tumor shows persistent enhancement, but the complete centripetal fill-in seen in hemangiomas is absent.

On MRI, the tumor is predominantly hypointense on T1-weighted images with internal foci of hyperintensity corresponding to intratumoral hemorrhage. On T2-weighted images, the tumor is generally heterogeneously hyperintense compared with background liver and may contain septa or fluid–fluid levels related to hemorrhage.[8,13] Metastasis is common, affecting up to 60% of patients, and most commonly involves the lungs and spleen.[8] It is critical to assess the dependent areas of the abdomen and pelvis to check for hemoperitoneum in cases of tumor rupture. The treatment of these lesions includes surveillance or surgical resection and liver transplant in unresectable cases.[13]

Epithelioid hemangioendothelioma

Hepatic epithelioid hemangioendothelioma (HEHE) is an extremely rare malignant vascular tumor that typically presents in individuals in their 40s, more commonly in women, and the typical presentation includes abdominal pain, jaundice, and hepatosplenomegaly.[14] Involvement of other organs has been observed in 36.6% of patients, most often affecting lungs, regional lymph nodes, and peritoneum, with bones frequently affected.[15] A HEHE can mimic other tumors, most commonly cholangiocarcinoma, HCC, metastatic carcinoma, and angiosarcoma. Definitive diagnosis requires pathologic assessment, which shows endothelial cells, identifiable by positive staining with antibodies against factor VIII, CD31, and CD34.[16]

Imaging features of HEHE demonstrate multiple hypoattenuating nodules on

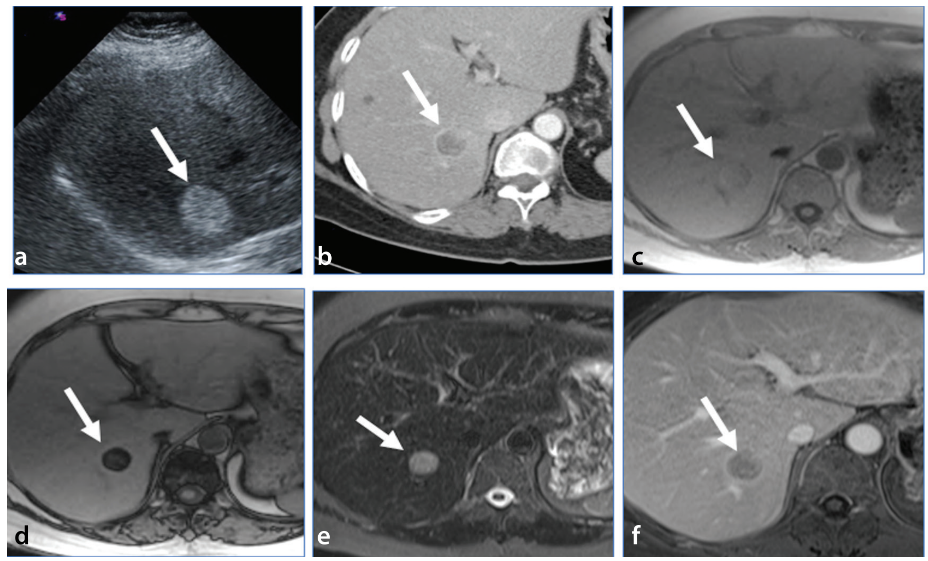


Figure 4. Angiomyolipoma. Transabdominal ultrasonography (**a**) shows a round echogenic mass (white arrow). Axial contrast-enhanced CT (**b**) shows a round, peripherally enhancing mass with poor central enhancement. Axial T1-weighted in-phase (**c**) and opposed-phase (**d**), axial T2-weighted (**e**), and post-gadolinium T1-weighted magnetic resonance images with fat suppression (**f**) show a round liver mass demonstrating loss of signal on opposed-phase images, high signal intensity on T2-weighted images, and rim enhancement following contrast administration.

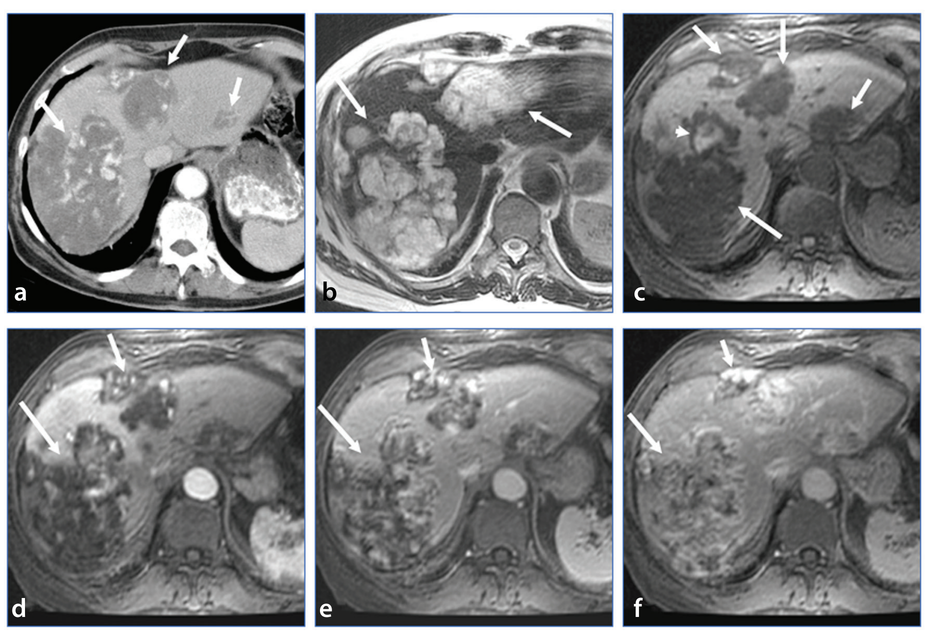


Figure 5. Angiosarcoma. Axial contrast-enhanced computed tomography (CT) (**a**), axial T2-weighted magnetic resonance imaging (MRI) (**b**), axial T1-weighted MRI (**c**), and dynamic post-gadolinium T1-weighted MRI with fat suppression (**d-f**) show multifocal liver angiosarcomas (long arrows). The masses demonstrate low attenuation relative to the liver on contrast-enhanced CT, high T2 signal intensity, low T1 signal intensity, and progressive enhancement following contrast administration. Early arterial enhancement is irregular and more central than in hemangiomas. High signal intensity within the lesion on T1-weighted imaging is due to hemorrhage. The T2 signal intensity is more heterogeneous than seen in hemangiomas.

non-contrast CT, which may or may not have calcifications. This tumor is most commonly subcapsular and can cause capsular retraction. Depending on the size of the lesion, these neoplasms can exhibit different patterns of contrast enhancement. Small lesions tend to demonstrate mild homogeneous enhancement; medium-size lesions can demonstrate ring enhancement, usually due to central necrosis; and large lesions demonstrate heterogeneous delayed enhancement.[17] Also helpful in the diagnosis are a "halo" sign and a "lollipop" sign, which show a branch of a hepatic vein draining the tumor.[18] Tumor thrombi may be present in the inferior vena cava.[19] Multiple lesions are more likely to occur in HEHE than other, more common hepatic tumors such as HCC, intrahepatic cholangiocarcinoma, and hepatic metastases (Figure 6).[15] However, as HEHE is a rare entity, multifocal liver masses are still more likely to represent these more common etiologies.

On MRI, HEHE tumors are typically T1 hypointense, heterogeneously T2 hyperintense, and diffusion restricting.[20] Ring enhancement is observed following intravenous contrast administration.[21] Relatively specific MRI features of HEHE are capsular retraction, lollipop sign, and "target" sign on both T2-weighted and portal phase imaging (Figure 7).[20]

On analysis with contrast-enhanced ultrasonography, HEHE demonstrates slower enhancement and more rapid washout time than more common hepatic tumors.[22] Moreover, HEHE can easily be misdiagnosed as hepatic metastases on ultrasonography given the common presentation of multiplicity and hypoechoic appearance.[23] Therefore, cross-sectional imaging is key for further evaluation.

Surgery is the treatment of choice for a confirmed case of unifocal HEHE and should be performed in centers with sarcoma surgery experience. There are no definitive guidelines for treating multifocal HEHE or metastatic EHE, and these cases are treated with a combination of chemotherapy, radiation therapy, surgery, and liver transplant.[24]

Hepatic schwannoma

Schwannomas (also called neurilemmomas) are benign, slow-growing nerve sheath tumors that typically occur in the head, neck, and upper extremities. These lesions can occur in all ages but are most common in women aged 20–50. Liver involvement of schwannomas is exceedingly uncommon, and when it occurs, it most often presents in patients with neurofibromatosis type 1 [50% of cases (25)] or following radiation. Hepatic schwannomas are believed to originate from nerve fibers that themselves originate from the plexus at the hepatic hilum. These fibers then branch out into the connective tissue along portal arteries and veins.[25]

On imaging, hepatic schwannomas demonstrate T1 hypointensity and T2 hy-

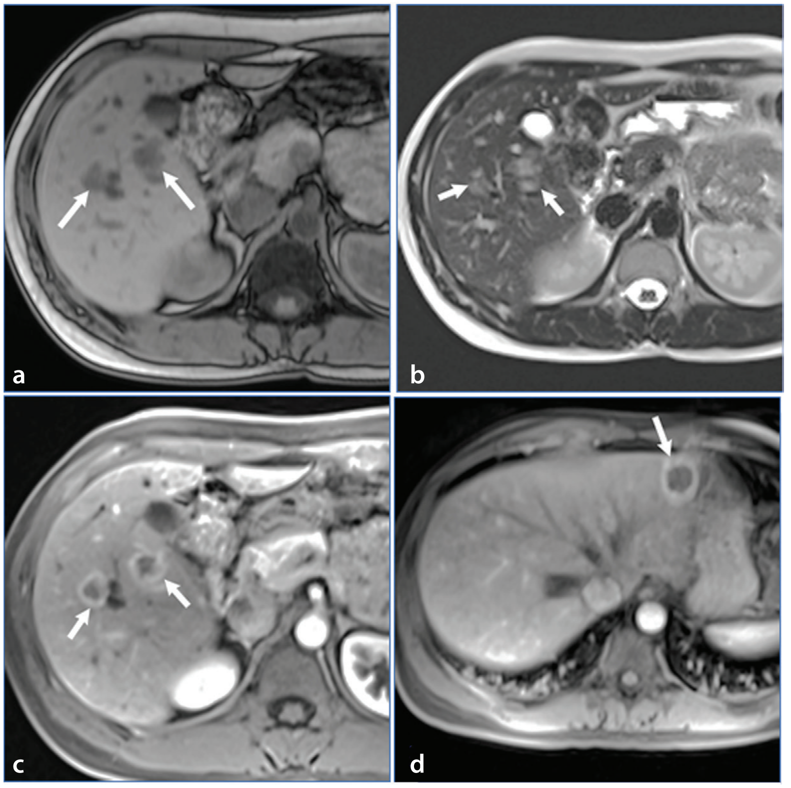


Figure 6. Epithelioid hemangioendothelioma, multiple. Axial T1-weighted (**a**) and axial T2-weighted (**b**) images and axial post-gadolinium T1-weighted images with fat suppression (**c, d**) show multiple liver lesions (arrows). The lesions demonstrate slightly high T2 signal intensity, low T1 signal intensity, and ring-like enhancement following contrast administration.

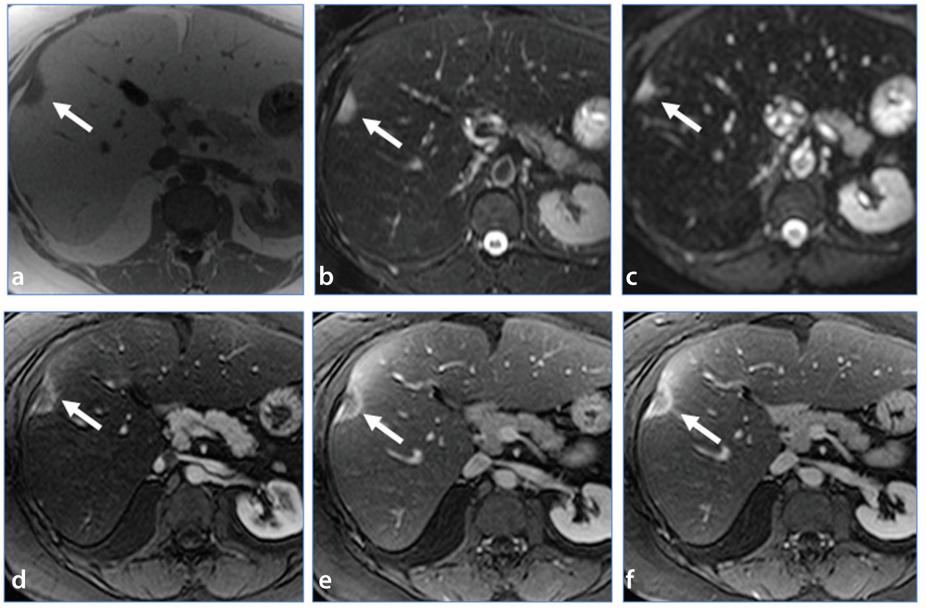


Figure 7. Epithelioid hemangioendothelioma, single. Axial T1-weighted imaging (**a**), axial T2-weighted imaging with fat suppression (**b**), axial diffusion-weighted imaging (**c**), and dynamic post-gadolinium T1-weighted imaging with fat suppression (**d-f**) show a subcapsular right lobe lesion (arrows) with capsular retraction. The mass demonstrates high T2 signal intensity, low T1 signal intensity, diffusion restriction, and progressive ring-like enhancement following contrast administration.

perintensity and have peripheral enhancement with central areas of patchy irregular enhancement on post-contrast imaging (Figure 8).[26] Rarely, these tumors can present with a multicystic appearance, with or without hemorrhage; this is more likely if the tumor is large.[25]

Schwannoma of the biliary tract can resemble cholangiocarcinoma, and patients may present with jaundice and abdominal pain, a situation that can lead to radiologic misdiagnosis and overtreatment of patients with these tumors.[27] Given the concern of biliary obstruction in certain cases, surgical resection is the preferred and curative treatment.[28]

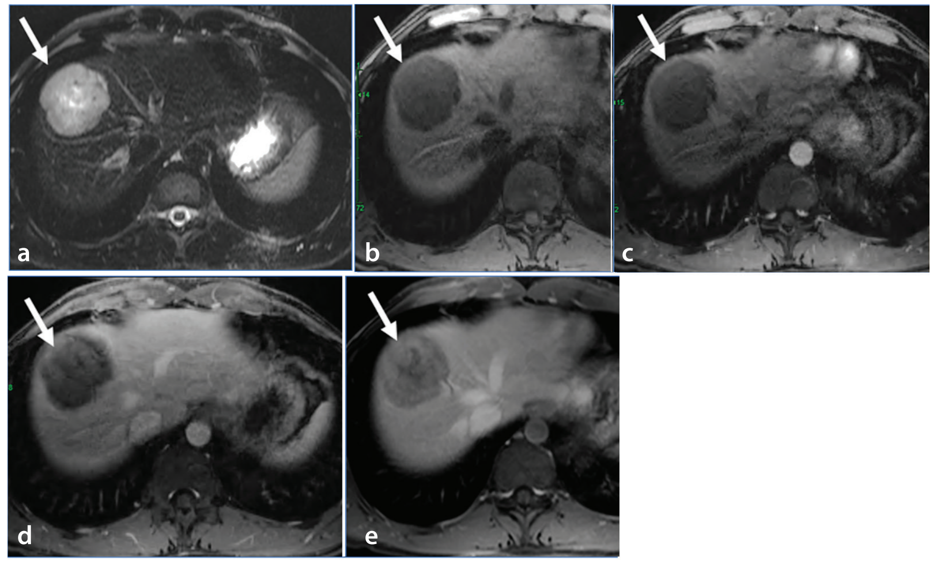


Figure 8. Schwannoma. Axial T2-weighted imaging with fat suppression **(a)** and dynamic post-gadolinium T1-weighted imaging with fat suppression (**b-e)** show a right hepatic lobe mass (arrows). The mass demonstrates high T2 signal intensity, low T1 signal intensity, and progressive enhancement following contrast administration.

Multiple myeloma and solitary plasmacytoma

Multiple myeloma is a malignancy of clonal plasma cell proliferation and is the second most common hematologic malignancy. Although plasma cell proliferation generally occurs inside the bone marrow, extramedullary involvement can also be observed. Extramedullary multiple myeloma (EMM) has a reported incidence of 7%–18% at presentation and 6%–20% during disease progression.[29] Liver involvement can be seen in up to 34% of patients with EMM. These patients can present with hepatomegaly, jaundice, ascites, and acute liver failure, and tend to have a poor prognosis.[29] Imaging features are variable, as EMM can present with a focal mass, multifocal lesions, or diffuse hepatomegaly. On ultrasonography, EMM lesions are usually hypoechoic (Figure 9). On CT, they appear hypoattenuating with mild enhancement, while they may present with low or high signal intensity in T1-weighted images and with a high T2 signal with mild enhancement. On FDG-PET/CT, EMM demonstrates moderate to intense FDG uptake.[30]

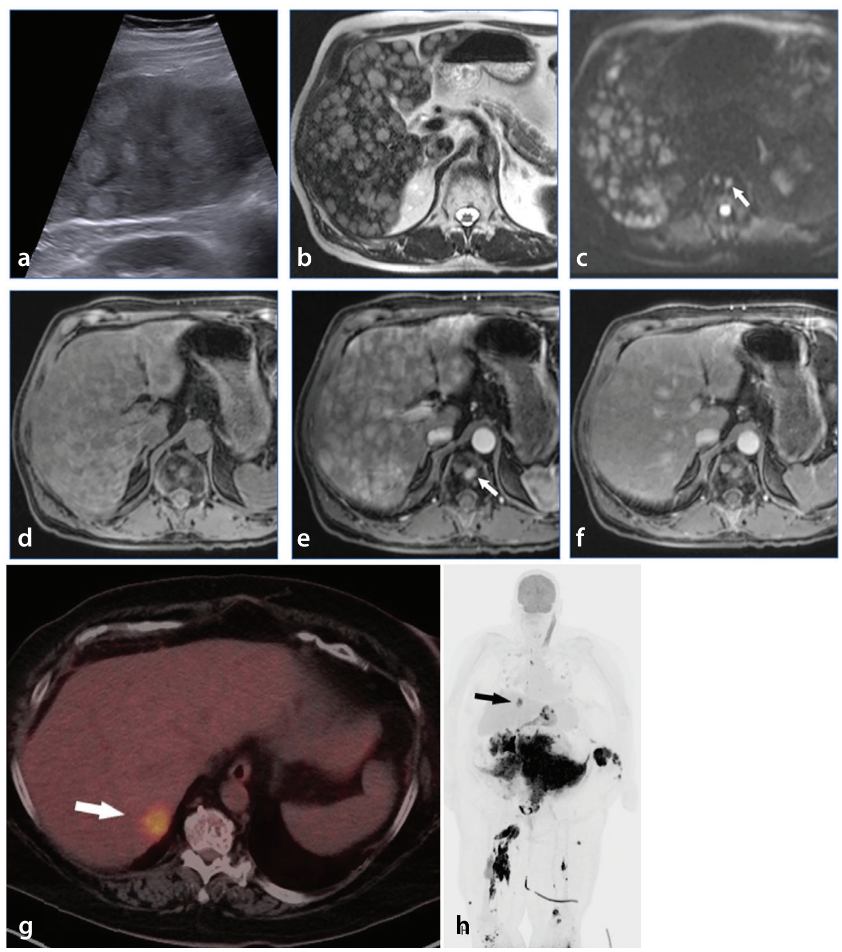


Figure 9. Multiple myeloma. Transabdominal ultrasonography **(a)** shows multiple hyperechoic liver lesions surrounded by a zone of low echogenicity resulting in target appearance. Axial T2-weighted imaging **(b)**, axial diffusion-weighted imaging (DWI) **(c)**, axial T1-weighted imaging **(d),** and post-gadolinium T1-weighted imaging with fat suppression during the arterial **(e)** and delayed **(f)** phases show numerous liver lesions demonstrating high T2 and low T1 signal intensity, with diffusion restriction on DWI. The lesions show intense enhancement in the arterial phase and are not visible in the delayed phase of contrast enhancement. A fluorodeoxyglucose-positron emission tomography/computed tomography scan **(g, h)** of a case of extramedullary multiple myeloma shows uptake (white and black arrows).

Solitary extramedullary plasmacytoma is a solitary mass of abnormal plasma cells in the absence of systemic myeloma. Hepatic solitary plasmacytoma is rare, and the imaging findings are variable. On FDG-PET/CT, the lesions are hypermetabolic. Patients with solitary plasmacytoma of the liver have a better prognosis than patients with systemic myeloma such as EMM.[31] Treatment includes autologous stem cell transplant and chemotherapy.

Hepatic lymphoma

Primary hepatic lymphoma (PHL) is an extremely uncommon variant of non-Hodgkin lymphoma (NHL), accounting for

0.016% of all NHL. PHL is confined to the liver and draining nodes, including the perihepatic and peripancreatic region. Unlike disseminated NHL with liver involvement, PHL shows no evidence of involvement of other visceral organs, distant lymph nodes, or bone marrow for at least 6 months after the onset of hepatic disease; PHL occurs more commonly in men and usually presents in patients in their mid-50s (range: 5–87). Patients may present with abdominal pain, constitutional symptoms, and B symptoms, such as fever and weight loss.[32]

The most common presentation of PHL is a solitary mass, while it can also present as multiple masses, and less commonly with diffuse hepatic involvement and a periportal pattern of distribution. On ultrasonography, these lesions are hypoechoic compared with normal liver parenchyma. On CT, the nodules are hypoattenuating with lower enhancement than the surrounding liver. On MRI, the nodules tend to be hypo- or isointense on T1-weighted images and hyperintense on T2-weighted images. Diffusion-weighted MRI is an important component of the imaging protocol due to the highly cellular nature of lymphoma, typically resulting in restricted diffusion in the diffusion-weighted sequences (Figure 10). The PET/CT technique is also helpful in diagnosis, and as with other types of lymphoma, hepatic lymphoma is typically extremely FDG avid.

A distinctive feature is that PHL tumoral masses have an insinuative growth behavior, encasing (not occluding) the vascular structures and bile ducts (Figure 11). Nonetheless, PHL patients are frequently misdiagnosed as having a primary liver tumor or metastatic cancer, and a definitive diagnosis can be achieved through histopathologic examination. Although PHL is an aggressive disease, it is resectable and responsive to chemotherapy and radiotherapy. Therefore, it should be considered in the differential diagnosis for patients presenting with mass lesions in the liver.[33,34]

Post-transplant lymphoproliferative disorder

Post-transplant lymphoproliferative disorder (PTLD) ranks as the second most common malignancy arising as a complication of solid organ transplant, following non-melanomatous skin cancer.[35] It is a disorder related to abnormal lymph node proliferation and encompasses a spectrum of disease processes ranging from benign lymphoid hyperplasia to high-grade malignant lymphomas.[36]

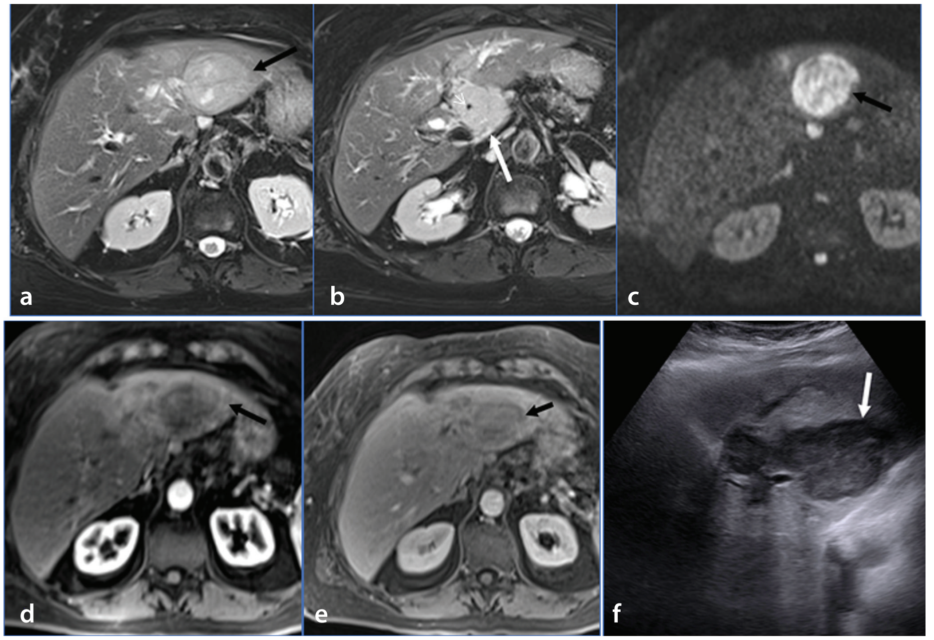


Figure 10. Hepatic lymphoma. Axial T2-weighted imaging **(a, b)**, diffusion-weighted b800 imaging **(c)**, and post-gadolinium T1-weighted imaging with fat suppression during the arterial **(d)** and delayed **(e)** phases. There is a well-circumscribed left lobe mass (black arrows) showing heterogeneous increased signal intensity on T2-weighted imaging, diffusion restriction, and poor enhancement following gadolinium administration. The adjacent left lobe demonstrates increased T2 signal intensity due to portal vein compression. A separate mass in the hepatic hilum (long white arrows) shows similar signal and enhancement characteristics. The hilar mass encases the hepatic artery (short white arrow). Transabdominal ultrasonography **(f)** shows a heterogeneous hypoechoic mass (white arrow).

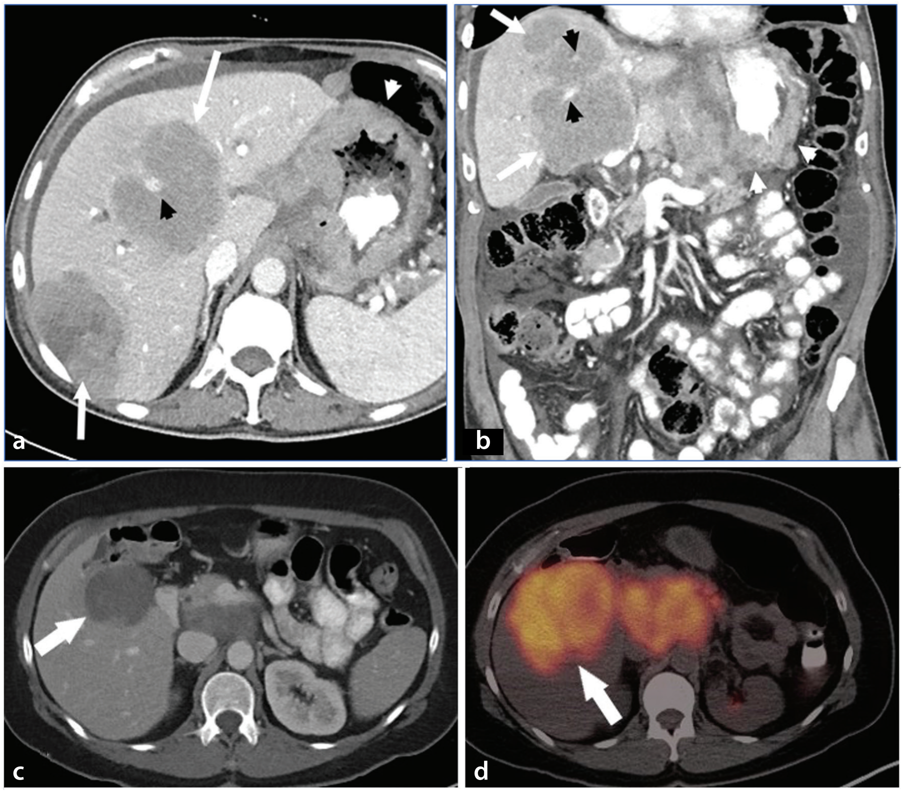


Figure 11. Hepatic lymphoma encasing vessels. Axial **(a)** and coronal **(b)** contrast-enhanced computed tomography (CT) shows multiple liver masses (long white arrows) and diffuse gastric wall thickening (short white arrows). As with lymphoma in other parts of the body, hepatic lymphoma tends to encase, rather than occlude, vascular structures. In this case, liver masses appear to encase branches of the portal vein (short black arrows). A companion case demonstrates a contrast enhanced CT scan **(c)** with a hypoattenuating lymphoma that is markedly fluorodeoxyglucose (FDG) avid on ^{18}F-FDG-positron emission tomography/CT scan **(d)** (white arrows).

Epstein–Barr virus (EBV) infection is a significant risk factor for the development of PTLD, particularly in transplant recipients who are EBV-seronegative prior to transplant. Other risk factors include young age, higher levels of immunosuppression following transplant, and having received a liver transplant within the past year.[37]

The reported incidence of PTLD in liver transplant recipients is variable. Taylor et al.[37] reported PTLD in up to 2.8% of adults and up to 15% of children following liver transplant. More recent studies showed a lower incidence of PTLD at 1.5% in adults and 4.3% in the pediatric population.[37] Generally, PTLD can be associated with significant morbidity and mortality, particularly in cases of high-grade lymphomas or when the disorder is diagnosed late.[35]

Unlike lymphoma, PTLD tends to involve extranodal sites such as the liver, and imaging is crucial to its evaluation. Hepatic involvement in PTLD can manifest in different forms and presentations. On CT imaging, PTLD may appear as multiple hypodense masses, a single infiltrating mass, or a heterogeneous mass at the liver hilum causing biliary obstruction (Figure 12). On MRI, the lesion or lesions often have isointense to low signal intensity on T1-weighted images and intermediate to high intensity on T2-weighted images. Dynamic T1-weighted post-contrast images may be characterized by peripheral enhancement, and DWI can show restricted diffusion. However, these imaging features can overlap with those of liver abscesses. This overlap can pose a diagnostic challenge, especially in patients who are at risk for both PTLD and disseminated infections.[36,38]

Early detection and management of PTLD are critical in improving outcomes for affected individuals. Treatment options such as reducing immunosuppression, antiviral therapy, rituximab (an anti-CD20 monoclonal antibody), chemotherapy, or radiation therapy may be considered depending on the severity and type of PTLD. Regular monitoring for EBV infection can help identify high-risk patients and allow for proactive interventions when necessary.

Hepatic benign cystic teratomas

Teratomas are germ cell tumors that originate from pluripotent cells that have been arrested along their migration pathway. They often contain components derived from all three germ cell layers and present as a cyst with fat, hair, and calcifications. Hepatic benign cystic teratomas are extremely rare, accounting for <1% of all body teratomas. They commonly occur in patients under 3 years old.[38,39]

Hepatic teratomas are often asymptomatic and may be discovered incidentally during imaging studies for unrelated reasons. However, large tumors can cause symptoms such as abdominal pain, discomfort, or fullness due to compression of neighboring organs. In exceptionally rare cases, hepatic cystic teratomas may rupture.[38,39]

On CT and MRI, hepatic teratomas typically appear as well-defined cystic lesions with heterogeneous internal components related to variable amounts of fat and calcifications (Figure 13).[38,39]

Surgical resection is the preferred treatment option for hepatic benign cystic teratomas, especially with large and symptomatic tumors. Complete surgical excision is usually curative, and recurrence is rare following successful resection.[39]

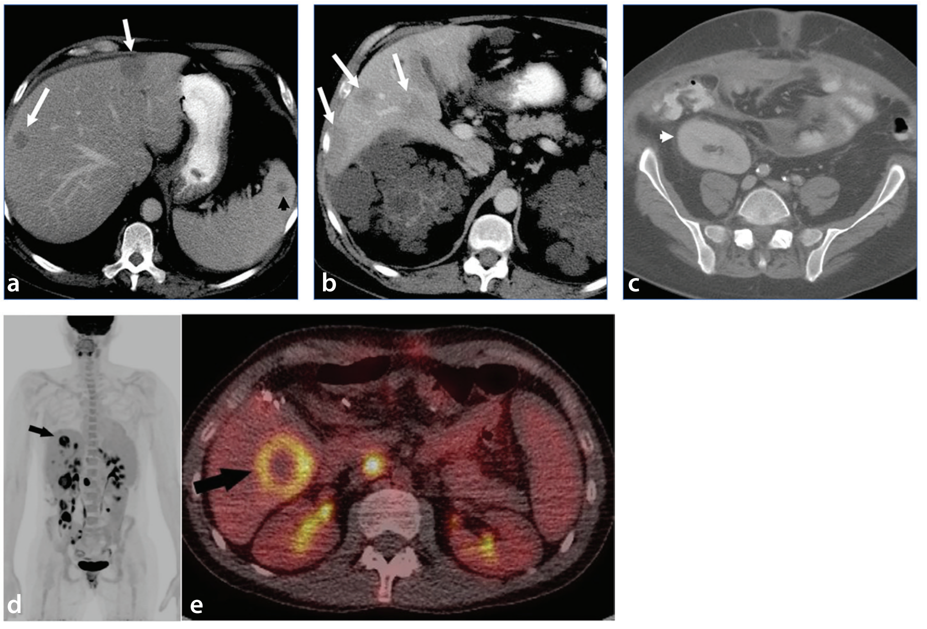


Figure 12. Axial contrast-enhanced computed tomography (CT) images in a 55-year-old man who had undergone renal transplant (short white arrow) 2 years earlier due to autosomal dominant polycystic renal disease and presented with abdominal pain and diarrhea. A contrast-enhanced CT scan **(a-c)** shows multiple hypoattenuating liver masses (long white arrows) and splenomegaly with one hypoattenuating splenic mass (black arrow). An fluorodeoxyglucose-positron emission tomography/CT scan **(d, e)** demonstrates uptake (black arrows) in another case of post-transplant lymphoproliferative disorder.

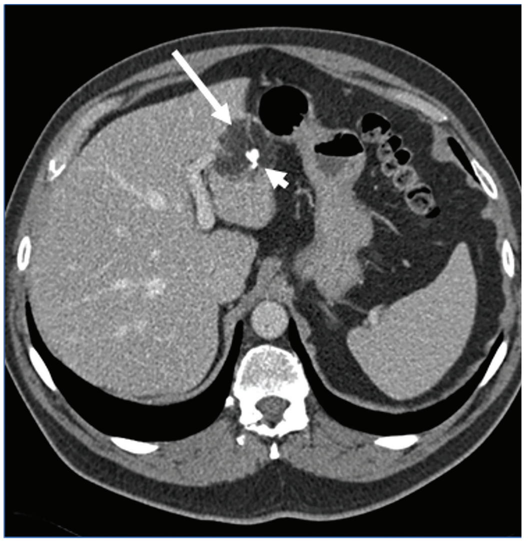


Figure 13. Axial contrast-enhanced computed tomography in the portal venous phase shows a left hepatic lobe mass (long white arrow). The mass was found to represent a teratoma. It is predominantly fatty with peripheral nodular calcifications (short white arrow).

In conclusion, familiarity with the typical appearance of unusual hepatic tumors is important for radiologists. While these tumors are infrequently seen, their inclusion in the differential diagnosis greatly aids the clinician in appropriately triaging patients. This awareness can also avoid unnecessary biopsies, thus improving patient care. This review of several such entities can serve as a useful guide for radiologists in their daily practice.

Acknowledgments

The manuscript was edited by Sarah Bronson, ELS, of the Research Medical Library at The University of Texas MD Anderson Cancer Center.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Tuan Linh L, Minh Duc N, Tu Minh H, et al. Primary hepatic neuroendocrine tumor. *Endocrinol Diabetes Metab Case Rep*. 2021;2021:20-0220. [CrossRef]
2. Yang K, Cheng YS, Yang JJ, Jiang X, Guo JX. Primary hepatic neuroendocrine tumors: multi-modal imaging features with pathological correlations. *Cancer Imaging*. 2017;17(1):20. [CrossRef]
3. Houat AP, von Atzingen AC, Velloni FG, et al. Hepatic neuroendocrine neoplasm: imaging patterns. *Radiol Bras*. 2020;53(3):195-200. [CrossRef]
4. Pan B, Wang SC, Chen ZK, Zou GC. ^{18}F-FDG-PET/CT findings of a primary hepatic neuroendocrine tumor: a case report and literature review. *J Clin Case Stu*. 2019;5(1). [CrossRef]
5. Brazg Ferro L, Wolf I, Peleg Hasson S, et al. Extrapulmonary small cell cancer: a new insight into a rare disease. *Oncology*. 2021;99(6):373-379. [CrossRef]
6. Berniker AV, Abdulrahman AA, Teytelboym OM, Galindo LM, Mackey JE. Extrapulmonary small cell carcinoma: imaging features with radiologic-pathologic correlation. *Radiographics*. 2015;35(1):152-163. [CrossRef]
7. Crider MH, Hoggard E, Manivel JC. Undifferentiated (embryonal) sarcoma of the liver. *Radiographics*. 2009;29(6):1665-1668. [CrossRef]
8. Chung EM, Lattin J, Cube R, et al. From the archives of the AFIP: Pediatric liver masses: radiologic-pathologic correlation. Part 2. Malignant tumors. *Radiographics*. 2011;31(2):483-507. [CrossRef]
9. Prasad SR, Wang H, Rosas H, et al. Fat-containing lesions of the liver: radiologic-pathologic correlation. *Radiographics*. 2005;25(2):321-331. [CrossRef]
10. Lee SJ, Kim SY, Kim KW, et al. Hepatic angiomyolipoma versus hepatocellular carcinoma in the noncirrhotic liver on gadoxetic acid-enhanced MRI: a diagnostic challenge. *AJR Am J Roentgenol*. 2016;207(3):562-570. [CrossRef]
11. Klompenhouwer AJ, Dwarkasing RS, Doukas M, et al. Hepatic angiomyolipoma: an international multicenter analysis on diagnosis, management and outcome. *HPB (Oxford)*. 2020;22(4):622-629. [CrossRef]
12. Gaballah AH, Jensen CT, Palmquist S, et al. Angiosarcoma: clinical and imaging features from head to toe. *Br J Radiol*. 2017;90(1075):20170039. [CrossRef]
13. Yi LL, Zhang JX, Zhou SG, et al. CT and MRI studies of hepatic angiosarcoma. *Clin Radiol*. 2019;74(5):406. [CrossRef]
14. Makhlouf HR, Ishak KG, Goodman ZD. Epithelioid hemangioendothelioma of the liver: a clinicopathologic study of 137 cases. *Cancer*. 1999;85(3):562-582. [CrossRef]
15. Mehrabi A, Kashfi A, Fonouni H, et al. Primary malignant hepatic epithelioid hemangioendothelioma: a comprehensive review of the literature with emphasis on the surgical therapy. *Cancer*. 2006;107(9):2108-2121. [CrossRef]
16. Gan LU, Chang R, Jin H, Yang LI. Typical CT and MRI signs of hepatic epithelioid hemangioendothelioma. *Oncol Lett*. 2016;11(3):1699-1706. [CrossRef]
17. Tan H, Zhou R, Yu H, et al. CT appearances and classification of hepatic epithelioid hemangioendothelioma. *Insights Imaging*. 2023;14(1):56. [CrossRef]
18. Virarkar M, Saleh M, Diab R, Taggart M, Bhargava P, Bhosale P. Hepatic hemangioendothelioma: an update. *World J Gastrointest Oncol*. 2020;12(3):248-266. [CrossRef]
19. Luo L, Cai Z, Zeng S, et al. CT and MRI features of hepatic epithelioid haemangioendothelioma: a multi-institutional retrospective analysis of 15 cases and a literature review. *Insights Imaging*. 2023;14(1):2. [CrossRef]
20. Liu X, Yu H, Zhang Z, et al. MRI appearances of hepatic epithelioid hemangioendothelioma: a retrospective study of 57 patients. *Insights Imaging*. 2022;13(1):65. [CrossRef]
21. Liu Z, Yi L, Chen J, et al. Comparison of the clinical and MRI features of patients with hepatic hemangioma, epithelioid hemangioendothelioma, or angiosarcoma. *BMC Med Imaging*. 2020;20(1):71. [CrossRef]
22. Xu Y, Chen K, Zhang Q, et al. Ultrasound findings of hepatic epithelioid hemangioendothelioma: comparison with other malignant hepatic tumors. *Abdom Radiol (NY)*. 2024;49:762-773. [CrossRef]
23. Banerjee B, Rennison A. Epitheloid haemangioendothelioma of liver: a vascular tumour easily mistaken for metastatic carcinoma on ultrasound imaging. *Br J Radiol*. 1992;65(775):611-613. [CrossRef]
24. Stacchiotti S, Miah AB, Frezza AM, et al. Epithelioid hemangioendothelioma, an ultra-rare cancer: a consensus paper from the community of experts. *ESMO Open*. 2021;6(3):100170. [CrossRef]
25. Haradome H, Woo J, Nakayama H, et al. Characteristics of hepatic schwannoma presenting as an unusual multi-cystic mass on gadoxetic acid disodium-enhanced MR imaging. *Magn Reson Med Sci*. 2018;17(1):95-99. [CrossRef]
26. Ota Y, Aso K, Watanabe K, et al. Hepatic schwannoma: imaging findings on CT, MRI and contrast-enhanced ultrasonography. *World J Gastroenterol*. 2012;18(35):4967-4972. [CrossRef]
27. Marin Campos C, Garcia Sanz I, Muñoz de Nova JL, Valdés de Anca A, Martín Pérez ME. Schwannoma of the biliary tract resembling cholangiocarcinoma: a case report and review. *Ann R Coll Surg Engl*. 2016;98(7):143-146. [CrossRef]
28. Wan DL, Zhai ZL, Ren KW, et al. Hepatic schwannoma: a case report and an updated 40-year review of the literature yielding 30 cases. *Mol Clin Oncol*. 2016;4(6):959-964. [CrossRef]
29. Bladé J, de Larrea CF, Rosiñol L. Extramedullary involvement in multiple myeloma. *Haematologica*. 2012;97(11):1618-1619. [CrossRef]
30. Tomasian A, Sandrasegaran K, Elsayes KM, Shanbhogue A, Shaaban A, Menias CO. Hematologic malignancies of the liver: spectrum of disease. *Radiographics*. 2015;35(1):71-86. [CrossRef]
31. Ng P, Slater S, Radvan G, Price A. Hepatic plasmacytomas: case report and review of imaging features. *Australas Radiol*. 1999;43(1):98-101. [CrossRef]
32. Noronha V, Shafi NQ, Obando JA, Kummar S. Primary non-Hodgkin's lymphoma of the liver. *Crit Rev Oncol Hematol*. 2005;53(3):199-207. [CrossRef]
33. Ippolito D, Porta M, Maino C, et al. Diagnostic approach in hepatic lymphoma: radiological imaging findings and literature review. *J Cancer Res Clin Oncol*. 2020;146(6):1545-1558. [CrossRef]
34. Colagrande S, Calistri L, Grazzini G, et al. MRI features of primary hepatic lymphoma. *Abdom Radiol (NY)*. 2018;43(9):2277-2287. [CrossRef]
35. Camacho JC, Moreno CC, Harri PA, Aguirre DA, Torres WE, Mittal PK. Posttransplantation lymphoproliferative disease: proposed imaging classification. *Radiographics*. 2014;34(7):2025-2038. [CrossRef]
36. Soliman M, Guys N, Liu P, et al. Multimodality imaging findings of infection-induced tumors.

Abdom Radiol (NY). 2022;47(11):3930-3953. [CrossRef]

37. Taylor AL, Marcus R, Bradley JA. Post-transplant lymphoproliferative disorders (PTLD) after solid organ transplantation. *Crit Rev Oncol Hematol*. 2005;56:155-167. [CrossRef]

38. Borhani AA, Hosseinzadeh K, Almusa O, Furlan A, Nalesnik M. Imaging of posttransplantation lymphoproliferative disorder after solid organ transplantation. *Radiographics*. 2009;29(4):1000-1002. [CrossRef]

39. Ramkumar J, Best A, Gurung A, et al. Resection of ruptured hepatic teratoma in an adult. *Int J Surg Case Rep*. 2018;53:414-419. [CrossRef]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242912

ABDOMINAL IMAGING
ORIGINAL ARTICLE

Grading portal vein stenosis following partial hepatectomy by high-frequency ultrasonography: an *in vivo* study of rats

Lin Ma[1,2]
Chihan Peng[1]
Lulu Yang[1]
Xiaoxia Zhu[1]
Hongxia Fan[1]
Jiali Yang[1]
Hong Wang[1]
Yan Luo[1]

[1]West China Hospital of Sichuan University, Department of Ultrasound, Sichuan, China

[2]West China Tian Fu Hospital, Sichuan University, Department of Ultrasound, Sichuan, China

PURPOSE
To evaluate the diagnostic value of ultrasound in grading portal vein stenosis (PVS) in a rat model of 70% partial hepatectomy (PH).

METHODS
A total of 96 Sprague-Dawley rats were randomly divided into a PH group and PVS groups with mild, moderate, and severe PVS following PH. Hemodynamic parameters were measured using high-frequency ultrasound (5–12 MHz high-frequency linear transducer), including pre-stenotic, stenotic, and post-stenotic portal vein diameters (PVD_{pre}, PVD_s, PVD_{post}); pre-stenotic and stenotic portal vein velocity (PVV_{pre}, PVV_s); hepatic artery peak systolic velocity (PSV); end-diastolic velocity; and resistive index. The portal vein diameter ratio (PVDR) and portal vein velocity ratio (PVVR) were calculated using the following formulas: $PVDR=PVD_{pre}/PVD_s$ and $PVVR=PVV_s/PVV_{pre}$. The value of these parameters in grading PVS was assessed.

RESULTS
Portal vein hemodynamics showed gradient changes as PVS aggravated. For identifying >50% PVS, PVD_s and PVDR were the best parameters, with areas under the curve (AUC) of 0.85 and 0.86, respectively. For identifying >65% PVS, PVD_s, PVDR, and PVVR were relatively better, with AUCs of 0.94, 0.85, and 0.88, respectively. The AUC of hepatic artery PSV for identifying >65% PVS was 0.733.

CONCLUSION
High-frequency ultrasonography can be used to grade PVS in rats, with PVD_s, PVDR, and PVVR being particularly useful. Hepatic artery PSV may help in predicting >65% PVS. These findings provide valuable information for PVS rat model research and offer an experimental basis for further studies on PVS evaluation in living-donor liver transplantation (LDLT).

CLINICAL SIGNIFICANCE
Ultrasonography serves as a first-line technology for diagnosing PVS following LDLT. However, the grading criteria for PVS severity remain unclear. Investigating the use of ultrasonic hemodynamics in the early diagnosis of PVS and grading stenosis severity is important for early postoperative intervention and improving recipient survival rates.

KEYWORDS
Portal vein stenosis, high-frequency ultrasonography, hemodynamics, portal vein, hepatic artery, rat

Corresponding author: Yan Luo

E-mail: yanluo@scu.edu.cn

Received 03 July 2024; revision requested 14 August 2024; accepted 29 September 2024.

Epub: 25.11.2024

Publication date: 03.03.2025

DOI: 10.4274/dir.2024.242912

As organ transplantation techniques mature and new immunosuppressants are developed, living-donor liver transplantation (LDLT) is becoming an effective treatment for end-stage liver disease. Compared with whole-liver transplantation, a unique characteristic of LDLT is that postoperative regeneration allows the liver volume to increase, resulting in successful reconstruction even though the graft volume is relatively small.[1,2] Sufficient portal blood flow is a prerequisite for the transplanted liver to regenerate and survive. In LDLT, the recipient's portal vein trunk is usually anastomosed to the portal vein branch of the graft

You may cite this article as: Ma L, Peng C, Yang L, et al. Grading portal vein stenosis following partial hepatectomy by high-frequency ultrasonography: an *in vivo* study of rats. *Diagn Interv Radiol.* 2025;31(2):68-74.

(left or right branch). Consequently, the donor's and recipient's portal vein diameters (PVD) often do not match, resulting in portal vein stenosis (PVS). Furthermore, portal vein angulation or torsion may lead to PVS after LDLT more frequently than after whole-liver transplantation, with an incidence rate of 0.5%–8.1%.[3-5] Mild PVS usually does not affect liver regeneration or function, but severe PVS can lead to portal hypertension, small-for-size syndrome, and acute liver failure. If PVS can be discovered early and clinical intervention is performed before liver regeneration and function are irreversibly affected, this defect may be reversed.[4-9]

The diagnosis of PVS mainly relies on imaging techniques such as ultrasonography, computed tomography, magnetic resonance imaging, and digital subtraction angiography. Among these techniques, ultrasonography can accurately assess PVD and hemodynamics and has advantages such as convenience, lack of radiation, repeatability, and bedside operation. Therefore, ultrasonography serves as a first-line imaging modality for diagnosing PVS in the early postoperative period and during long-term follow-up. Generally, a diagnosis of significant stenosis is made when the portal vein trunk diameter is <2.5–3.5 mm, the blood flow velocity at the stenotic site is >150 cm/s, or the velocity ratio between stenotic and pre-stenotic flow is ≥4.[10-13] However, to date, the grading criteria in ultrasonography for PVS severity remain unclear. In addition, when portal blood flow volume decreases, hepatic artery flow volume will show varying degrees of increase due to the hepatic arterial buffer response (HABR).[14,15] There are no reports on how hepatic artery flow changes under different severities of PVS or whether its hemodynamic parameters can aid in PVS evaluation. Therefore, studying the application of ultrasonic hemodynamics for the early diagnosis of PVS and the grading of stenosis severity is important for early postoperative intervention and for increasing the survival rate of liver transplant recipients.

Main points

- Portal vein hemodynamic parameters-portal vein diameter at stenosis (PVD_s), portal vein diameter ratio (PVDR), portal vein velocity at stenosis (PVV_s), and portal vein velocity ratio (PVVR)-show significant gradient changes among different degrees of portal vein stenosis (PVS), with stenosis rate (SR) ≤50%, 50%< SR ≤65%, and SR >65% (all P < 0.0001).
- PVD at stenosis and PVDR are the best parameters for PVS grading [all areas under the curve (AUCs) >0.80].
- PVV_s can effectively diagnose the presence/absence of PVS (AUC: 0.958), but the diagnostic performance in PVS grading is relatively low (AUC <0.80). The PVVR showed good performance in the identification of >65% PVS (AUC: 0.880).
- A significant increase in hepatic artery peak systolic velocity may be helpful for PVS evaluation, especially in predicting >65% PVS (AUC: 0.733).

Due to ethical constraints and the diversity of liver diseases, we conducted animal experiments in this study. The rat model of 70% partial hepatectomy (PH) is a classical model for studying liver regeneration,[16,17] and partial portal vein ligation is the most commonly used method for producing the PVS model.[18,19] In this study, varying degrees of partial portal vein ligation were performed based on the 70% PH rat model to simulate different degrees of PVS following LDLT. Ultrasonography was used to measure the hemodynamic parameters of the portal vein and hepatic artery to assess the effectiveness of ultrasonography in diagnosing and grading PVS, thereby providing an experimental basis for further studies on early PVS evaluation and intervention.

Methods

Study subjects

All rats and procedures used in this research were approved by the Animal Ethics Committee of West China Hospital, Sichuan University (no: 2020101A). Ninety-six healthy male Sprague-Dawley rats (7–14 weeks old, weighing 200–400 g, specific-pathogen-free grade) were purchased from Chengdu Dashuo Biotechnology Co., Ltd, and given ad libitum access to food and water at the animal experiment center of West China Hospital. All rats were housed at a constant temperature under a 12-h light–dark cycle to acclimate for at least 1 week before the experiment.

The rats were randomly divided into a PH group and PVS groups with mild, moderate, and severe PVS following PH (n = 24 for each group). The PH group was established as a model of 70% hepatectomy without portal vein ligation, whereas the PVS groups were created through varying degrees of partial portal vein ligation after PH. Mild, moderate, and severe PVS were respectively defined as ≤50% stenosis, 50%–65% stenosis, and >65% stenosis, approaching near occlusion.[19]

Construction of rat models

Construction of the 70% partial hepatectomy rat model

The standard method for 70% PH in rats developed by Higgins and Anderson[16] was used as a reference.[17] The specific procedure was as follows: (1) Continuous inhalational anesthesia with ether was administered before the rat was placed in the supine position. The rats were immobilized, and the abdomen was shaved using an electric hair remover. (2) Iodine was used to disinfect the surgical site, and an abdominal midline incision was made below the xiphoid process. The skin and muscles were dissected layer by layer to access the abdominal cavity, and the liver was exposed. (3) A suture was used to ligate and then resect the left lateral lobe and middle lobe. The resected liver accounted for approximately 70% of the entire liver. (4) The liver pedicle ligation site was inspected for bleeding, and the residual liver lobes were examined for congestion.

Construction of portal vein stenosis models with varying severity after partial hepatectomy

After PH, varying degrees of partial ligation of the portal vein trunk were performed to construct PVS models of different severity.[18,19] The specific steps were as follows: (1) The portal vein trunk was dissociated, and a microvascular caliper was used to measure the PVD. (2) Needles of different sizes were selected and placed parallel to the portal vein. A silk suture was used to ligate the portal vein and the needle together. At this point, significant congestion could be observed in the gastrointestinal tract. After ligation, the needle was slowly withdrawn, alleviating the congestion in the gastrointestinal tract. The PVD of the stenotic segment was equal to the external diameter of the needle. Needles of varying sizes were used for partial ligation of the portal vein to create PVS models of different severity. The sizes of needles used in this study were 18G, 19G, 20G, 21G, and 22G, with outer diameters of 1.2 mm, 1.0 mm, 0.9 mm, 0.8 mm, and 0.7 mm, respectively. The PV stenosis rate (SR) was calculated using the formula $SR=(1-D_{needle}/PVD) \times 100\%$.[3] After evaluating the intestinal congestion status and vital signs, 32,000 units of penicillin and 5 mL of NaCl (0.9%) were administered via peritoneal injection, and then the abdomen was sealed layer by layer.[4] The rats were labeled and housed in individual cages after surgery, kept warm, and given ad libitum access to food and water.

Ultrasonography examination

Duplex Doppler ultrasound examinations were performed using an IU22 US system (Philips Healthcare, Bothell, WA), equipped with a 5-12 MHz transducer. At 24 h post-surgery, scans were conducted with the rats ether-anesthetized and stably positioned in the supine position, using both grayscale and color Doppler imaging to identify vascular landmarks. Doppler tracings were acquired, and the best tracing was selected for analysis. In the PH group, the PVD and maximum portal vein velocity (PVV) were measured at a site approximately 5 mm below the bifurcation of the hilum. In the PVS groups, pre-stenotic, stenotic, and post-stenotic PVD (PVD_{pre}, PVD_s, PVD_{post}) and pre-stenotic and stenotic PVV (PVV_{pre}, PVV_s) were measured (Figure 1). The PVD ratio (PVDR) and the PVV ratio (PVVR) were calculated using the following formula: $PVDR=PVD_{pre}/PVD_s$ and $PVVR=PVV_s/PVV_{pre}$. Hepatic artery peak systolic velocity (PSV) and end-diastolic velocity (EDV) were measured in all rats, and the resistive index (RI) was calculated using the following formula: RI=(PSV−EDV)/PSV. The sampling volume was adjusted based on the course of the blood vessel and its inner diameter. The gain was adjusted to maximum sensitivity without noise, and the angle between the sound beam and blood flow was ≤60°. The aforementioned scanning and image storage were performed by an experienced physician who was blinded to the grouping. The mean of three measurements was calculated for all results.

Research ethics standards compliance

This study was carried out in accordance with the principles of the Basel Declaration and was approved by the Animal Ethics Committee of West China Hospital (decision no: 2020101A, date: March 24th, 2020).

Statistical analysis

SPSS 25.0 and GraphPad Prism 8 were used for statistical analysis. A value of $P < 0.05$ indicated a statistically significant difference. One-Way analysis of variance was used to compare the hemodynamic parameters among different groups for source data with a normal distribution. When inter-group differences were present, the least significant difference test was used for pairwise comparisons when variances were homogeneous, and Dunnett's T3 test was used for pairwise comparisons when variances were heterogeneous. Values are expressed as mean ± standard deviation ($\bar{x} \pm s$). Non-parametric rank tests were used to compare non-normally distributed source data, and pairwise comparisons were performed when inter-group differences were present. These values are expressed as medians.

The receiver operating characteristic (ROC) curve was plotted, and the area under the curve (AUC), standard error, asymptotic significance (b), asymptotic 95% confidence interval, best cut-off, sensitivity, and specificity were calculated to evaluate the value of the various ultrasound parameters in the diagnosis of PVS and in predicting stenosis severity.

Results

Model construction

In this study, the 70% PH models with no PVS were successfully constructed in 24 rats, whereas PVS models of different severities following PH were constructed in 72 rats. The SRs of the mild, moderate, and severe PVS groups were (45.16 ± 3.40)%, (59.21 ± 3.84)%, and (69.46 ± 2.17)%, respectively.

$D_{stenosis}$ (i.e., outer needle diameter) in PVS models with different severities showed significant gradient changes. When SR was >65%, the portal vein trunk diameter was extremely narrow, and the needle used for model construction was significantly thinner: mainly 21G (outer diameter: 0.8 mm). An 18G (outer diameter: 1.2 mm) needle was mostly used for model construction in rats with SR ≤50%, and an 18G (outer diameter: 1.2 mm) or 20G (outer diameter: 0.9 mm) needle was mostly used for model construction in rats with 50%< SR ≤65%. $D_{stenosis}$ in rats with SR >65% was significantly lower than that in the SR ≤50% and 50%< SR ≤65% groups, and the $D_{stenosis}$ of the 50%< SR ≤65% group was also significantly lower than that of the SR ≤50% group (Figure 2).

Hemodynamic changes

Portal vein hemodynamic changes

Residual liver and portal vein after 70% PH in rats can be observed using conventional ultrasound. In PVS rats, grayscale ultrasound clearly showed PVS, whereas Doppler ultrasound revealed turbulence of blood flow at the stenosis site, with the stenotic flow significantly faster and the pre-stenotic flow slower (Figure 1b, c). When PVS occurred, the lumen diameter of the stenotic site was significantly smaller, whereas the lumen diameters at two ends of the stenotic site showed varying degrees of expansion. As shown in Table 1 and Figure 3a-d, the PVD_{pre} of the moderate and severe PVS groups was significantly higher than that of the PH and mild PVS groups, and the PVD_{post} of the moderate PVS group was significantly higher than that of the PH group (all $P < 0.05$). The PVD_s among the mild, moderate, and severe PVS groups were significantly lower than that of

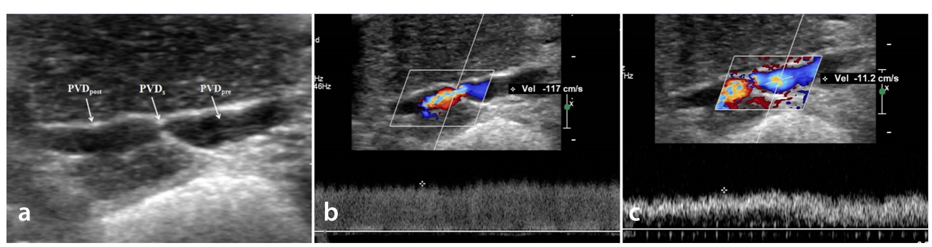


Figure 1. Measurement of portal vein blood flow parameters in the PVS group **(a)**. Pre-stenotic, stenotic, and post-stenotic PVD (PVD_{pre}, PVD_s, PVD_{post}) were measured using ultrasound (arrows). **(b, c)**. Stenotic and pre-stenotic PVV (PVV_s, PVV_{pre}) were measured using ultrasound. PVS, portal vein stenosis; PVD, portal vein diameter; PVV, portal vein velocity.

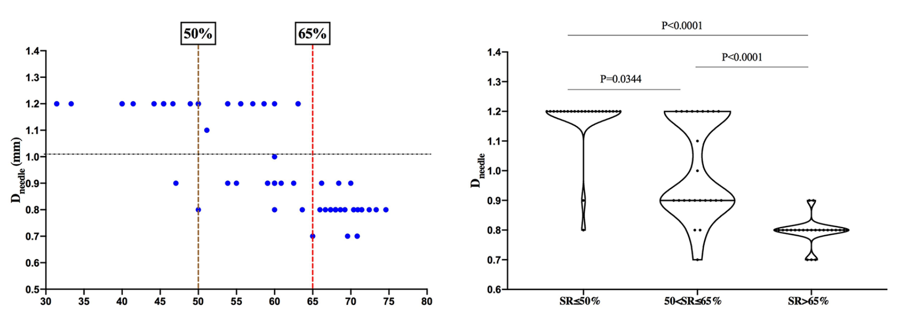


Figure 2. D_{needle} used in the PVS groups with different severities. D_{needle}, the diameter of the needle; PVS, portal vein stenosis.

the PH group (all $P < 0.05$). As PVS severity increased, PVD_s gradually decreased, and PVDR conversely increased. The differences in PVD_s and PVDR among the mild, moderate, and severe PVS groups were statistically significant (all $P < 0.05$).

When PVS occurred, the flow velocity at the stenotic site significantly increased, and pre-stenotic flow velocity showed varying magnitudes of decrease. As shown in Table 1 and Figure 3e-g, the PVV_{pre} of the moderate PVS group was significantly lower than the PVV of the PH group, and the PVV_{pre} of the severe PVS group was significantly lower than that of the PH and mild PVS groups (all $P < 0.05$). The PVV_s among the mild, moderate, and severe PVS groups were significantly higher than the PVV of the PH group (all $P < 0.05$). As PVS severity increased, PVV_s and PVVR increased. The differences in PVV_s and PVVR among the mild, moderate, and severe PVS groups were statistically significant (all $P < 0.05$).

Hepatic artery hemodynamic changes

When PVS occurred, hepatic artery PSV showed varying degrees of increase, and the PSV of the severe PVS group was significantly higher than that of the 70% PH group ($P < 0.05$, Table 1 and Figure 3h). There were no significant differences in EDV or RI among the various groups (all $P > 0.05$).

Evaluation of ultrasonography in portal vein stenosis diagnosis and grading of stenosis severity

Surgical PVS severity is the gold standard for PVS diagnosis. When diagnosing PVS, an ROC curve was plotted with the PH group as negative and the mild, moderate, and severe PVS groups as positive results. For identifying >50% PVS, the mild PVS group was used as the negative samples, and the moderate and severe PVS groups were used as positive samples for plotting the ROC curve. For identifying >65% PVS, the mild and moderate PVS groups were used as the negative samples, and the severe PVS group was used as the positive samples to plot the ROC curve.

The AUCs of PVD_{pre}, PVD_s, PVD_{post}, PVV_{pre}, and PVV_s in PVS diagnosis were significantly larger than the diagnostic reference AUC ($P < 0.05$ vs. AUC: 0.05, Table 2 and Figure 4a, b). The AUCs of PVD_s and PVV_s were 0.998 and 0.958, respectively. When PVD_s was <1.37 mm or PVV_s was >25.85 cm/s, their sensitivity and specificity were 98.61% and 100% or 83.33% and 100%, respectively.

Table 1. US parameters in different groups

US parameter	PH	Mild PVS	Moderate PVS	Severe PVS
PVD_s	1.94 ± 0.38	1.10 ± 0.17^{a}	0.97 ± 0.16^{a,b}	0.76 ± 0.06^{a,b,c}
PVD_{pre}	1.94 ± 0.38	2.10 ± 0.40	2.58 ± 0.63^{a,b}	2.60 ± 0.35^{a,b}
PVD_{post}	1.94 ± 0.38	2.13 ± 0.37	2.30 ± 0.41^{a}	2.13 ± 0.41
PVDR	-	1.97 ± 0.60	2.73 ± 0.85^{b}	3.43 ± 0.51^{b,c}
PVV_s	14.80 ± 4.70	43.15 ± 18.64^{a}	51.86 ± 30.73^{a}	80.50 ± 38.49^{a,b,c}
PVV_{pre}	14.80 ± 4.70	11.56 ± 4.18	9.72 ± 3.48^{a}	8.43 ± 3.67^{a,b}
PVVR	-	4.03 ± 1.91	5.33 ± 3.09	10.33 ± 4.90^{b,c}
HA PSV	42.65 ± 16.37	53.61 ± 16.55	56.86 ± 25.44	59.69 ± 17.37^{a}

[a]$P < 0.05$ vs. PH group; [b]$P < 0.05$ vs. mild PVS group; [c]$P < 0.05$ vs. moderate group. US, ultrasound; PVS, portal vein stenosis; PH, partial hepatectomy; HA PSV, hepatic artery peak systolic velocity.

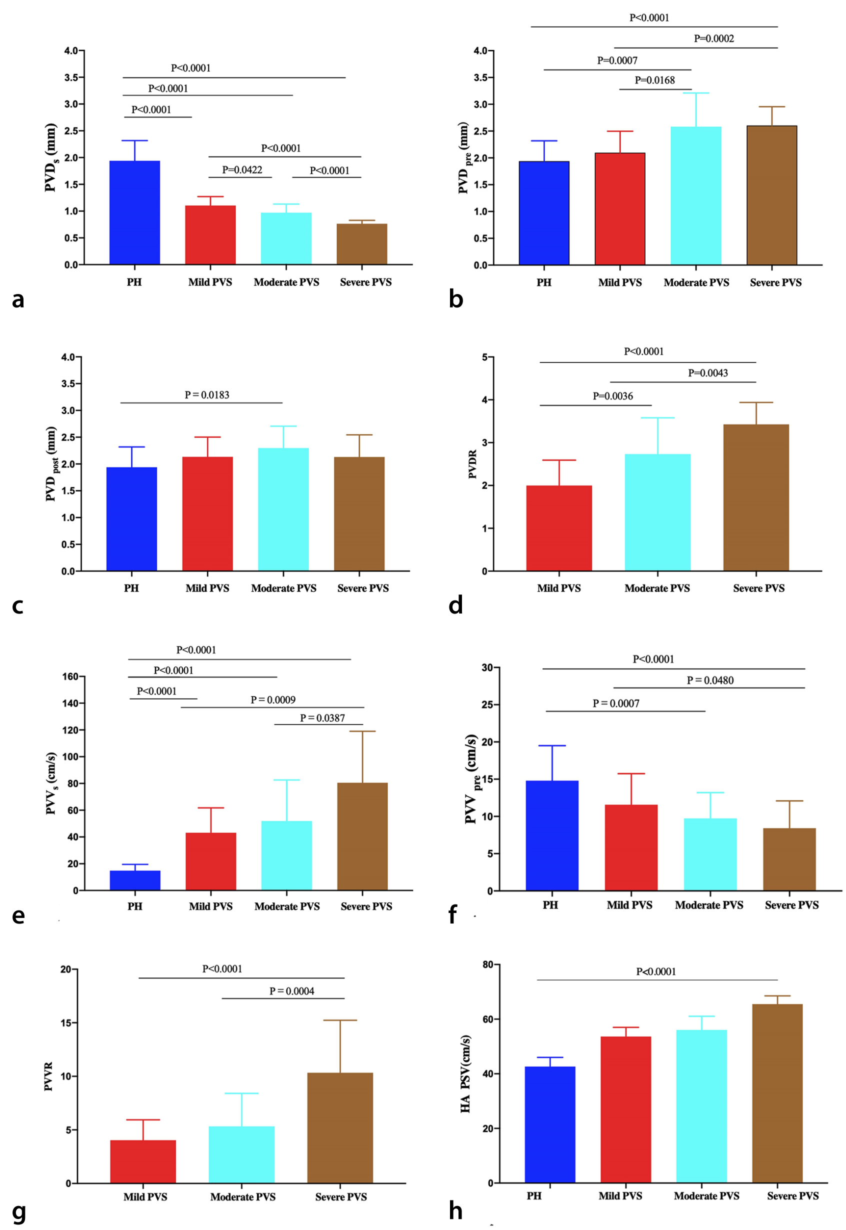


Figure 3. Ultrasound parameters in different groups. **(a)** Portal vein diameter at the stenotic site (PVD_s). **(b)** Portal vein diameter at the pre-stenotic site (PVD_{pre}). **(c)** Portal vein diameter at the post-stenotic site (PVD_{post}). **(d)** Portal vein diameter ratio (PVDR, PVD_{pre}/PVD_s). **(e)** Portal vein velocity at the stenotic site (PVV_s). **(f)** Portal vein velocity at the pre-stenotic site (PVV_{pre}). **(g)** Portal vein velocity ratio (PVVR, PVV_s/PVV_{pre}). **(h)** Hepatic artery peak systolic velocity (PSV).

With regards to PVS stenosis severity grading, the AUCs of PVD_s, PVD_{pre}, PVDR, PVV_s, $PVV_{pre,}$ and PVVR were significantly higher than the diagnostic reference AUC when used to identify >50% PVS and >65% PVS. For identifying >50% PVS, PVD_s and PVDR were better than other parameters, with AUCs of 0.85 and 0.86, respectively. When PVD_s was <0.95 mm or PVDR >2.51, their sensitivity and specificity were 75.00% and 83.33% or 77.08% and 87.50%, respectively (Table 2 and Figure 4c, d). For identifying >65% PVS, PVD_s, PVDR, and PVVR were relatively better than other parameters, with AUCs of 0.94, 0.85, and 0.88, respectively. When PVD_s was <0.87 mm, PVDR was >2.82, or PVVR was >5.43, their sensitivity and specificity were 100% and 81.25%, 91.67% and 77.08%, or 87.50% and 72.92%, respectively (Table 2 and Figure 4e, f).

As shown in Table 2 and Figure 5, the AUC of hepatic artery PSV in predicting PVS was 0.711, and when used to identify >50% PVS and >65% PVS, the AUC of hepatic artery PSV was 0.666 and 0.733, respectively (all *P* < 0.05 vs. AUC: 0.05). When PSV was >51.15 cm/s, the sensitivity of identifying >65% PVS reached 87.50%, whereas the specificity was only 55.56%.

Discussion

Liver regeneration after LDLT is key to postoperative patient survival. Portal vein blood flow accounts for 75%–80% of the total blood flow volume in the liver. On one hand, this provides nutrient-rich blood from the intestines to liver tissues. On the other hand, this blood acts as a carrier of hepatocyte growth factors, hormones, and related receptors that play a vital role in liver regeneration. Therefore, sufficient portal vein blood supply is one of the prerequisites for the survival of the graft. PVS is one of the major vascular complications after LDLT. It may occur within 1 month after liver transplantation or may be late-onset (≥3 months after surgery).[20,21] Although PVS is not as acute as hepatic artery complications, its early clinical manifestations are not specific, and severe PVS significantly reduces liver blood supply and severely impairs the function of the transplanted liver, leading to graft failure.[3-6] In addition, the incidence of PVS is relatively high in pediatric LDLT due to factors such as small recipient portal vein size, dysplasia, and mismatched donor-recipient PVD.[4,5] In clinical practice, symptomatic treatment (such as balloon dilatation or stent implantation) is usually performed when there is significant hepatic dysfunction or portal hypertension.[7-9] However, hepatocyte structure and function may have undergone irreversible damage at this point, resulting in grafts being in a state of poor regeneration for a long time, even after treatments are applied. Therefore, early diagnosis of PVS and accurate grading of stenosis severity promote early intervention and thus increase the survival rate of patients.

Ultrasonography is the preferred imaging method for the early diagnosis of vascular complications after liver transplantation. Conventional grayscale ultrasound can clearly show the liver parenchyma and portal vein and accurately measure the PVD. Doppler ultrasound can monitor portal vein blood flow for disturbances, observe the blood flow direction, and obtain blood flow velocity infor-

Table 2. Results of ROC analysis in grading PVS by ultrasound

	US index	AUC	Standard error	*P* value	95% confidence interval: Lower bound	95% confidence interval: Upper bound	Best cut-off	Sensitivity (%)	Specificity (%)
PVS	PVD_s (mm)	0.998	0.002	<0.0001	0.994	1.000	<1.37	98.61	100.00
	PVD_{pre} (mm)	0.776	0.051	<0.0001	0.675	0.877	>2.13	68.06	79.17
	PVD_{post} (mm)	0.694	0.062	0.0045	0.574	0.815	>2.11	62.50	75.00
	PVV_s (cm/s)	0.958	0.018	<0.0001	0.922	0.994	>25.85	83.33	100.00
	PVV_{pre} (cm/s)	0.793	0.048	<0.0001	0.699	0.886	<11.60	69.44	79.17
	HA PSV (cm/s)	0.711	0.058	0.0020	0.5969	0.8250	>46.90	73.61	62.50
>50% PVS	PVD_s (mm)	0.850	0.050	<0.0001	0.760	0.940	<0.95	75.00	83.33
	PVD_{pre} (mm)	0.790	0.060	<0.0001	0.680	0.900	>2.54	62.50	87.50
	PVD_{post} (mm)	0.520	0.070	0.7335	0.380	0.670	-	-	-
	PVDR	0.860	0.050	<0.0001	0.760	0.950	>2.51	77.08	87.50
	PVV_s (cm/s)	0.690	0.060	0.0094	0.570	0.810	>52.60	62.50	75.00
	PVV_{pre} (cm/s)	0.670	0.070	0.0169	0.540	0.810	<8.00	45.83	83.33
	PVVR	0.770	0.050	0.0002	0.660	0.880	>4.79	72.92	75.00
	HA PSV (cm/s)	0.666	0.047	0.0010	0.574	0.759	>48.45	72.73	56.06
>65% PVS	PVD_s (mm)	0.940	0.030	<0.0001	0.880	1.000	<0.87	100.00	81.25
	PVD_{pre} (mm)	0.650	0.060	0.0417	0.520	0.770	>2.14	95.83	47.92
	PVD_{post} (mm)	0.580	0.070	0.2900	0.440	0.710	-	-	-
	PVDR	0.850	0.040	<0.0001	0.760	0.940	>2.82	91.67	77.08
	PVV_s (cm/s)	0.767	0.060	0.0002	0.650	0.880	>52.60	79.17	64.58
	PVV_{pre} (cm/s)	0.670	0.070	0.0169	0.540	0.800	<8.85	70.83	62.50
	PVVR	0.880	0.040	<0.0001	0.810	0.960	>5.43	87.50	72.92
	HA PSV (cm/s)	0.733	0.051	0.0007	0.633	0.832	>51.15	87.50	55.56

PVS, portal vein stenosis; US, ultrasound; PVD_s, portal vein diameter at the stenotic site; PVD_{pre}, portal vein diameter at the pre-stenotic site; PVD_{post}, portal vein diameter at the post-stenotic site; PVV_s, portal vein velocity at the stenotic site; $PVV_{pre,}$ portal vein velocity at the pre-stenotic site; PVDR, portal vein diameter ratio (PVD_{pre}/PVD_s); PVVR, portal vein velocity ratio (PVV_s / PVV_{pre}); HA PSV, hepatic artery peak systolic velocity.

mation. Mild stenosis (SR <50%) at the portal vein anastomosis usually does not lead to significant hemodynamic changes. When significant PVS occurs, grayscale ultrasound will show local lumen narrowing, whereas Doppler ultrasound will demonstrate disturbance of blood flow at the stenotic site with a faster blood flow velocity. Currently, there are no unified ultrasonic diagnostic criteria for PVS in clinical practice. In China, a PVD of <2.5–3.5 mm at the stenotic site, a blood flow velocity at the stenotic site >150 cm/s, or a velocity ratio between stenotic and pre-stenotic flow ≥4 is regarded as the diagnostic criterion for PVS.[6,10-12] Mullan et al.[13] defined a maximal blood velocity >80 cm/s at the stenotic segment of the portal vein as the diagnostic criterion for PVS, with a sensitivity of 100% and a specificity of 84%. Chong et al.[22] used a maximal blood velocity >125 cm/s at the stenotic segment of the portal vein as the PVS diagnostic criterion, which had a specificity of 95% and a sensitivity of 73%. Moreover, the grading criteria in ultrasonography for PVS severity are not clear.

In this study, partial portal vein ligation was carried out based on the 70% PH rat model to simulate different degrees of PVS after LDLT. This model is easy to construct, stable, and facilitates hemodynamic monitoring. When PVS occurred, PVD decreased at the stenotic site, and PVD at the pre-stenotic and post-stenotic sites showed varying degrees of increase. Furthermore, stenotic PVV significantly increased, whereas pre-stenotic PVV showed varying degrees of decrease. The PVD$_s$, PVDR, PVV$_s$, and PVVR of the mild, moderate, and severe PVS groups showed significant gradient changes. More severe stenosis led to lower PVD$_s$, higher PVV$_s$, and larger PVDR and PVVR. Among the various portal vein hemodynamic parameters, PVD$_s$ and PVV$_s$ showed good performance in diagnosing PVS, followed by PVD$_{pre}$ and PVV$_{pre}$, whereas PVD$_{post}$ showed relatively poor performance. In grading PVS severity, PVD$_s$, PVD$_{pre}$, PVDR, PVV$_s$, PVV$_{pre}$, and PVVR demonstrated some diagnostic efficacy. Regarded as the standard with a high diagnostic value, an AUC >0.80 indicates that PVD$_s$ and PVDR can effectively differentiate mild, moderate, and severe PVS, whereas PVVR showed good diagnostic performance in identifying >65% PVS. In contrast, PVD$_{pre}$, PVV$_s$, and PVV$_{pre}$ showed relatively poor performance in grading PVS severity. After PH, the residual liver will be in a hyperdynamic circulatory state, and portal vein blood flow volume and velocity will increase. In this study, the construction of the surgical model and the

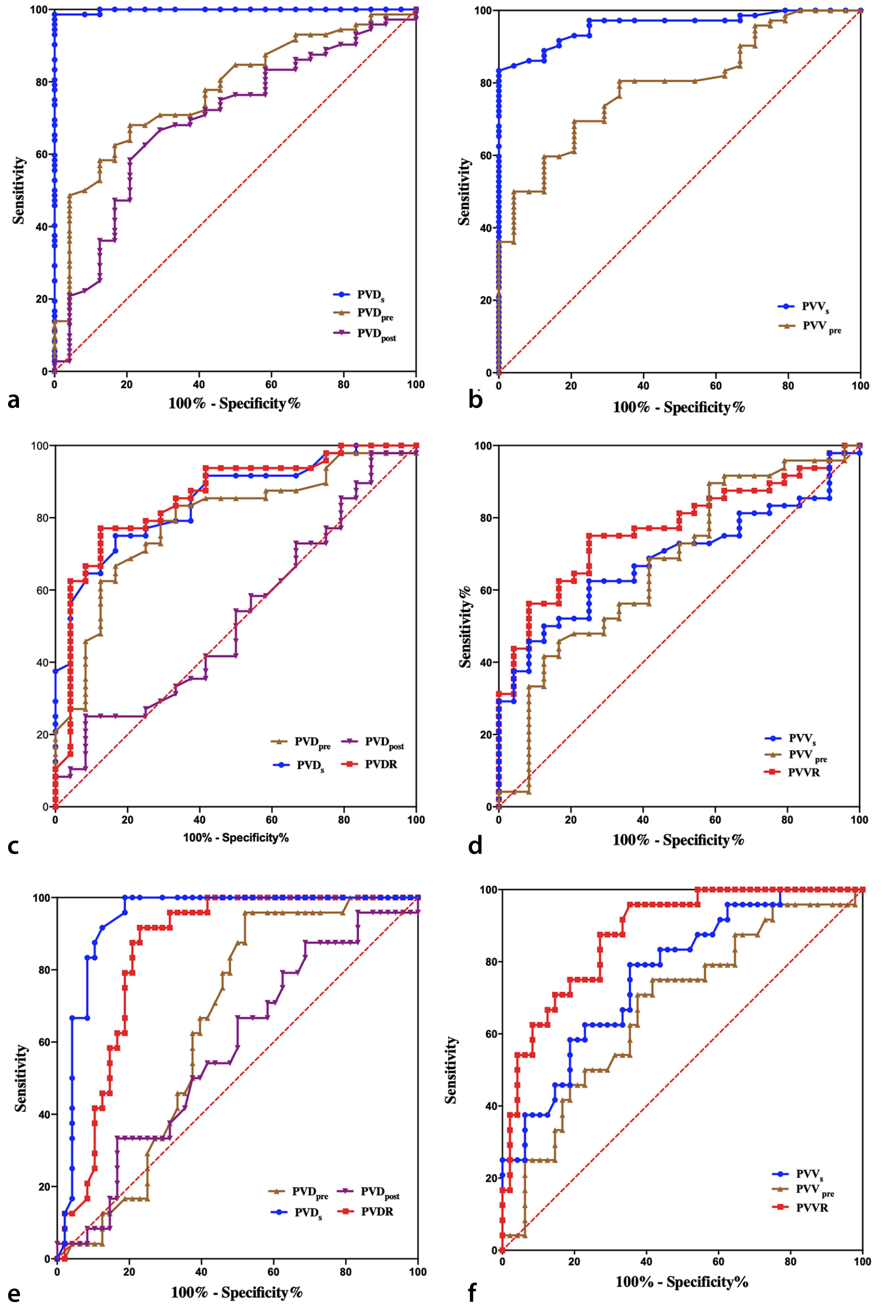


Figure 4. Receiver operating characteristic (ROC) curves of portal vein parameters in grading PVS. **(a, b)** ROC curves of PVD$_s$, PVD$_{pre}$, PVD$_{post}$, PVV$_s$, and PVV$_{pre}$ in diagnosing PVS. **(c, d)** ROC curves of PVD$_s$, PVD$_{pre}$, PVD$_{post}$, PVDR, PVV$_s$, PVV$_{pre}$, and PVVR in identifying >50% PVS. **(e, f)** ROC curves of PVD$_s$, PVD$_{pre}$, PVD$_{post}$, PVDR, PVV$_s$, PVV$_{pre}$, and PVVR in identifying >65% PVS. PVS, portal vein stenosis; PVD$_s$, portal vein diameter at the stenotic site; PVD$_{pre}$, portal vein diameter at the pre-stenotic site; PVD$_{post}$, portal vein diameter at the post-stenotic site; PVV$_s$, portal vein velocity at the stenotic site; PVV$_{pre}$, portal vein velocity at the pre-stenotic site; PVDR, portal vein diameter ratio (PVD$_{pre}$/PVD$_s$); PVVR, portal vein velocity ratio (PVV$_s$/PVV$_{pre}$).

grading and diagnostic criteria for PVS were all based on portal vein blood flow after PH. Additionally, there were inter-individual differences in parameters. Hence, the sample size should be expanded to further validate the PVS grading criteria.

When significant changes in portal vein blood flow volume occur, the hepatic artery buffers these effects by adjusting the blood flow volume to maintain relative stability in the total blood flow volume of the liver. This phenomenon is known as the HABR. Under different severities of PVS, portal vein blood flow volume will exhibit varying degrees of decrease, and HABR can result in a compensatory increase in hepatic artery blood flow, leading to corresponding increases in blood flow volume and velocity.[14,15] In this study,

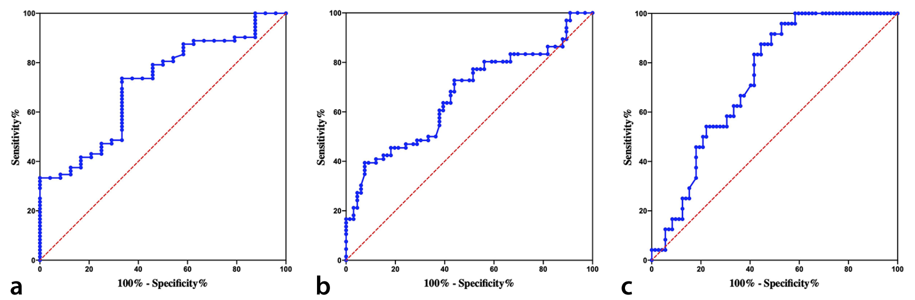


Figure 5. Receiver operating characteristic (ROC) curves of hepatic artery PSV in predicting PVS grade. **(a)** ROC curves of hepatic artery PSV in diagnosing PVS. **(b)** ROC curves of hepatic artery PSV in identifying >50% PVS. **(c)** ROC curves of hepatic artery PSV in identifying >65% PVS. PVS, portal vein stenosis; PSV, peak systolic velocity.

hepatic artery blood flow velocity in rats with different severities of PVS showed varying degrees of increase, with the most significant increase observed in cases of >65% PVS. The ROC analysis indicated that when hepatic artery PSV exceeded 51.15 cm/s, the sensitivity for identifying >65% PVS reached 87.50%. Therefore, significant increases in hepatic artery flow velocity can help predict >65% PVS. However, since the rat hepatic artery has a small inner diameter and a tortuous course, it tends to be influenced by heart rate and respiratory rate, leading to potential errors in the measurement of hemodynamic parameters by ultrasound. Consequently, the quantitative evaluation of hepatic artery compensation post-PVS requires further validation.

In conclusion, high-frequency greyscale and Doppler ultrasound can accurately demonstrate PVS and the hemodynamic changes it causes in rats. Portal vein hemodynamic parameters exhibit significant gradient changes among different degrees of PVS, classified as SR ≤50%, 50%< SR ≤65%, and SR >65%. PVDs and the PVDR are the best parameters for grading PVS. PVV can effectively diagnose the presence or absence of PVS, but its diagnostic performance in grading PVS is relatively low. The PVVR showed good performance in identifying >65% PVS. A significant increase in hepatic artery PSV may help evaluate PVS, particularly in predicting >65% PVS. These findings provide valuable information for PVS rat model research and an experimental basis for further studies on early PVS evaluation in LDLT.

Footnotes

Funding

This study was supported by the National Key R&D Program of China (2020YFA0714002).

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Bozkurt B, Dayangac M, Tokat Y. Living donor liver transplantation. *Chirurgia (Bucur)*. 2017;112(3):217-228. [CrossRef]
2. Goldaracena N, Barbas AS. Living donor liver transplantation. *Curr Opin Organ Transplant*. 2019;24(2):131-137. [CrossRef]
3. Yagi S, Hirata M, Miyachi Y, Uemoto S. Liver Regeneration after hepatectomy and partial liver transplantation. *Int J Mol Sci*. 2020;21(21):8414. [CrossRef]
4. Gad EH, Abdelsamee MA, Kamel Y. Hepatic arterial and portal venous complications after adult and pediatric living donor liver transplantation, risk factors, management and outcome (a retrospective cohort study). *Ann Med Surg (Lond)*. 2016;828-839. [CrossRef]
5. Yin C, Zhu ZJ, Wei L, Sun LY, Zhang HM, Wu HR. Risk factors for portal vein stenosis in pediatric liver transplantation. *Clin Transplant*. 2020;34(8):e13992. [CrossRef]
6. Kimura Y, Tapia Sosa R, Soto-Trujillo D, Kimura Sandoval Y, Casian C. Liver transplant complications radiologist can't miss. *Cureus*. 2020;12(6):e8465. [CrossRef]
7. Narita Y, Sugawara Y, Ibuki S, et al. Portal vein stent placement in living-donor liver transplantation: a single-center experience. *Transplant Proc*. 2019;51(5):1522-1524. [CrossRef]
8. Naidu SG, Alzubaidi SJ, Patel IJ, et al. Interventional radiology management of adult liver transplant complications. *Radiographics*. 2022;42(6):1705-1723. [CrossRef]
9. Karakaya E, Akdur A, Ayvazoğlu Soy EH, Boyvat F, Moray G, Haberal M. Vascular complications in pediatric liver transplants and their management. *Exp Clin Transplant*. 2022;20(Suppl 3):72-75. [CrossRef]
10. Ma L, Lu Q, Luo Y. Vascular complications after adult living donor liver transplantation: evaluation with ultrasonography. *World J Gastroenterol*. 2016;22(4):1617-1626. [CrossRef]
11. Delgado-Moraleda JJ, Ballester-Vallés C, Marti-Bonmati L. Role of imaging in the evaluation of vascular complications after liver transplantation. *Insights Imaging*. 2019;10(1):78. [CrossRef]
12. Byun J, Kim KW, Choi SH, et al. Indirect Doppler ultrasound abnormalities of significant portal vein stenosis after liver transplantation. *J Med Ultrason (2001)*. 2019;46(1):89-98. [CrossRef]
13. Mullan CP, Siewert B, Kane RA, Sheiman RG. Can Doppler sonography discern between hemodynamically significant and insignificant portal vein stenosis after adult liver transplantation? *AJR Am J Roentgenol*. 2010;195(6):1438-1443. [CrossRef]
14. Spaggiari M, Martinino A, Ray CE Jr, et al. Hepatic arterial buffer response in liver transplant recipients: implications and treatment options. *Semin Intervent Radiol*. 2023;40(1):106-112. [CrossRef]
15. Liu C, Song JL, Lu WS, et al. Hepatic arterial buffer response maintains the homeostasis of graft hemodynamics in patient receiving living donor liver transplantation. *Dig Dis Sci*. 2016;61(2):464-473. [CrossRef]
16. Higgins GM, Anderson RM. Experimental pathology of liver: restoration of liver of white rat following partial surgical removal. *Arch Pathol*. 1931. [CrossRef]
17. Topcu A, Yildiz A, Ozkan OF. Effect of L-carnitine on regeneration in experimental partial hepatectomy model in rats. *Ulus Travma Acil Cerrahi Derg*. 2022;29(1): 9-16. [CrossRef]
18. Van Thiel DH, Gavaler JS, Slone FL, et al. Is feminization in alcoholic men due in part to portal hypertension: a rat model. *Gastroenterology*. 1980;78(1):81-91. [CrossRef]
19. Ma L, Chen K, Yang L, et al. Ultrasonographic hemodynamics for prediction of poor liver regeneration induced by severe portal vein stenosis in rats. *Ann Transl Med*. 2020;8(8):527. [CrossRef]
20. Kawano Y, Mizuta K, Sugawara Y, et al. Diagnosis and treatment of pediatric patients with late-onset portal vein stenosis after living donor liver transplantation. *Transpl Int*. 2009;22(12):1151-1158. [CrossRef]
21. Chang WT, Kuo YT, Lee KT, et al. The value of primary vascular stents in management of early portal vein stenosis after liver transplantation. *Kaohsiung J Med Sci*. 2016;32(3):128-134. [CrossRef]
22. Chong WK, Beland JC, Weeks SM. Sonographic evaluation of venous obstruction in liver transplants. *AJR Am J Roentgenol*. 2007;188(6):515-521. [CrossRef]

DIR

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242854

ARTIFICIAL INTELLIGENCE AND INFORMATICS

REVIEW

Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects

Burak Koçak[1]
Andrea Ponsiglione[2]
Arnaldo Stanzione[2]
Christian Bluethgen[3]
João Santinha[4]
Lorenzo Ugga[2]
Merel Huisman[5]
Michail E. Klontzas[6]
Roberto Cannella[7]
Renato Cuocolo[8]

[1]University of Health Sciences, Başakşehir Çam and Sakura City Hospital, Clinic of Radiology, İstanbul, Türkiye

[2]University of Naples Federico II, Department of Advanced Biomedical Sciences, Naples, Italy

[3]University of Zurich, University Hospital Zurich, Diagnostic and Interventional Radiology, Zurich, Switzerland

[4]Digital Surgery LAB, Champalimaud Research, Champalimaud Foundation; Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

[5]Radboud University Medical Center, Department of Radiology and Nuclear Medicine, Nijmegen, Netherlands

[6]University of Crete, School of Medicine, Department of Radiology; University Hospital of Heraklion, Department of Medical Imaging,Crete, Greece; Karolinska Institute, Department of Clinical Science Intervention and Technology (CLINTEC), Division of Radiology, Solna, Sweden

[7]University of Palermo, Department of Biomedicine, Neuroscience and Advanced Diagnostics, Section of Radiology, Palermo, Italy

[8]University of Salerno, Department of Medicine, Surgery and Dentistry, Baronissi, Italy

Corresponding author: Burak Koçak

E-mail: drburakkocak@gmail.com

Received 11 May 2024; accepted 11 June 2024.

Epub: 02.07.2024

Publication date: 03.03.2025

DOI: 10.4274/dir.2024.242854

ABSTRACT

Although artificial intelligence (AI) methods hold promise for medical imaging-based prediction tasks, their integration into medical practice may present a double-edged sword due to bias (i.e., systematic errors). AI algorithms have the potential to mitigate cognitive biases in human interpretation, but extensive research has highlighted the tendency of AI systems to internalize biases within their model. This fact, whether intentional or not, may ultimately lead to unintentional consequences in the clinical setting, potentially compromising patient outcomes. This concern is particularly important in medical imaging, where AI has been more progressively and widely embraced than any other medical field. A comprehensive understanding of bias at each stage of the AI pipeline is therefore essential to contribute to developing AI solutions that are not only less biased but also widely applicable. This international collaborative review effort aims to increase awareness within the medical imaging community about the importance of proactively identifying and addressing AI bias to prevent its negative consequences from being realized later. The authors began with the fundamentals of bias by explaining its different definitions and delineating various potential sources. Strategies for detecting and identifying bias were then outlined, followed by a review of techniques for its avoidance and mitigation. Moreover, ethical dimensions, challenges encountered, and prospects were discussed.

KEYWORDS

Artificial intelligence, machine learning, medical imaging, bias, fairness, radiology

Bias, with its various definitions depending on the context, often denotes systematic errors due to existing inappropriate models, whether intentional or unintentional.[1] Extensive studies of bias in human cognition have included the field of radiology and medical imaging, addressing biases at both personal (e.g., bias during reporting) and societal levels.[2] It is typically linked to conscious or subconscious cognitive preconceptions that may arise during clinical practice, particularly in rapid decision-making scenarios.[3,4]

Advances in artificial intelligence (AI) related to medical imaging, particularly in radiology, present new avenues to enhance patient care across different stages of the patient journey, such as triage, selecting imaging modalities, image quality improvements, risk assessment, diagnosis, and prognostication.[5-7] However, increasing integration of AI into clinical practice comes with new challenges for radiologists, who may not be accustomed to potential biases or systematic errors introduced into their workflow, thereby risking the integrity of outcomes.[8-13]

Medical publication trends indicate a growing interest in bias in AI (Figure 1). This international collaborative review effort aims to provide readers with the fundamental knowledge and potential tools or strategies necessary to navigate bias when dealing with AI for medical imaging, thus mitigating negative impacts on patient management. This study comprehensively reviews bias in AI for medical imaging, covering its fundamentals, detection techniques, prevention strategies, mitigation methods, encountered challenges, ethical concerns, and prospects.

You may cite this article as: 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.

Definition of bias in artificial intelligence

The concept of bias in machine learning (ML) research and more generally in the field of predictive modeling is intrinsically tied to the concept of variance.[14] In this context, bias can be defined as the distance (or error) between the prediction and the actual target variable, whereas variance signifies the dependence of predictions on the randomness in the training data sampling (Figure 2).[15] Hypothetically, a predictive model can present any combination of high or low bias and variance. From a statistical point of view, the sum of bias (squared) and variance is represented by the mean squared error metric.[16] Interestingly, the concepts of bias and variance are not limited to the domain of statistical or ML modeling alone, but they also affect human learning and have been extensively studied in cognitive sciences.[15]

From a mathematical point of view, noise (the joint probability distribution between training and test/inference samples), bias, and variance are the three components that lead to model performance degradation and negatively affect generalization to new data.[17] Given the somewhat irreducible nature of noise, ML has focused mostly on addressing bias and variance when optimizing model performance during the hyperparameter tuning process. However, it should be made clear that these two entities are interdependent, and reducing one (e.g., variance) typically comes at the expense of increasing the other (i.e., bias), which gives birth to the concept of a bias-variance tradeoff. In recent years, the technical evolution of ML models, and especially the rise of large neural network architectures, has begun to challenge the traditional approach of validation (or cross-validation) error minimization as the ideal strategy to optimize the bias-variance tradeoff during model training.[17-20]

Types and sources of bias

One way to comprehend imaging AI bias is by examining sources of bias related to fundamental components of the AI life cycle: study design and dataset (formulating the research question, collection, annotation,

Main points

- In the medical artificial intelligence (AI) context, “bias” refers to systematic errors leading to a distance between prediction and truth, to the potential detriment of all or some patients.
- AI in medical imaging is at risk of being compromised by several types of biases, which could adversely affect patient outcomes.
- Understanding that medical imaging AI systems are prone to biases in various forms is key for their successful incorporation into real-world clinical settings, with greater satisfaction of end-users.
- Proactively identifying and addressing AI bias may prevent its potential negative consequences from being realized later.
- Increasing community awareness about all aspects of bias, such as fundamentals, mitigation strategies, and ethics, may contribute to the development of more effective regulatory frameworks.

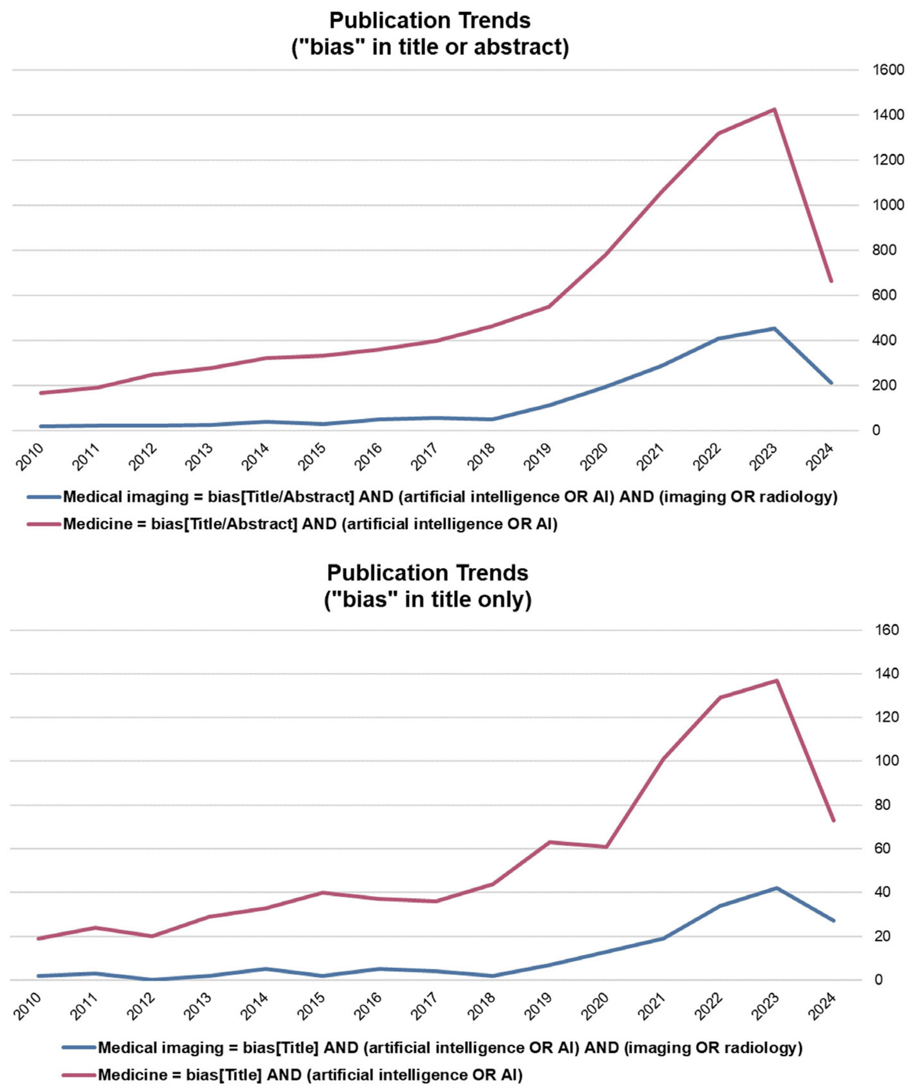


Figure 1. Publication trends about bias in medical imaging artificial intelligence (AI) in comparison with AI in medicine, with different search syntaxes to identify the occurrences of the term “bias” in the title or abstract versus the title alone. Source: PubMed; date of search: May 7, 2024.

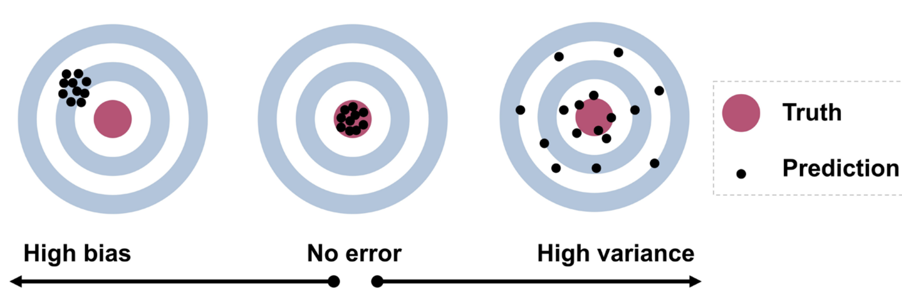


Figure 2. Over-simplified illustration of bias (i.e., systematic error) in contrast to variance, such as random noise.

preprocessing, etc.), modeling (development and evaluation before using in real-world settings), and deployment (implementation in real-world settings). This section focuses on the most common sources of bias that medical imaging professionals, particularly radiologists, may encounter. Accordingly, types and sources of bias and concepts mentioned in this review are given in Figure 3. Table 1 provides a glossary of definition of other bias sources as well, including other related concepts. Table 2 presents fictional examples for selected bias sources.

Bias related to study design and dataset

Bias can emerge when taking the very first step into the development of AI solutions for medical imaging, which is the correct identification of an unmet and relevant clinical need.[21] A valid research question must also

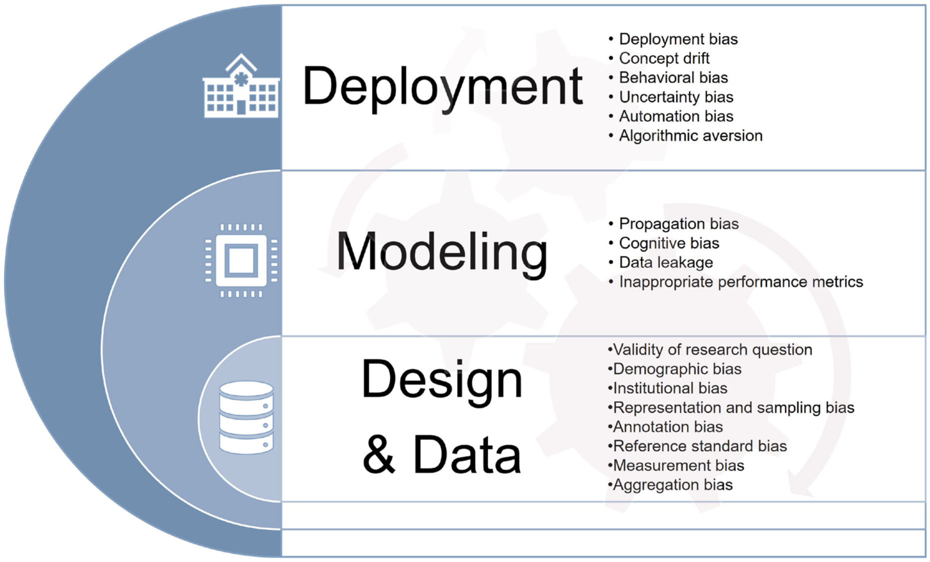


Figure 3. Main types and sources of bias and related concepts highlighted throughout this review. For other common types and sources of bias, please refer to Table 1.

Table 1. Common terminology and concepts related to bias

Terminology	Definition
Aggregation bias	False conclusions or assumptions about individuals compared with the whole population based on inappropriate combinations of distinct groups.
Algorithm fairness	Ensuring equitable outcomes across different demographic groups.
Algorithmic bias	Systematic errors or prejudices in the algorithms.
Algorithmic aversion	Reluctance or skepticism toward relying on artificial intelligence (AI) algorithms.
Annotation bias	Systematic errors mostly introduced by human annotators during the labeling process of training data, mostly related to their experience, subjective interpretation, and cognitive biases concerning the annotation task.
Automation bias	Overreliance on AI results, leading to the neglect of human decision-making.
Behavioral bias	Distortions in user behavior seen across various platforms, contexts, or datasets.
Class imbalance	Disproportionate representation of certain classes within or between the data partitions.
Cognitive bias	Systematic subjective patterns in thinking that can affect the decision-making of individuals due to reliance on heuristics (i.e., shortcut strategies derived from previous experiences to solve a problem or reach a goal).
Concept drift	Changes in correlation between input variables and output predictions over time due to fluctuations in data.
Confirmation bias	Tendency to interpret AI model results in a way that confirms their existing beliefs or expectations.
Data leakage	Exposure of target features or information to the model during training, leading to poor generalizability.
Demographic bias	Systematic errors in models that disproportionately affect specific demographic groups based on factors such as age, gender, or ethnicity.
Deployment bias	Misalignment between the envisioned purpose of a system or algorithm and its actual application.
Distributional shift	Discrepancies between the distribution of data used to train AI models and the distribution encountered in real-world deployment.
Feedback loop bias	Increase of systematic errors over time as the AI model continues to learn from its own predictions and feedback.
Institutional bias	Systematic errors led by differences in practices, protocols, or equipment across institutions.
Measurement bias	Systematic errors related to how particular features are chosen, used, or measured.
Omitted variable bias	Systematic errors appear when one or more relevant variables are omitted, or context is neglected.
Overfitting	Phenomenon where the AI model learns to memorize the training data instead of generalizing on new data.
Propagation bias	Increase of potential systematic errors present in any algorithm or pipeline and being inherited by the final model or even amplified in it.
Representation and sampling bias	Systematic errors in the collection of data, resulting in an unrepresentative sample.
Statistical bias	Discrepancies between actual and predicted values when approximating a specific statistical measure.
Temporal bias	Systematic errors arising over time, such as from the changes in medical imaging technology, protocols, or patient demographics.
Temporal drift	Changes in the distribution or characteristics of data over time, leading to discrepancies between the development and deployment AI performance.
Uncertainty bias	Influence of uncertainty on decision-making stemming from AI models.
Underfitting	Phenomenon where the AI model is too simplistic, failing to adequately capture the complexity of data.

Table 2. Examples based on fictional scenarios for selected bias sources related to medical imaging

Bias source	Example
Annotation bias	A breast artificial intelligence (AI) tool is being developed to assist in analyzing mammograms. As radiologists annotate the images to be used for its development, they primarily focus on identifying malignant masses due to their significance in cancer diagnosis. Benign calcifications, less concerning but still important, may be underrepresented in the annotations made by the radiologists. The resultant tool may have this annotation bias, being more inclined to detect malignant masses and neglecting to adequately recognize benign calcifications, leading to an increased risk of false negatives.
Automation bias	A radiologist or a clinician relies on an AI tool to interpret chest computed tomography (CT) scans. If the AI model is trained on datasets that predominantly include lung nodules, it may develop a bias toward detection of these nodules over other clinically significant findings (e.g., consolidations). By developing a tendency to prioritize the AI tool's output over the entire clinical evaluation, end-users may show an over-reliance on the AI tool, trusting it without thoroughly considering other important information present in the CT scans. This automation bias can result in missing important findings beyond lung nodules.
Confirmation bias	An experienced radiologist uses an AI tool to analyze a prostate magnetic resonance imaging (MRI) scan of a patient with a history of urinary symptoms and elevated prostate-specific antigen levels. As the radiologist examines the imaging results, they may identify certain features that appear to support their initial suspicion of benign prostatic hyperplasia (BPH) based on the observed prostatic enlargement and nodularity. However, the tool also flags some potential small focal lesions or suspicious tissue characteristics, suggestive of prostate cancer. Despite these, the radiologist's focus on confirming their preliminary diagnosis of BPH may lead them to ignore the important alerts provided by the tool. The cognitive bias of the radiologist toward confirming their previous suspicion of BPH influences their MRI interpretation.
Demographic bias	Radiologists utilize an AI tool to analyze abdominal CT scans. If the AI model is trained on datasets that primarily includes younger patients, the AI tool may not be effectively trained to recognize age-related conditions commonly found in older individuals, such as diverticulosis. Consequently, when presented with abdominal CT scans from older patients, the model may experience difficulty in accurately identifying and assessing these pathologies, due to age-related demographic bias.
Feedback loop bias	Radiologists rely on an AI algorithm to assist in analyzing brain MRI scans. If the algorithm is initially trained on datasets mostly featuring images with clear and prominent lesions, such as large tumors, it may develop a bias toward identifying these abnormalities with high accuracy. Users of this tool may subconsciously prioritize confirming the presence of these well-defined lesions, providing feedback that reinforces the AI's accuracy in detecting such cases. Consequently, the model may improve its performance at identifying large lesions while potentially ignoring smaller, subtler, early-stage abnormalities, especially if they were underrepresented in the initial training data. This feedback loop between the AI model and the end-users, such as radiologists, can perpetuate bias, leading to a situation where the AI becomes increasingly adept at detecting certain types of abnormalities while potentially missing others.

be properly formulated so that it can be effectively translated into a fitting task for AI.[22] Any flaw in these essential starting points inevitably generates a bias in the subsequent steps, such as the selection of training datasets, AI model development, and/or deployment.

Bias in the dataset collection and preparation phases can significantly affect the outcomes of AI systems, particularly in the critical domain of medical imaging. This bias can stem from a variety of sources and can lead to disparities in the performance of AI systems across different patient groups, potentially exacerbating existing health inequalities.[23]

One of the primary sources of bias in medical imaging datasets is demographic imbalance. For example, if a dataset predominantly consists of images from a particular racial or ethnic group, the AI model trained on this dataset may exhibit reduced accuracy when applied to individuals from other groups. This situation can lead to misdiagnoses or delayed diagnoses for underrepresented groups. Similar issues arise with gender, age, and socio-economic status, where AI systems may perform better for the demographic groups that are overrepresented in the training data (Figure 4).[24]

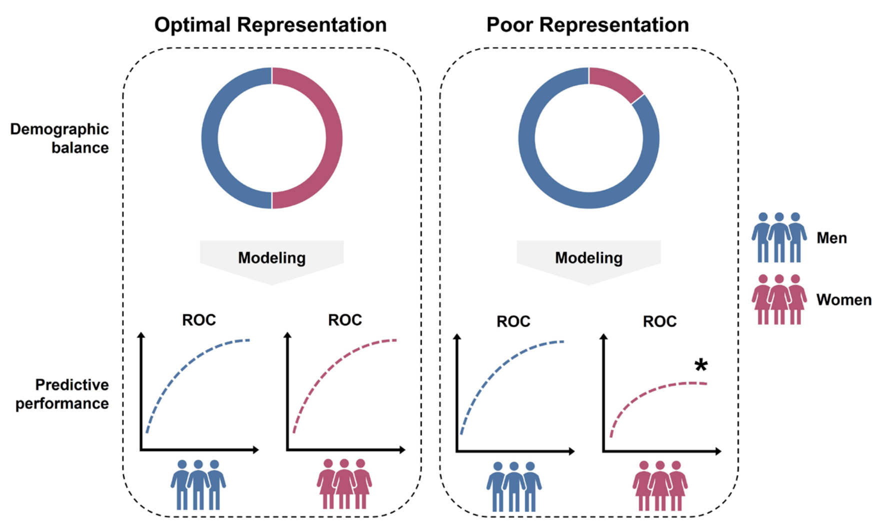


Figure 4. Over-simplified illustration of optimal and poor representation of subgroups, such as gender in this case, and their effect (*) in subsequent modeling. ROC, receiver operating characteristics.

Another critical aspect is the quality and source of the medical images. Bias can be introduced if the images come from a limited number of institutions or geographic locations, as different places may use varying equipment, protocols, and standards for image capture. This can ultimately contribute to covariate shifts (distributional differences of features between training and test sets) (Figure 5). Such variations can cause AI systems to become overfitted to the characteristics specific to the data they were trained on, reducing their generalizability and effectiveness when deployed in different settings.

The preparation of datasets also introduces potential biases (Figure 6). The process of labeling medical images, which is often performed by human experts, can lead to inconsistencies due to subjective interpretation of what the images represent and in turn to annotation bias. Moreover, if a small group of experts annotates the dataset, their individual biases and level of expertise can influence the labels, affecting the AI model's learning process. A broader concept than annotation bias is reference standard bias, affecting the way instances are labeled and consequently impacting algorithm development.[25] Different reference standards are often available to confirm radiological diagnosis, which may also lead to systematic errors.[26] Some could be highly accurate but also costly and poorly available, whereas others could neglect intermediate findings or be operator-dependent,[27] potentially reducing label applicability and reliability. Additionally, the choice of data preprocessing techniques, such as normalization, augmentation, or cropping, can also influence the model's output by emphasizing certain features over others.[28]

Moreover, bias can stem from broader historical and societal inequities that are reflected in the data. For example, certain diseases may be more prevalent in specific populations due to factors such as access to healthcare, environmental exposures, or genetic predispositions. If these factors are not adequately considered during dataset collection and AI model training, the resulting models may not only perpetuate but also amplify existing disparities.

Bias related to modeling

The development of AI models is a multi-step process, and different AI algorithms are frequently employed at different stages, such as image segmentation, feature reduction, and selection.[29] Therefore, potential bias present in any of the algorithms will propagate down the pipeline and be inherited by the final model or even amplified in it, resulting in propagation bias. It should also be considered that, since humans are developing AI models, the latter can also inherit cognitive bias from the former.[3] This is not specific to the model development stage alone and can potentially occur at any point in the AI lifecycle (Figure 7).[30]

AI modeling also includes a validation step, necessary to confirm the performance of the algorithms before actual deployment. This should ideally be verified on publicly available benchmark datasets to ensure a common ground for model testing, as seen in AI challenges. Nevertheless, further testing on independent data remains pivotal to verify that all requisites for deployment are met. In this context, a common and serious source of bias in model validation lies in data leakage.[31] An example of data leak in medical imaging is represented by the inclusion of different scans from the same patient both in the training and validation dataset, which increases the risk of overfitting.

Another aspect to carefully consider is the choice of metrics used to estimate the model's performance, which could introduce bias if those selected do not match the information needed. A case example is the validation of automated segmentation tools, for which specific parameters should be selected based on the segmentation task characteristics (e.g., is it more important to have an accurate segmentation or a precise localization for the task?).[32]

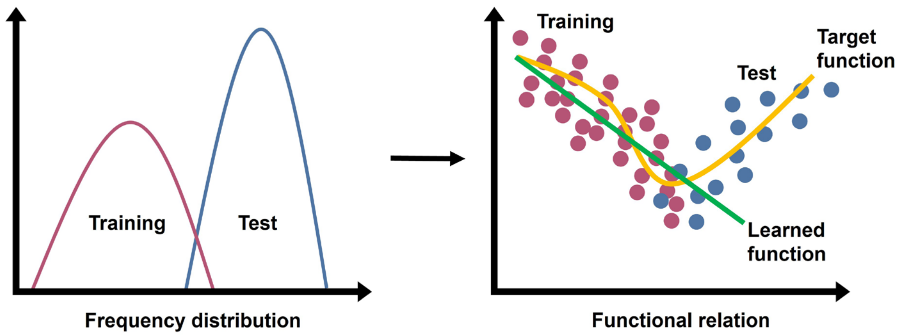


Figure 5. Over-simplified illustration of covariate shift. Distributional differences between training and test sets lead to poor test performance (i.e., poor generalizability) or significant deviation from the learned function.

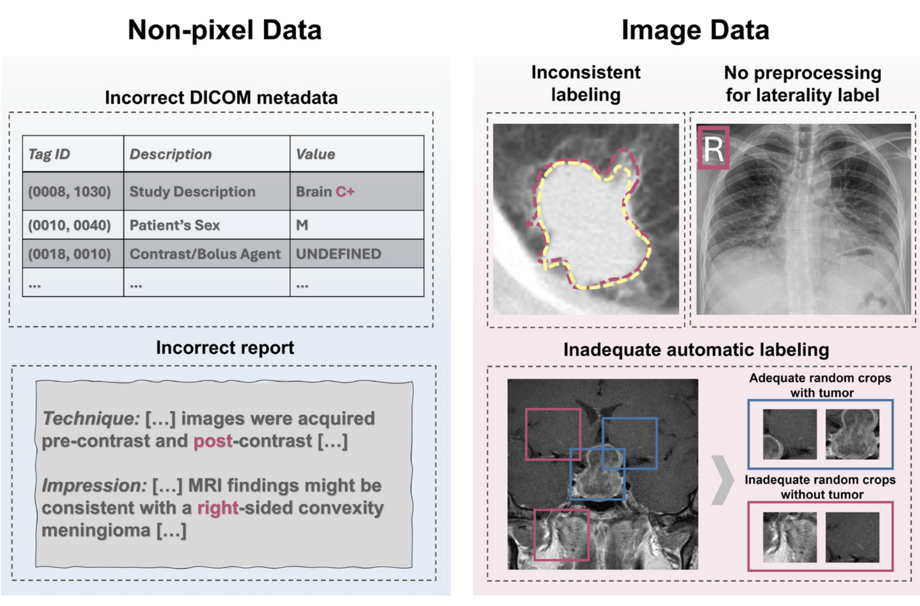


Figure 6. Potential and practical bias sources relevant to medical imaging artificial intelligence based on data type (i.e., non-pixel and image data). Radiological images belong to chest computed tomography (upper left panel), chest X-ray (upper right panel), and pituitary magnetic resonance imaging (lower panel).

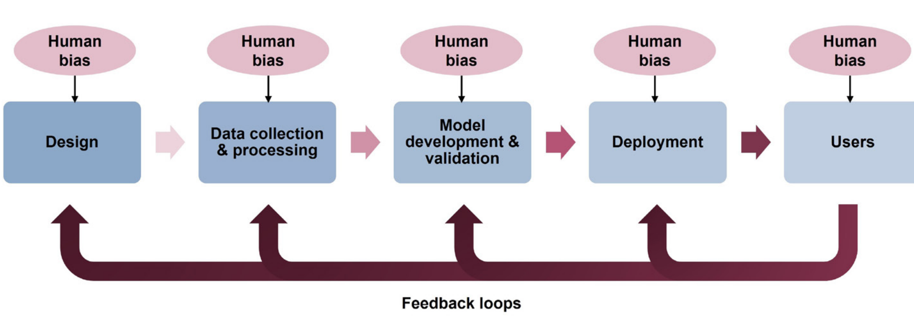


Figure 7. Human bias in the artificial intelligence life cycle.

Finally, the model's performance needs to be put into context, correctly selecting valid baseline alternatives for comparison, such as already recognized diagnostic tests, and formally evaluating with statistical approaches the added value that the model may bring.[33]

Bias related to deployment

Model deployment represents the final phase of AI/ML algorithms for medical imaging, following data collection and evaluation.[34] It involves assessing the model's performance in real-world scenarios, including potential application in clinical practice.[35]

A deployment bias emerges when there is a misalignment between the envisioned purpose of a system or algorithm and its actual application.[36] In medical imaging, this bias can manifest when an algorithm designed for segmentation tasks is utilized by human operators, whether intentionally or inadvertently, as a detection tool instead.[37] Additionally, improper utilization by end-users can also arise when utilizing systems to analyze images from anatomical districts or imaging modalities that differ from those they have been trained and validated with-for example, employing abdominal computed tomography images instead of abdominal magnetic resonance images.

Concept drift represents an additional source of bias for model deployment (Figure 8). Specifically, it arises when the correlation between input variables, such as images, and output predictions, such as diagnoses, evolves due to fluctuations in data, such as variations in image acquisition hardware or protocols, shifts in disease prevalence, or advancements in gold-standard technologies.[38]

Behavioral bias pertains to the potential distortions in user behavior seen across various platforms, contexts, or datasets.[39] Factors such as past experiences, social stigma, exposure to misinformation, limited healthcare access, and historical context play a role in shaping this bias. In particular, this bias can lead to skewed data cohorts, incomplete information, heightened uncertainty in outcomes, and potential dismissal of algorithm-assisted medical advice.[40]

Uncertainty bias encompasses the influence of uncertainty on decision-making stemming from AI/ML models.[39] Precisely characterizing and estimating uncertainty is pivotal in ensuring the thorough evaluation and transparent reporting of AI/ML models. Nonetheless, human observer decisions relying on AI/ML model outputs and their reported uncertainty may be unduly swayed by the uncertainties inherent in the model's output.[41] Consider this scenario: AI/ML models can be "confidently wrong," meaning they may yield incorrect outcomes with a high level of certainty. Consequently, humans may place greater importance on a prediction that exhibits high certainty, even if it happens to be incorrect, compared with one with lower certainty that is actually correct.

Automation bias refers to the tendency of individuals to rely excessively on automated systems, such as AI algorithms, and to disregard or underutilize their own judgment or critical thinking skills.[42] In the context of AI in medical imaging, automation bias can manifest when clinicians or radiologists place undue trust in the outputs or recommendations provided by AI algorithms, leading them to overlook potentially important information or make errors in diagnosis or treatment planning.[43] Automation bias can occur in busy clinical settings where clinicians may feel pressure to make rapid decisions, leading them to rely on AI-generated results as a shortcut rather than engaging in thorough analysis.[44] Additionally, clinicians may tend to seek out or interpret information in a way that confirms their preexisting beliefs or expectations. If an AI algorithm's recommendation aligns with their initial impressions, they may be more likely to accept it without question. A lack of adequate training or education on how to effectively integrate AI algorithms into workflow may favor automation bias.[45]

Algorithmic aversion refers to a phenomenon where clinicians or healthcare professionals exhibit reluctance or skepticism toward relying on AI algorithms for making diagnostic or treatment decisions in medical imaging.[46] This bias can manifest due to several reasons, such as trust issues on algorithms' reliability, transparency, or interpretability or a lack of familiarity, fear of job displacement, or even ethical and legal concerns.

Bias detection/identification

Detecting bias in AI algorithms necessitates awareness of all sources of bias, including those that have to do with the dataset and the development and evaluation of AI algorithms as well as those related to the deployment of these algorithms, such as human user biases and inference. Methods for bias detection vary according to the type of bias. One of the first strategies that can be used to identify bias related to the dataset is dataset evaluation against a set of predefined criteria (searching for exclusion bias, selection bias, recall bias, observer bias, and prejudice bias) and comprehensive data analysis.[47] Unsupervised analysis of the training dataset, using methods such as principal component analysis and hierarchical clustering, can be used for the detection of patterns in the training dataset that may be otherwise occult, highlighting data skewness. Statistical comparison of model output according to different patient groups or confounders that may exist in the training dataset, such as the gender or age of patients. Potential discrepancies in group results could indicate a source of bias that can affect the final results.[48] Visualization of algorithm output with methods such as class activation heatmaps can help detect discrepancies related to such potential confounders.

The next step in bias detection is the evaluation of the model development process. This starts with a code review that can be

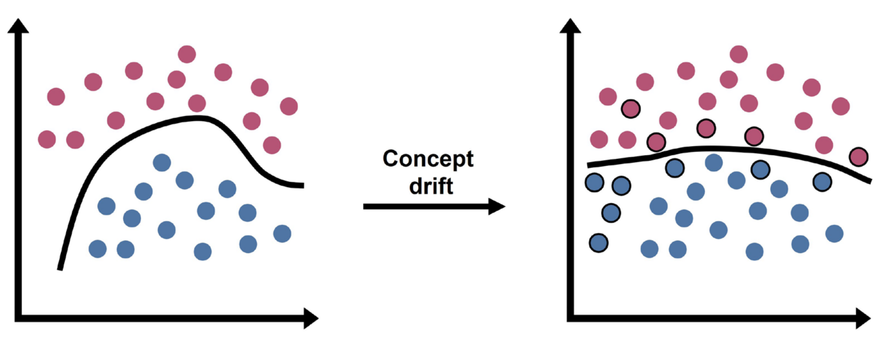


Figure 8. Over-simplified illustration of true concept drift while adding new data over time, resulting in changes in the relationship of input features and the target variable and ultimately in model behavior. In medical imaging, this may result from, for instance, a change of reference standard (e.g., new guidelines) in determining the target variable or a difference in the distribution of underlying data. It is also possible that such changes, particularly changes in data distribution, may result in virtual drifts with no obvious difference in model behavior.

carried out by an independent experienced coder/auditor.[49] Companies such as Google have developed methods for anonymous code review by several experts.[49] Such a code review can be also performed retrospectively by the scientific community for manuscripts published with open-access code.[50] Once the code has been scrutinized for potential bias, comprehensive testing should be initiated. This testing should extend from the evaluation of model performance in populations unseen in the training dataset (e.g., assessment of model performance in a pediatric population even though the algorithm was not trained with child data) to explainability analysis.[51] Simulation methods testing algorithms in various scenarios with Bayesian parameter search have been proposed to identify bias sources of algorithmic performance reduction.[52] Several explainability methods have been used that include saliency maps, such as gradient-weighted class activation mapping (CAM) and integrated gradients. Evaluation of the results of saliency maps necessitates extra care, as concerns have been raised about the reliability of these methods.[53,54]

To detect bias related to the use of the developed algorithm, human factors as well as economic, ethical, and legal factors need to be evaluated. Testing by a variety of user groups with variable experiences and backgrounds can identify human user bias. Receiving feedback with user interviews and monitoring the results per user group can help locate performance outliers or imbalances related to human factors. In addition, deep learning systems that reduce the variability in human actions leading in turn to bias reduction can be useful.[55] Auditing by legal and ethics experts can also reveal issues related to the successful deployment of the model.[56,57]

To identify and flag bias in AI publications, tools have been developed to assist the writing process of AI manuscripts.[58,59] One of these tools is the Prediction Model Risk of Bias ASsesment Tool (PROBAST), which was developed in 2019 to enable the critical evaluation of studies presenting predictive models. The current version of PROBAST evaluates the risk of bias in four potential bias categories: participants, predictors, outcomes, and analysis.[60] Nonetheless, the current version of PROBAST is not suitable for the evaluation of ML studies, and this is the reason that the PROBAST group has initiated the process of developing an AI-specific version of PROBAST called PROBAST-AI, which is still under development.[61] For systematic reviews of AI studies, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) has been widely used to detect the risk of bias.[62] The QUADAS-2 tool includes 14 questions and provides an estimate of the risk of bias in the study, categorizing it as high, low, or unclear. Reporting guidelines, such as the Fairness Universality Traceability Usability Robustness Explainability-AI and TRIPOD-AI, can assist authors of AI manuscripts in reporting their studies according to the Fairness principle, promoting the identification of bias sources.[58,63,64] When dealing with radiomics studies, the CheckList for EvaluAtion of Radiomics (CLEAR) and METhodological RadiomICs Score (METRICS) have been developed to evaluate the reporting and methodological study quality.[65,66] Among the items evaluated, CLEAR item#7 and METRICS item#1 require adherence to reporting guidelines similar to those mentioned above; CLEAR item#36 and METRICS item#19 require the consideration of confounding factors related to dataset preparation that are closely related to bias.

Avoidance strategies

Ideally, bias should be prevented before it becomes embedded within AI systems. The focus of strategies employed during the planning, data collection, and model training phases of creating AI systems is on prevention, setting a course that avoids the pitfalls of bias rather than correcting for it post-hoc.

To mitigate bias and potentially avoid it, medical AI system development should adhere to ethical AI design principles. Guiding principles, such as transparency, fairness, non-maleficence, and respect for privacy from the outset, are widely included in recommendations and position papers and can help to prevent bias (Figure 9).[67] Transparency increases explainability, interpretability, and similar acts of communication and disclosure, which in the context of bias mitigation applies to the explicit, proactive thought about which training data are used, and how they are collected, processed, and employed. Fairness refers to an impartial treatment without favoritism or discrimination. In the context of preventing bias, fairness can be pursued by creating and upholding design standards that respect diversity, equity, and inclusion. Non-maleficence is a core medical principle. AI systems should never cause foreseeable or unintentional harm, for instance through discrimination or suboptimal patient management, which can be a direct result of biased models.[13] Respect for privacy is an important ethical principle, particularly in healthcare. In the context of mitigating bias, upholding this principle requires careful risk-benefit analyses to balance incorporating more data with the need to provide individuals control over their own data.

By incorporating the above-mentioned considerations early into the design phase, developers can create systems that are less likely to perpetuate or amplify biases. This involves rigorous ethical review processes and early stakeholder consultations to guide the decision-making process. The composition of the involved teams can influence the AI's propensity for bias. Teams that are diverse in terms of gender, ethnicity, culture, and professional background bring a wide array of perspectives to the table, which can help identify and eliminate potential biases early in the development process.

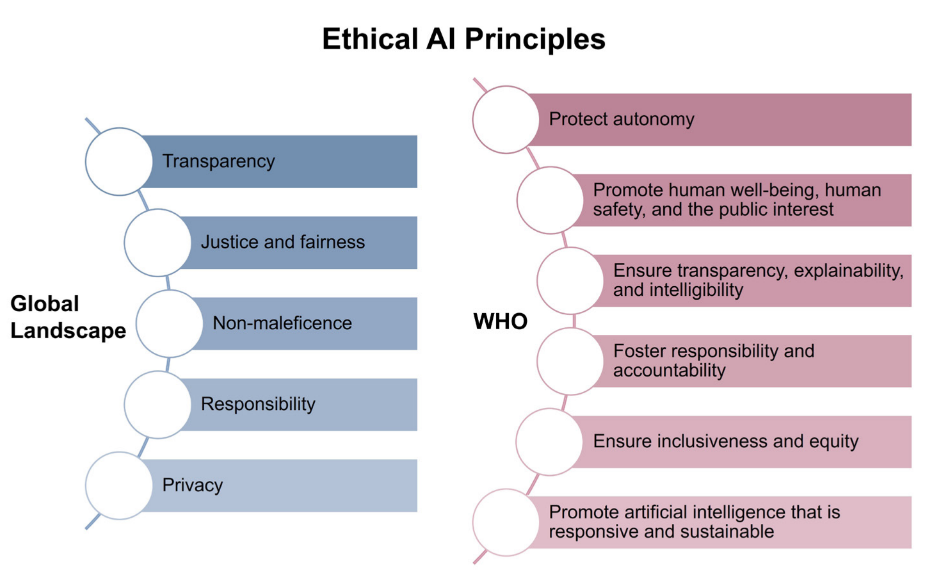


Figure 9. Key ethical artificial intelligence principles. WHO, World Health Organization.

AI systems may transport various types of bias stemming from their underlying training data.[68,69] At the data collection and processing phase, these include measurement bias (how particular features are chosen, used, or measured), omitted variable bias (when one or more relevant variables are omitted or context is neglected), representation and sampling bias (incorrect sampling leads to insufficiently diverse or otherwise non-representative datasets), and aggregation bias (false conclusions about individuals from observing whole populations).[69,70] These issues warrant thoughtful data collection and processing to ensure that datasets are representative of the diversity of the population or phenomena they are intended to model. It requires sourcing data from a wide range of demographics, geographies, and contexts to capture a broad spectrum. Nonetheless, even data collected following these principles may still reflect existing structural and historical biases.

Apart from collecting more data, strategies at the data processing stage may include the creation of more representative training datasets by data augmentation (e.g., by specifically adding underrepresented examples to the data through additional sampling or data generation) or data filtering (e.g., actively undersampling or filtering out undesirable or non-representative samples).[68] Generative AI models, such as large language models or vision-language models capable of synthesizing images, additionally allow for tailored data augmentation by creating new examples that meet a set of targeted criteria.[71-73] An overview of bias avoidance strategies at the data processing phase is presented in Figure 10.

The way data are presented to the model during training (affected by the problem formulation and the labeling methodology) and how model parameters are updated (defined through training setup including the objective function) can introduce bias into the model.[13,68] A classic example is optimizing a model for overall accuracy, which may severely impact the model performance on minority class samples in imbalanced setups. Other techniques, such as pruning, aiming to compress the model may also disproportionately impact underrepresented subsets in the data.[74] Careful design of the training setup can help avoid biases at this stage.

Transparent and comprehensive documentation of the AI system's design choices, data sources, and any assumptions made during development (e.g., through model cards)[75] is crucial and can help spot sources of bias before, during, and after training. Additionally, especially in the context of foundation models, detailed documentation may help developers seeking to use larger models' outputs to train smaller models to prevent propagating bias existing in the teacher model to downstream models.

Mitigation strategies

This section reviews different approaches and algorithms to mitigate biases. Bias mitigation algorithms can be divided into three types according to the phase in which they are applied: in a preprocessing phase, during model training, or after model training.[76] Additionally, algorithms can be categorized according to whether they explicitly or implicitly address bias by accessing or not accessing the bias variables during training.[77]

The bias mitigation algorithms applied in the preprocessing phase are motivated by the fact that many of the errors in ML models arise from biases inherent in the data used to train them. Additionally, these are independent of the model and can be used in a black-box setting by altering the data distribution to increase model fairness.[76] To achieve this effect, discriminatory effects within data are first quantified and then removed or accounted for. Several specific mechanisms for handling discrimination have been proposed to create a fair training distribution.[76]

Re-sampling and re-weighting algorithms focus on rebalancing the class distribution by adjusting the sample probability/loss weight for majority/minority samples.[78-83] Nabi and Shpitser[84] rely on causal inference to estimate the effects of specific variables on the outcome, allowing them to transform the inference problem on a specific distribution into another fair distribution to train the model. Despite addressing what can be considered the root of the fairness issue, this approach may need unrealistic assumptions about the training distribution or result in the loss of information that is implicit in the original data.

Other algorithms, such as distributionally robust optimization[85] and variations,[86] ensembling approaches,[87-89] adversarial debiasing,[90-95] invariant risk minimization,[96] invariant causal predictors,[96,97] limited capacity models,[98-100] and gradient starvation mitigation,[101] have been proposed to mitigate bias during model training by updating the objective function or imposing constraints on the model, with the last two methods implicitly achieving this.[77]

Finally, another set of methods mitigates bias in a post-processing phase after model training by changing prediction based on fairness constraints.[76] Hardt et al.[102] proposed a methodology for achieving equalized odds and equality of opportunity, whereas Pleiss et al.[103] proposed calibrated equalized odds. Woodworth et al.[104] used equalized odds to propose learning non-discriminatory predictors, and Kamiran et al.[105] used decision theory to suggest reject option-based classification and discrimination-aware ensemble for discrimination-aware classification. Lohia et al.[106] proposed a post-processing method for individual and group debiasing. These post-processing methods can be used in black-box settings, similar to preprocessing methods, as they do not require access to model parameters.[76]

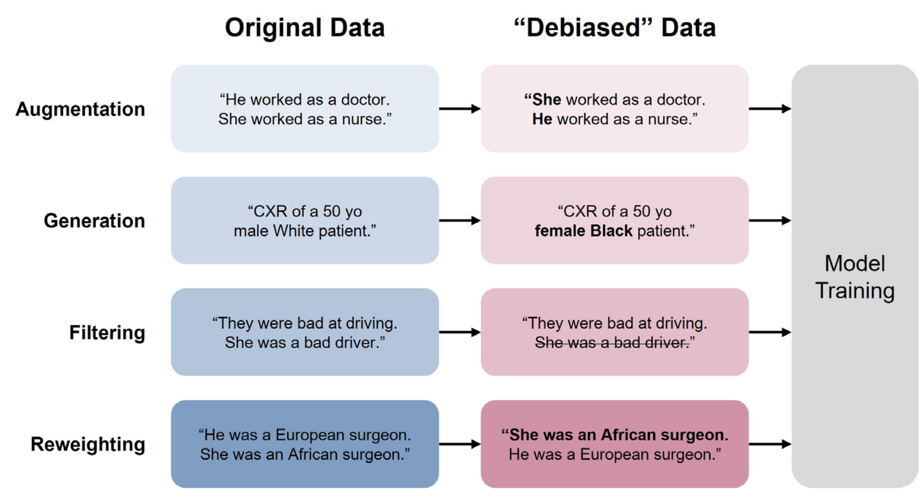


Figure 10. Overview of bias avoidance strategies at the data processing phase. Adapted from Gallegos et al.[68] CXR, chest X-ray.

In addition to active bias mitigation techniques, explainable artificial intelligence (XAI) methods offer insights into the key features influencing a model's predictions and identify and understand the significance of features driving a model's decisions. This understanding is crucial for uncovering limitations and biases in AI applications within medical imaging. These methods help us discern if confounders or biases are present in the model, allowing for their control or removal.[107] In general, XAI methods can be categorized into two main groups: perturbation-based and backpropagation-based explanations. Perturbation-based methods include occlusion,[108] LIME,[109] SHAP,[110] and various forms of perturbations.[111-113] Backpropagation-based methods encompass well-known techniques, such as saliency map visualization,[114] CAMs,[115] and their extensions.[116-118]

Potential challenges

Handling bias in AI systems is crucial for ensuring fairness and equity in decision-making processes. However, there are several challenges in handling bias that can be related to ambiguities in interpreting results, limited diversity in benchmark datasets, and the subjectivity of detecting bias.

Ambiguities in interpreting results can pose significant challenges in the development and clinical use of AI software. These refer to situations where the interpretation of the results is not unique or is open to multiple meanings by the users. Ambiguities can also originate during the different applications of the AI tools from the intended use statement provided by the AI developers, increasing the risk of off-label or erroneous applications of AI in clinical practice.[119] For example, AI software trained for adult fracture detection is at risk of erroneous results if applied in a pediatric population.

Limited diversity in benchmark datasets can represent a significant challenge in AI development and generalizability. This can occur when some diseases or events are collected with underrepresentation or overrepresentation compared with their prevalence in the general population or clinical practice due to the limited patient diversity included in the training data; this causes a class imbalance due to an uneven distribution between the training data and the actual population to which the AI model is applied.[120] As AI tools learn from archival data, a narrowed data source results in AI models that are not generalizable in heterogeneous patient populations with different demographics, clinical characteristics, and disease prevalence, leading to perpetuated bias in the final AI model.[120,121] Publicly accessible benchmarks are essential for comparison for AI models and represent a crucial element of open science.[122] Multicentric databases can potentially overcome this challenge by collecting a large number of diverse and representative data in rarer conditions. Currently, these publicly available datasets are limited to a narrow spectrum of diseases or countries of origin of the patient population.[123] Different demographic and clinical characteristics should be included to ensure a real-world representation in benchmark datasets.[48] However, although sharing data is essential for developing robust AI tools, patient privacy when collecting medical information can pose significant challenges.[124] Furthermore, real-world data are affected by missing or incomplete clinical values in retrospective cohorts and heterogeneity of clinical and laboratory parameters with their standard of reference. Image quality, noise, and acquisition parameters represent additional challenges in handling bias in multicentric cohorts. In the current radiological literature, there are ongoing difficulties in sharing benchmark datasets, with fewer than approximately 6% of all published articles in radiology journals partially or completely sharing the experimental data used to build the AI models.[125] Finally, data labeling for model training can be affected by the human image interpretation and diagnostic performance of the selected reference standard for the investigated condition.[121]

Identifying the source of bias in AI tools is also a relevant challenge. Subjectivity in the detection of bias can be related to personal interpretation and individual perspectives related to the identification of the bias itself. The complexity of AI tools makes it difficult to detect. Moreover, different sources of bias can contribute to the generation of bias, including the data source, algorithm, and users, which makes the identification more cumbersome.[124] Ultimately, identifying and addressing bias in AI will require significant effort for algorithm transparency, data source and processing, and final model utilization.

Ethical considerations

Ethical considerations are important in all steps of the AI pathway, from identifying a use case to post-market surveillance. It is important to ensure the technology promotes well-being, minimizes harm, and distributes benefits and harms justly among all stakeholders.[126] The World Health Organization (WHO) poses six key ethical principles for AI in healthcare in their framework (Figure 9): (1) protect autonomy, (2) promote human well-being, human safety, and the public interest, (3) ensure transparency, explainability, and intelligibility, (4) foster responsibility and accountability, (5) ensure inclusiveness and equity, and (6) promote AI that is responsive and sustainable.[127]

The WHO principles 2 and 5 address bias, mandating that AI tools prioritize human well-being, safety, and public interest. Ensuring AI's safety and efficacy in medical imaging demands rigorous testing, validation, and ongoing monitoring to mitigate harms and biases. Cost-effectiveness analyses and environmental awareness are both crucial to prevent unnecessary burdens on society, patients, and our environment.

Addressing biases in AI, particularly those affecting inclusivity and equity based on gender (identity), ethnicity, and socio-economic status, requires thorough subgroup analyses. The 2020 Dutch case against the "system risk indication" tool, which violated privacy laws and wrongly identified innocent people as fraud suspects, underscores the impact of such biases.[128]

Additionally, the lack of diversity among developers and researchers can worsen these issues, as teams may unconsciously favor perspectives similar to their own. Therefore, enhancing team diversity and unconscious bias training is crucial for mitigating bias in AI development.

Central to data ethics in AI are principles such as informed consent, privacy, data protection, and transparency. Currently, patients can decline being evaluated by AI-based tools according to the right to informed consent for any procedure in the hospital.[129] Patients should be given comprehensive information about how AI is used in their care, including any limitations or biases of the AI system that may affect their treatment. This may, however, eventually become infeasible when AI is deeply integrated into healthcare, and refusing AI may then compromise an individual's access to care. An alternative may then be a human-in-the-loop and a rigorous monitoring system.[130]

Ultimately, to protect patients, the ethical use of AI including mitigating biases needs to be captured in regulations. The recent European Union's AI Act serves as a pioneering legal framework aimed at regulating AI use, particularly in high-risk applications such as healthcare (as defined in Article 6). Set to fully take effect in 2026, the act governs the development, deployment, and use of AI, ensuring safety, transparency, and adherence to ethical standards across the EU. Article 10 mandates that for high-risk AI systems, training, validation, and testing datasets must be relevant, representative, error-minimal, and complete for their intended use. Additionally, it requires rigorous data governance, including bias examination and mitigation measures, to prevent impacts on health, safety, fundamental rights, or unlawful discrimination, particularly when data outputs affect future inputs. Concerning monitoring, Article 61 of the legislation mandates that developers of high-risk AI systems establish ongoing, systematic post-market surveillance mechanisms. Critiques of the act highlight liability gaps and tension between its vague yet stringent stipulations, potentially stifling innovation and escalating healthcare costs through the compliance burden.[131]

Prospects

Despite the above challenges, proactive efforts are expected to avoid and mitigate bias in AI for medical imaging in the future. Addressing bias in medical imaging AI is a dynamic landscape with many opportunities for innovation. Before going into detail, it should be acknowledged that expecting completely bias-free systems may be unrealistic.

Developing new bias detection methods is a promising future direction. More sophisticated algorithms that can identify and measure biases, including subtle discrimination, may be developed by researchers. Even though AI models are assumed to be biased, AI-based bias auditing tools may be leveraged to help mitigate bias.[132,133] To reflect diverse healthcare landscapes and disparities across countries and regions, initiatives to improve diversity and representativeness in datasets, possibly globally, may support this goal.[123] Such initiatives should aim to reduce AI system biases by compiling larger and more inclusive data repositories from diverse demographic groups and geographic regions.[123]

Additionally, bias or fairness-aware algorithms for medical imaging applications may be promising.[134] These algorithms can ensure equitable outcomes across patient populations. Because collaboration across disciplines is key to progress in this field, experts from computer science, medicine, ethics, and policymaking can collaborate to address bias in AI medical imaging from multiple perspectives.[39] Resultant algorithms must be explainable with transparent methods so these can be further studied and debated in the future.[135-137] AI companies should be encouraged to actively participate in independent research on AI biases and algorithms to improve fairness.

After training, an AI algorithm can be locked or adaptive.[138] Instead of becoming outdated after a few years, the AI model could be updated continuously as it learns from new data. Continuous learning can increase bias if the new data are biased.[139] Continuous monitoring of models should address biases that may arise over time to ensure the integrity of AI medical imaging systems in real-world clinical settings.[10,48,140] By identifying and addressing biases, these systems can improve healthcare outcomes and equity. Independent experts or organizations can audit these regularly.

AI system development and deployment in healthcare should require adherence to certain ethical guidelines and standards, which need to be improved over time considering the dynamic nature of these tools. These guidelines should explicitly deal with AI bias and fairness as well. Stronger regulatory oversight and accountability mechanisms, such as the Food and Drug Administration's action plans and the European Union's AI act, are needed to ensure that AI medical imaging systems meet bias and trustworthiness standards without hindering AI innovation.[141-143]

Final remarks

Understanding that medical imaging AI systems are sensitive to biases is key for their effective real-world integration into clinical practice. As technology progresses, the AI community should prioritize addressing bias throughout the entire AI lifecycle, starting from the research question to data collection, data processing, model development, model evaluation, and eventual real-world deployment. For this purpose, we present collective recommendations in Table 3.

Despite the aspiration for unbiased AI, complete inclusivity of all data types and sources remains an unattainable goal in model development. Nevertheless, by leveraging diverse datasets, integrating fairness-aware systems or bias assessment tools, and promoting interpretability and explainability methods, the future -and also the AI itself- may hold great promise to mitigate bias and enhance patient care outcomes. Even so, developers and clinicians must acknowledge the inherent limitations of AI methodologies and potential biases, similar to traditional diagnostic tools, to ensure the ultimate clinical decisions are based on clinical context and benefit all patients equitably. Being at the forefront of AI implementation, medical imaging professionals, particularly radiologists, are positioned to lead efforts toward unbiased AI integration in healthcare.

By offering a comprehensive review of critical aspects, but without a detailed technical discussion, we hope this review effort raises awareness within the medical imaging community about the importance of identifying and addressing AI bias proactively to prevent its impact from being realized later.

Acknowledgement

Language of this manuscript was checked and improved by Grammarly and partly by QuillBot, which are technologies powered by generative AI. The authors conducted strict supervision when using these tools.

Conflict of interest disclosure

Burak Koçak, MD, is Section Editor in 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. Roberto Cannella has received support for attending meetings from Bracco and Bayer; research collaboration with Siemens Healthcare. Christian Bluethgen has received support for attending conferences from Bayer AG, CH. He has also received research support from the Promedica Foundation, Chur, CH. Merel Huisman has received speaker honoraria from industry (Bayer); consulting fees (Capvision, MedicalPhit). Other authors have nothing to disclose.

Funding

This study has not directly received any funding. Roberto Cannella: co-funding by the European Union - FESR or FSE, PON Research and Innovation 2014-2020 - DM 1062/2021.

Table 3. Recommendations for addressing bias in artificial intelligence (AI) for medical imaging

Stage of AI	Recommendation
Design	• Ensure that the project team represents a range of perspectives, including radiologists, clinicians, data scientists, engineers, and department administrators, preferably from different demographic backgrounds. • Encourage the entire team for transparency in detecting and reporting potential biases. • Scrutinize research questions to identify any inherent biases or inequalities and address them proactively in the study design. • Consider adhering to established reporting and methodological quality guidelines to ensure transparency and reproducibility.
Data	• Collect data from a wide range of sources to capture diverse patient populations. • Conduct in-depth exploratory data analysis to identify any potential systematic errors that may exist, informing subsequent modeling and mitigation strategies. • Standardize data to ensure consistency across datasets, with effective harmonization techniques. • Implement rigorous quality control measures to maintain the accuracy and reliability of labels and annotations, following established protocols and guidelines. • Continuously monitor data quality and update annotations as needed to reflect any changes or improvements.
Modeling and Evaluation	• Divide the dataset into training, validation, and test sets before any modeling begins, ensuring that each subset is representative of the overall population. • Select evaluation metrics that account for disparities in outcomes across different demographic groups, avoiding metrics that may mask underlying systematic errors. • Consider techniques such as fairness-aware machine learning algorithms and model interpretability methods to mitigate bias and enhance transparency. • Evaluate model fairness using a variety of methods to capture different aspects of bias. • Assess model performance separately for different demographic subgroups to identify any disparities in predictive accuracy or bias. • Continuously retrain and update models to account for evolving datasets and mitigate the perpetuation of historical biases.
Deployment	• Continuously monitor model performance in real-world settings, paying particular attention to disparities in outcomes among different demographic groups. • Conduct thorough evaluation of model performance after any updates or modifications to ensure that biases have not been inadvertently introduced or amplified. • Engage with regulatory bodies to ensure compliance with relevant standards and guidelines and seek periodic audits to validate the fairness and effectiveness of the deployed models. • Try to collect effective feedback from the end-users to identify potential biases or shortcomings in the deployed system and address them promptly.

References

1. Hammond MEH, Stehlik J, Drakos SG, Kfoury AG. Bias in medicine: lessons learned and mitigation strategies. *JACC Basic Transl Sci*. 2021;6(1):78-85. [CrossRef]
2. Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. *Radiology*. 1988;167(2):565-569. [CrossRef]
3. Busby LP, Courtier JL, Glastonbury CM. Bias in radiology: the how and why of misses and misinterpretations. *Radiographics*. 2018;38(1):236-247. [CrossRef]
4. Gopal DP, Chetty U, O'Donnell P, Gajria C, Blackadder-Weinstein J. Implicit bias in healthcare: clinical practice, research and decision making. *Future Healthc J*. 2021;8(1):40-48. [CrossRef]
5. Oren O, Gersh BJ, Bhatt DL. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *Lancet Digit Health*. 2020;2(9):e486-e488. [CrossRef]
6. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*. 2023;13(17):2760. [CrossRef]
7. Tang X. The role of artificial intelligence in medical imaging research. *BJR Open*. 2019;2(1):20190031. [CrossRef]
8. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pac Symp Biocomput*. 2021;26:232-243. [CrossRef]
9. 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]
10. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019;322(24):2377-2378. [CrossRef]
11. Vrudhula A, Kwan AC, Ouyang D, Cheng S. Machine learning and bias in medical imaging: opportunities and challenges. *Circ Cardiovasc Imaging*. 2024;17(2):e015495. [CrossRef]
12. Banerjee I, Bhattacharjee K, Burns JL, et al. "Shortcuts" causing bias in radiology artificial intelligence: causes, evaluation, and mitigation. *J Am Coll Radiol*. 2023;20(9):842-851. [CrossRef]
13. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. [CrossRef]
14. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. *Neural Comput*. 1992;4(1):1-58. [CrossRef]
15. Doroudi S, Rastegar SA. The bias-variance tradeoff in cognitive science. *Cogn Sci*. 2023;47(1):e13241. [CrossRef]
16. Doroudi S. The bias-variance tradeoff: how data science can inform educational debates. *AERA Open*. 2020;6(4):2332858420977208. [CrossRef]
17. Guan X, Burton H. Bias-variance tradeoff in machine learning: theoretical formulation and implications to structural engineering applications. *Structures*. 2022;46:17-30. [CrossRef]
18. Bouchard G. Bias-variance tradeoff in hybrid generative-discriminative models. In: *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*. Cincinnati, OH, USA. 2007:124-129. [CrossRef]
19. Rocks JW, Mehta P. Memorizing without overfitting: bias, variance, and interpolation in overparameterized models. *Phys Rev Res*. 2022;4(1):013201. [CrossRef]
20. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc Natl Acad Sci*. 2019;116(32):15849-15854. [CrossRef]
21. Gong J, Currano RM, Sirkin D, Yeung S, Holsinger FC. NICE: Four Human-Centered AI principles for bridging the AI-to-clinic translational gap. In: 2021. Accessed April 8, 2024. [CrossRef]
22. Koçak B, Cuocolo R, dos Santos DP, Stanzione A, Ugga L. Must-have qualities of clinical research on artificial intelligence and machine

learning. *Balkan Med J*. 2023;40(1):3-12. [CrossRef]

23. Chen P, Wu L, Wang L. AI fairness in data management and analytics: a review on challenges, methodologies and applications. *Appl Sci*. 2023;13(18):10258. [CrossRef]
24. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci U S A*. 2020;117(23):12592-12594. [CrossRef]
25. Jayakumar S, Sounderajah V, Normahani P, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. *NPJ Digit Med*. 2022;5(1):11. [CrossRef]
26. van der Pol CB, McInnes MDF, Salameh JP, et al. Impact of reference standard on CT, MRI, and contrast-enhanced US LI-RADS diagnosis of hepatocellular carcinoma: a meta-analysis. *Radiology*. 2022;303(3):544-545. [CrossRef]
27. Chen PC, Mermel CH, Liu Y. Evaluation of artificial intelligence on a reference standard based on subjective interpretation. *Lancet Digit Health*. 2021;3(11):e693-e695. [CrossRef]
28. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. *Med Image Anal*. 2020;63:101693. [CrossRef]
29. Stanzione A, Cuocolo R, Ugga L, et al. Oncologic imaging and radiomics: a walkthrough review of methodological challenges. *Cancers (Basel)*. 2022;14(19):4871. [CrossRef]
30. Ashmore R, Calinescu R, Paterson C. Assuring the machine learning lifecycle: desiderata, methods, and challenges. Published online May 10, 2019. [CrossRef]
31. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (N Y)*. 2023;4(9):100804. [CrossRef]
32. Müller D, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res Notes*. 2022;15(1):210. [CrossRef]
33. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med*. 2022;5(1):48. [CrossRef]
34. Kulkarni V, Gawali M, Kharat A. Key Technology considerations in developing and deploying machine learning models in clinical radiology practice. *JMIR Med Inform*. 2021;9(9):e28776. [CrossRef]
35. Malerbi FK, Nakayama LF, Gayle Dychiao R, et al. Digital education for the deployment of artificial intelligence in health care. *J Med Internet Res*. 2023;25(1):e43333. [CrossRef]
36. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38. [CrossRef]
37. 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]
38. Roland T, Böck C, Tschoellitsch T, et al. Domain shifts in machine learning based Covid-19 diagnosis from blood tests. *J Med Syst*. 2022;46(5):23. [CrossRef]
39. Drukker K, Chen W, Gichoya J, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *J Med Imaging(Bellingham)*. 2023;10(6):061104. [CrossRef]
40. Olteanu A, Castillo C, Diaz F, Kıcıman E. Social data: biases, methodological pitfalls, and ethical boundaries. *Front Big Data*. 2019;2:13. [CrossRef]
41. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn*. 2021;110(3):457-506. [CrossRef]
42. Khera R, Simon MA, Ross JS. Automation Bias and Assistive AI: Risk of Harm From AI-driven clinical decision support. *JAMA*. 2023;330(23):2255-2257. [CrossRef]
43. Bitterman DS, Aerts HJWL, Mak RH. Approaching autonomy in medical artificial intelligence. *Lancet Digit Health*. 2020;2(9):e447-e449. [CrossRef]
44. 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]
45. 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]
46. Cadario R, Longoni C, Morewedge CK. Understanding, explaining, and utilizing medical artificial intelligence. *Nat Hum Behav*. 2021;5(12):1636-1642. [CrossRef]
47. Rouzrokh P, Khosravi B, Faghani S, et al. Mitigating bias in radiology machine learning: 1. data handling. *Radiol Artif Intell*. 2022;4(5):e210290. [CrossRef]
48. Gichoya JW, Thomas K, Celi LA, et al. AI pitfalls and what not to do: mitigating bias in AI. *Br J Radiol*. 2023;96(1150):20230023. [CrossRef]
49. Murphy-Hill E, Jaspan CN, Egelman C, Cheng L. The Pushback effects of race, ethnicity, gender, and age in code review. *Commun ACM*. 2022;65(3):52-57. [CrossRef]
50. A Akinci D'Antonoli T, Cuocolo R, Baessler B, Pinto Dos Santos D. Towards reproducible radiomics research: introduction of a database for radiomics studies. *Eur Radiol*. 2024;34(1):436-443. [CrossRef]
51. Zhang K, Khosravi B, Vahdati S, et al. Mitigating bias in radiology machine learning: 2. model development. *Radiol Artif Intell*. 2022;4(5):e220010. [CrossRef]
52. McDuff D, Cheng R, Kapoor A. Identifying bias in AI using simulation. Published online September 30, 2018. [CrossRef]
53. 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]
54. 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]
55. Sengupta E, Garg D, Choudhury T, Aggarwal A. Techniques to elimenate human bias in machine learning. In: *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*. Moradabad, India: 2018:226-230. [CrossRef]
56. 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]
57. Chen Z. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanit Soc Sci Commun*. 2023;10(1):1-12. [CrossRef]
58. 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]
59. Koçak B, Keleş A, Köse F. Meta-research on reporting guidelines for artificial intelligence: are authors and reviewers encouraged enough in radiology, nuclear medicine, and medical imaging journals? *Diagn Interv Radiol. 2024*. [CrossRef]
60. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med*. 2019;170(1):51-58. [CrossRef]
61. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008. [CrossRef]
62. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011;155(8):529-536. [CrossRef]
63. Lekadir K, Osuala R, Gallin C, et al. FUTURE-AI: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. Published online October 30, 2023. [CrossRef]
64. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. [CrossRef]
65. Kocak B, Akinci D'Antonoli T, Mercaldo N, et al. METhodological RadiomICs Score (METRICS):

a quality scoring tool for radiomics research endorsed by EuSoMII. *Insights Imaging*. 2024;15(1):8. [CrossRef]

66. 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]
67. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1(9):389-399. [CrossRef]
68. Gallegos IO, Rossi RA, Barrow J, et al. Bias and Fairness in large language models: a survey. Published online 2023. [CrossRef]
69. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on bias and fairness in machine learning. *ACM Comput Surv*. 2022;54(6):1-35. [CrossRef]
70. Suresh H, Guttag J. A framework for understanding sources of harm throughout the machine learning life cycle. In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM; 2021:1-9. [CrossRef]
71. Carrillo-Perez F, Pizurica M, Zheng Y, et al. Generation of synthetic whole-slide image tiles of tumours from RNA-sequencing data via cascaded diffusion models. *Nat Biomed Eng*. 2024. [CrossRef]
72. Chambon P, Bluethgen C, Delbrouck JB, et al. RoentGen: Vision-language foundation model for chest X-ray generation. Published online 2022. [CrossRef]
73. Wiles O, Albuquerque I, Rebuffi SA, et al. Generative models improve fairness of medical classifiers under distribution shifts. Published online May 31, 2023. [CrossRef]
74. Hooker S, Moorosi N, Clark G, Bengio S, Denton E. Characterising bias in compressed models. Published online 2020. [CrossRef]
75. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM; 2019:220-229. [CrossRef]
76. Feldman T, Peake A. End-to-end bias mitigation: removing gender bias in deep learning. Published online June 20, 2021. [CrossRef]
77. Shrestha R, Kafle K, Kanan C. An investigation of critical issues in bias mitigation techniques. In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE; Waikoloa, HI, USA: 2022:2512-2523. [CrossRef]
78. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Association for Computing Machinery; 2015:259-268. [CrossRef]
79. Cui Y, Jia M, Lin TY, Song Y, Belongie S. Class-balanced loss based on effective number of samples. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: 2019:9260-9269. [CrossRef]
80. Zou Y, Yu Z, Kumar BVKV, Wang J. Domain adaptation for semantic segmentation via class-balanced self-training. Published online October 25, 2018. [CrossRef]
81. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: 2017:2999-3007. [CrossRef]
82. He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. Hong Kong, 2008:1322-1328. [CrossRef]
83. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357. [CrossRef]
84. Nabi R, Shpitser I. Fair inference on outcomes. Published online January 21, 2018. [CrossRef]
85. Delage E, Ye Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper Res*. 2010;58(3):595-612. [CrossRef]
86. Sagawa S, Koh PW, Hashimoto TB, Liang P. Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. Published online April 2, 2020. [CrossRef]
87. Cadene R, Dancette C, Ben-younes H, Cord M, Parikh D. RUBi: reducing unimodal biases in visual question answering. Published online March 23, 2020. [CrossRef]
88. Clark C, Yatskar M, Zettlemoyer L. Don't take the easy way out: ensemble based methods for avoiding known dataset biases. Published online September 9, 2019. [CrossRef]
89. He H, Zha S, Wang H. Unlearn dataset bias in natural language inference by fitting the residual. Published online November 24, 2019. [CrossRef]
90. Kim B, Kim H, Kim K, Kim S, Kim J. Learning not to learn: training deep neural networks with biased data. Published online April 15, 2019. [CrossRef]
91. Grand G, Belinkov Y. Adversarial regularization for visual question answering: strengths, shortcomings, and side effects. Published online June 20, 2019. Accessed April 8, 2024. [CrossRef]
92. Ramakrishnan S, Agrawal A, Lee S. Overcoming language priors in visual question answering with adversarial regularization. Published online November 8, 2018. [CrossRef]
93. Adeli E, Zhao Q, Pfefferbaum A, et al. Representation learning with statistical independence to mitigate bias. Published online October 8, 2019. Accessed April 8, 2024. [CrossRef]
94. Zhang BH, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. Published online January 22, 2018. [CrossRef]
95. Ponsiglione A, Cuocolo R. Radiology under siege? Adversarial attacks against deep learning algorithms. *Eur J Radiol*. 2023;169:111156. [CrossRef]
96. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. Published online March 27, 2020. [CrossRef]
97. Santinha J, Matos C, Figueiredo M, Papanikolaou N. Improving performance and generalizability in radiogenomics: a pilot study for prediction of IDH1/2 mutation status in gliomas with multicentric data. *J Med Imaging (Bellingham)*. 2021;8(3):031905. [CrossRef]
98. Nam J, Cha H, Ahn S, Lee J, Shin J. Learning from failure: training debiased classifier from biased classifier. Published online November 23, 2020. [CrossRef]
99. Sanh V, Wolf T, Belinkov Y, Rush AM. Learning from others' mistakes: avoiding dataset biases without modeling them. Published online December 2, 2020. [CrossRef]
100. Utama PA, Moosavi NS, Gurevych I. Towards Debiasing NLU Models from Unknown Biases. Published online October 13, 2020. [CrossRef]
101. Pezeshki M, Kaba SO, Bengio Y, Courville A, Precup D, Lajoie G. Gradient starvation: a learning proclivity in neural networks. Published online November 24, 2021. [CrossRef]
102. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Published online October 7, 2016. [CrossRef]
103. Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. On Fairness and calibration. In: *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017. Accessed April 8, 2024. [CrossRef]
104. Woodworth B, Gunasekar S, Ohannessian MI, Srebro N. Learning non-discriminatory predictors. Published online November 1, 2017. [CrossRef]
105. Kamiran F, Karim A, Zhang X. Decision theory for discrimination-aware classification. In: *2012 IEEE 12th International Conference on Data Mining*. 2012:924-929. [CrossRef]
106. Lohia PK, Ramamurthy KN, Bhide M, Saha D, Varshney KR, Puri R. Bias mitigation post-processing for individual and group fairness. Published online December 14, 2018. [CrossRef]
107. Borys K, Schmitt YA, Nauta M, et al. Explainable AI in medical imaging: an overview for clinical practitioners - Beyond saliency-based XAI approaches. *Eur J Radiol*. 2023;162:110786. [CrossRef]
108. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Published online November 28, 2013. [CrossRef]

109. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier. Published online August 9, 2016. [CrossRef]

110. Lundberg S, Lee SI. A unified approach to interpreting model predictions. Published online November 24, 2017. [CrossRef]

111. Fong R, Patrick M, Vedaldi A. Understanding deep networks via extremal perturbations and smooth masks. Published online October 18, 2019. [CrossRef]

112. Dabkowski P, Gal Y. Real time image saliency for black box classifiers. Published online May 22, 2017. [CrossRef]

113. Fong RC, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017:3449-3457. [CrossRef]

114. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. Published online April 19, 2014. [CrossRef]

115. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Published online December 13, 2015. [CrossRef]

116. Wang H, Wang Z, Du M, et al. Score-CAM: score-weighted visual explanations for convolutional neural networks. Published online April 13, 2020. [CrossRef]

117. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: improved visual explanations for deep convolutional networks. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018:839-847. [CrossRef]

118. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*. 2020;128(2):336-359. [CrossRef]

119. van Leeuwen KG, Hedderich DM, Harvey H, Schalekamp S. How AI should be used in radiology: assessing ambiguity and completeness of intended use statements of commercial AI products. *Insights Imaging*. 2024;15(1):51. [CrossRef]

120. Tejani AS, Ng YS, Xi Y, Rayan JC. Understanding and Mitigating bias in imaging artificial intelligence. *Radiographics*. 2024;44(5):e230067. [CrossRef]

121. Flory MN, Napel S, Tsai EB. Artificial intelligence in radiology: opportunities and challenges. *Semin Ultrasound CT MRI*. Published online February 2024:S0887217124000052. [CrossRef]

122. Bell LC, Shimron E. Sharing Data Is Essential for the Future of AI in Medical Imaging. *Radiol Artif Intell*. 2024;6(1):e230337. [CrossRef]

123. Tripathi S, Gabriel K, Dheer S, et al. Understanding Biases and Disparities in Radiology AI Datasets: a review. *J Am Coll Radiol*. 2023;20(9):836-841. [CrossRef]

124. Ferrara E. Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. *Sci*. 2023;6(1):3. [CrossRef]

125. Venkatesh K, Santomartino SM, Sulam J, Yi PH. Code and data sharing practices in the radiology artificial intelligence literature: a meta-research study. *Radiol Artif Intell*. 2022;4(5):e220081. [CrossRef]

126. 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]

127. Ethics and Governance of Artificial Intelligence for Health: WHO guidance. Published online 2021. [CrossRef]

128. SyRI legislation in breach of European Convention on Human Rights. Accessed April 11, 2024. [CrossRef]

129. Ploug T, Holm S. The right to a second opinion on artificial intelligence diagnosis—remedying the inadequacy of a risk-based regulation. *Bioethics*. 2023;37(3):303-311. [CrossRef]

130. Pruski M. AI-enhanced healthcare: not a new paradigm for informed consent. *J Bioeth Inq*. 2024. [CrossRef]

131. Duffourc MN, Gerke S. The proposed EU Directives for AI liability leave worrying gaps likely to impact medical AI. *NPJ Digit Med*. 2023;6(1):77. [CrossRef]

132. Chen Y, Clayton EW, Novak LL, Anders S, Malin B. Human-centered design to address biases in artificial intelligence. *J Med Internet Res*. 2023;25(1):e43251. [CrossRef]

133. van Assen M, Beecy A, Gershon G, Newsome J, Trivedi H, Gichoya J. Implications of bias in artificial intelligence: considerations for cardiovascular imaging. *Curr Atheroscler Rep*. 2024;26(4)91-102. [CrossRef]

134. Ferrara C, Sellitto G, Ferrucci F, Palomba F, De Lucia A. Fairness-aware machine learning engineering: how far are we? Empir Softw Eng? *Empir Softw Eng*. 2023;29(1):9. [CrossRef]

135. Palatnik de Sousa I, Vellasco MMBR, Costa da Silva E. Explainable artificial intell. *Sensors (Basel)*. 2021;21(16):5657. [CrossRef]

136. Theunissen M, Browning J. Putting explainable AI in context: institutional explanations for medical. *Ethics Inf Technol*. 2022;24(2):23. [CrossRef]

137. Alikhademi K, Richardson B, Drobina E, Gilbert JE. Can explainable AI explain unfairness? A framework for evaluating explainable AI. Published online June 14, 2021. [CrossRef]

138. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med*. 2020;3:118. [CrossRef]

139. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. *J Am Med Inform Assoc*. 2020;27(12):2020-2023. [CrossRef]

140. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385(3):283-286. [CrossRef]

141. Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. *Lancet Digit Health*. 2021;3(6):e337-e338. [CrossRef]

142. Gasser U. An EU landmark for AI governance. *Science*. 2023;380(6651):1203. [CrossRef]

143. Mazzini G, Bagni F. Considerations on the regulation of AI systems in the financial sector by the AI Act. *Front Artif Intell*. 2023;6:1277544. [CrossRef]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242830

ARTIFICIAL INTELLIGENCE AND INFORMATICS

INVITED REVIEW

Artificial intelligence in musculoskeletal applications: a primer for radiologists

Michelle W. Tong[1,2,3]
Jiamin Zhou[4]
Zehra Akkaya[1,5]
Sharmila Majumdar[1,2]
Rupsa Bhattacharjee[1]

[1]University of California San Francisco, Department of Radiology and Biomedical Imaging, San Francisco, USA

[2]University of California San Francisco, Department of Bioengineering, San Francisco, USA

[3]University of California Berkeley, Department of Bioengineering, Berkeley, USA

[4]University of California San Francisco, Department of Orthopaedic Surgery, San Francisco, USA

[5]Ankara University Faculty of Medicine, Department of Radiology, Ankara, Türkiye

Corresponding author: Rupsa Bhattacharjee

E-mail: rupsa.bhattacharjee1@gmail.com

Received 05 May 2024; revision requested 03 June 2024; accepted 11 July 2024.

Epub: 19.08.2024

Publication date: 03.03.2025

DOI: 10.4274/dir.2024.242830

ABSTRACT

As an umbrella term, artificial intelligence (AI) covers machine learning and deep learning. This review aimed to elaborate on these terms to act as a primer for radiologists to learn more about the algorithms commonly used in musculoskeletal radiology. It also aimed to familiarize them with the common practices and issues in the use of AI in this domain.

KEYWORDS

Artificial intelligence, deep learning, machine learning, musculoskeletal, review

Approximately 1.71 billion people have musculoskeletal (MSK) conditions worldwide.[1] The need for imaging on MSK disorders is increasing in parallel with the rising and progressively aging global population,[2] posing a significant threat of fatigue in radiologists and unmet needs for patients.[3,4] The evolution of MSK radiology traces back to the inception of the field of radiology itself with the discovery of X-rays in 1895. On a separate trajectory, the 1950s witnessed the introduction of the first programming languages and software, raised by Turing's[5] question, "can machines think?". However, it was not until 1992, nearly a century later, that these two fields merged, culminating in the first research into artificial intelligence (AI) in radiology.[6] Today, AI has become an ever-growing field and is reshaping the world, including medicine, with radiology at the forefront, evidenced by Food and Drug Administration (FDA)-approved AI-based tools. The first AI-based algorithm was approved by the FDA in 2017. By 2022, radiology dominated the medical field by a striking 87% of all FDA-authorized AI-based devices.[7] In 2017, MSK applications were the second most common subject of AI-related publications in radiology, second only to neuroradiology.[8]

Thus far, AI research in radiology has primarily focused on interpretive tasks, including fracture detection, osteoarthritis detection and grading (cartilage and meniscal lesions), bone age determination, osteoporosis and bone quality assessment, tissue/region identification and segmentation, radiographic angle and bone measurements, clinical decision making on various bone and ligament anomalies, lesions characterization and diagnosis of infectious, oncological or rheumatological diseases, quantitative analysis and radiomics, and estimation of patient demographics.[9] However, AI also offers promising solutions for non-interpretive tasks, which aim to ensure high-quality care and time-efficient outputs for the rising demands on imaging.[10,11] Indeed, non-interpretive tasks, such as protocoling, quality control, and overseeing imaging studies, comprise 44% of a radiologist's daily workload.[12] However, most of these tasks are neglected where productivity is mainly assessed by the number of produced reports. Research in the emergency radiology department shows that for every 1 minute spent on the phone by radiologists, the report turnaround time increases by approximately 4 minutes.[12] Therefore, it is imperative to create time-efficient solutions to meet the rising demand in the field, where AI offers revolutionary solutions.

Therefore, radiologists must embrace a comprehensive understanding of AI and machine learning (ML) to integrate these technologies into their practice effectively, as described in Figure 1. Proficiency in data interpretation and validation will ensure the accuracy and reli-

You may cite this article as: Tong MW, Zhou J, Akkaya Z, Majumdar S, Bhattacharjee R. Artificial intelligence in musculoskeletal applications: a primer for radiologists. *Diagn Interv Radiol.* 2025;31(2):89-101.

ability of AI algorithms beneficial for clinical practice. Understanding the underlying principles of ML models, such as neural networks and deep learning (DL) architectures, is essential for critical appraisal and informed decision-making. Radiologists must also grasp the limitations and potential biases inherent in AI systems, emphasizing the importance of human oversight in clinical decision-making. Furthermore, knowledge of regulatory frameworks and ethical considerations surrounding AI adoption in healthcare is imperative to navigate legal and ethical challenges.

Algorithms

Alongside advancements in computational power, computer algorithms, and data availability, AI has gained popularity as a rapidly developing tool that can transform industries. Broadly defined, AI refers to computer systems that can perform assigned tasks, such as learning, decision-making, and problem-solving, with satisfactory or better-than-expected performance within a given context. Subsets of AI include the following: artificial narrow intelligence, which can perform specific tasks well but cannot transfer knowledge; artificial general intelligence, which can transfer knowledge across systems or tasks; and artificial superintelligence, which functions beyond the capability of human beings and is currently mainly conceptualized.[13] Commonly used AI concepts and descriptions are listed in Table 1.

ML essentially entails all techniques that can be employed to train a machine to mimic human performance. In the current context, it refers[14] to the development of algorithms that predict discrete labels (classification), continuous quantities (regression), data subgroups (clustering), or important features (dimensionality reduction) based on previous experiences using probability, statistics, and linear algebra. Traditional ML algorithms include linear classifiers, logistic regression, decision trees, and nearest-neighbor searches. Each of these algorithms seeks to learn a mapping between input and output variables by defining decision boundaries between labeled data or clustering of the data.

DL refers to a subset of ML that utilizes neural networks to learn new high-level feature representations of data for computer vision tasks, such as object segmentation, classification, and detection, with high efficiency.[15] Neural networks are composed of multiple layers of interconnected nodes with internal weights modeled after biological neural systems. The network learns to perform tasks by iteratively performing complex, non-linear transforms, involving passing forward input data through the network to predict a desired output and then using the discrepancy between the predicted and expected output to update the internal weights of the nodes in the network to improve task performance.

Convolutional neural networks (CNNs) perform convolution operations over local regions using shared convolution weights such that networks achieve translational invariance (i.e., objects can be detected regardless of location). Additional pooling operations down-sample data representations, automatically extracting relevant spatial hierarchical features. Variational CNNs have modified the underlying network structure to improve versatility and effectiveness. The two-dimensional (2D) U-Net was a significant breakthrough for medical imaging tasks, particularly segmentation. In 2015, Ronneberger et al.[16] proposed a unique U-shaped architecture (Figure 2), which down-sampled and up-sampled input images of varying image modalities to predict regions of interest with "very good performance," even after training with a very limited amount of training data. Despite their successes, CNNs are prone to overfitting, meaning CNN-based models do not perform as well on new unseen data. They also suffer from a requirement for large amounts of data for training and a lack of in-

Main points

- Proficiency in data interpretation and validation will ensure the accuracy and reliability of artificial intelligence (AI) algorithms beneficial for radiologists in their clinical practice.
- Understanding the underlying principles of machine learning (ML) models, such as neural networks and deep learning architectures, is essential for critical appraisal and informed decision-making and has been covered in this article.
- This article also discusses the limitations and potential biases inherent in AI systems, emphasizing the importance of human oversight in clinical decision-making.
- Furthermore, knowledge of regulatory frameworks and ethical considerations surrounding AI adoption in healthcare is imperative to navigate legal and ethical challenges.
- Continuous learning and collaboration with data scientists and AI experts are essential for radiologists to harness the full potential of AI and ML in improving diagnostic accuracy, efficiency, and patient care while upholding professional standards and ethical principles.

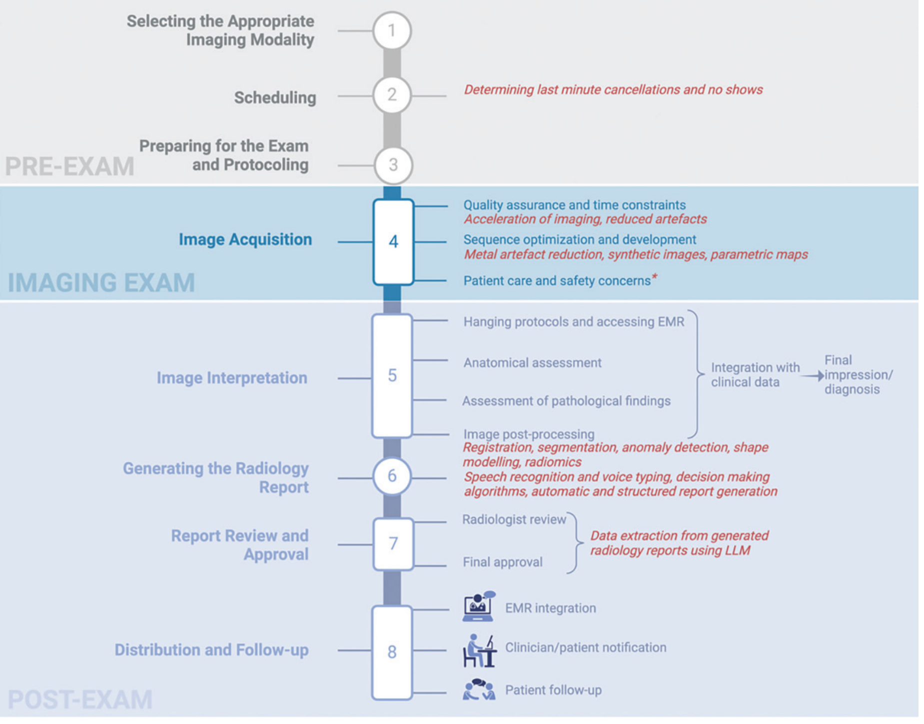


Figure 1. A schematic diagram of the usage of artificial intelligence (AI) in multiple levels of musculoskeletal radiology before, during, and after examination. *It is important to emphasize that continuous input from radiologists is crucial to minimize risks from AI in real-world clinical scenarios and to provide uncompromised patient safety at every step in the flowchart where AI-based solutions are being tested. The figure has been created with the help of the Biorender tool (https://www.biorender.com).

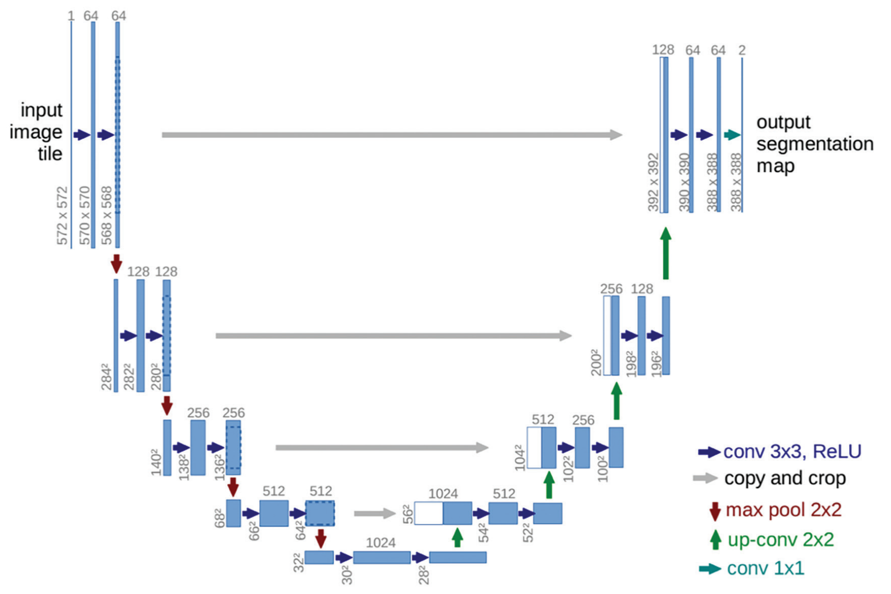


Figure 2: An introduction to the seminal U-Net architecture. Reproduced via Creative Commons License from.[16]

Table 1. A list of commonly used artificial intelligence concepts and descriptions	
Concepts	Meanings in one line
Artificial intelligence (AI)	The simulation of human intelligence processes by machines, particularly computer systems.
Machine learning (ML)	A subset of AI that allows systems to learn from data and improve over time without being explicitly programmed.
Deep learning	A subset of ML where artificial neural networks mimic the structure and function of the human brain to process data.
Neural networks	A system of algorithms modeled after the human brain, used to recognize patterns.
Natural language processing	The ability of computers to understand, interpret, and generate human language.
Computer vision	The field of AI that enables computers to interpret and understand visual information from the real world.
Reinforcement learning	A type of ML where an agent learns to make decisions by trial and error, receiving feedback in the form of rewards or penalties.
Supervised learning	A type of ML where the model is trained on labeled data, with input–output pairs provided.
Unsupervised learning	A type of ML where the model is trained on unlabeled data and must find patterns and relationships on its own.
Semi-supervised learning	A hybrid approach where the model is trained on a small amount of labeled data and a large amount of unlabeled data.
Transfer learning	An ML technique where a model trained on one task is repurposed or fine-tuned for a similar task.
Generative adversarial networks	A class of algorithms used in unsupervised learning to generate new data instances similar to a given dataset.
Overfitting	When a model learns to perform well on the training data but fails to generalize to new, unseen data.
Bias and variance	Bias refers to the error introduced by approximating a real-world problem with a simplified model, while variance refers to the error introduced by sensitivity to fluctuations in the training set.
Feature engineering	The process of selecting and transforming variables or features to improve the performance of ML algorithms.
Hyperparameters	Parameters that are set prior to training and control the learning process of ML algorithms.
Ensemble learning	A technique that combines multiple models to improve the performance of the overall system.

terpretability due to the model's architectural complexity.

Federated learning proposes a framework to address challenges with model generalizability, with special benefits when using medical data. An aggerate model encapsulates shared model weights from multiple collaborators who trained the model on private datasets.[17]

Generative adversarial networks (GANs) are popular for image-to-image translation, consisting of two opposing networks: a generator and a discriminator.[18] The generator creates an image to fool the discriminator, while the discriminator attempts to discern real or synthetic images.[19] Due to the oppositional nature of the network, GANs can be challenging to train and often require careful consideration of hyperparameters. Mode collapse occurs when the generator produces similar images that may not capture the full distribution of the training data and the discriminator is unable to provide useful feedback to guide training.

Recently, large language models and vision transformers (ViTs)[20] have spurred a new wave of innovation. Both of these DL architectures are based on transformers, which consist of an encoder, which extracts meaningful features from input data, and a decoder network, which uses the features to generate outputs. Transformers process data as a sequence of tokens, enabling the model to capture global relationships between the data (Figure 3). For ViTs, images are vectorized into tokens, which can be combined with text.[21]

A typical workflow to develop an ML algorithm involves several distinct stages. It begins with problem definition and data collection where a specific object is identified, and relevant data is gathered. Subsequently, data preprocessing involves cleaning, transforming, and processing the dataset for training. Common preprocessing techniques include image normalization and clipping to achieve favorable image intensity ranges and contrast for ML models. Before model development, data is split into training, validation, and testing subsets, often with balanced distributions of relevant metadata, such as age, for proper evaluation of model performance. During training, models may be prone to overfitting if highly sensitive to patterns in the training dataset. The validation dataset allows for the evaluation of model performance during training, while the test set is only used to assess performance on the final selected model for unbiased assessment.

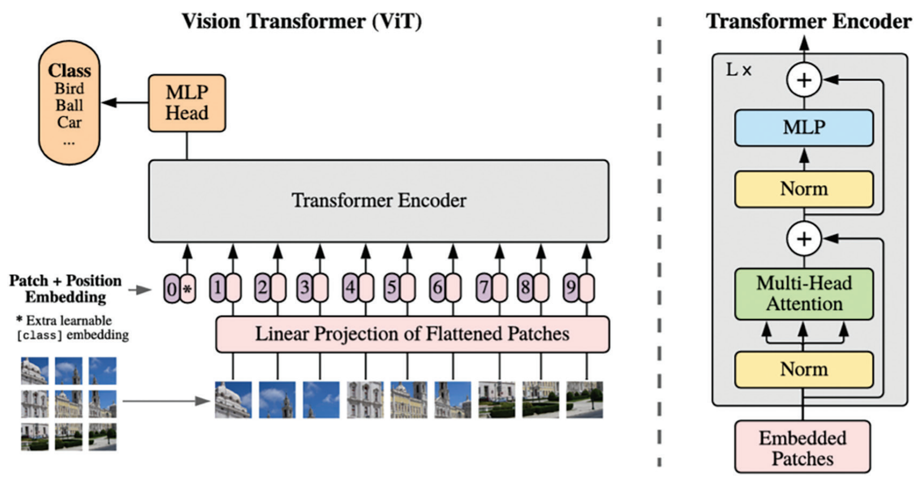


Figure 3. An introduction to vision transformers. Reproduced via Creative Commons License from.[20]

Next, model selection and training occur, where various algorithms are evaluated, and a suitable model is chosen. Existing models may offer excellent zero-shot capabilities such that no modification of model weights is needed. On the other hand, models may be trained for a specific use case by fine-tuning, which involves further training of a pre-trained model on a smaller, targeted data set. After training, the model is evaluated on the test dataset using appropriate metrics specific to the objects of the model. Finally, the model is deployed and undergoes monitoring and maintenance to ensure optimal performance over time. This iterative process requires collaboration between domain experts, data scientists, and computer programmers to achieve successful outcomes. Some of the crucial technical terms and metrics used in everyday ML, and what they mean, are listed in Table 2. Although AI seems to be an omnipresent tool in current radiology practices, many users remain unfamiliar with the basic concepts, utilities, challenges, processes, and biases associated with it. We aim to provide comprehensive starting content that prepares the community of medical experts to become tuned to the vocabulary and its nuances and to get a sense of how AI can be integrated into their daily MSK radiology practice.

Applications in musculoskeletal radiology

Image acquisition

Imaging acceleration

Extensive research dedicated to reducing the time required to acquire medical images has led to the development of unique data sampling and reconstruction techniques in MSK radiology, primarily for computed tomography (CT) and magnetic resonance imaging (MRI). In particular, MRI is an important modality for radiologists to diagnose many MSK conditions, but it suffers from increased cost and increased time to acquire images compared with other modalities. AI-based image acceleration techniques aim to break those Nyquist limits, though this must be done while considering any in-domain and domain-shift artifacts. Reconstruction, therefore, is equally essential to ensure the quality of images is clinically preserved in rapidly acquired MRI. AI researchers have developed algorithms that achieve both high accelerations for faster imaging and excellent reconstitution with comparable or improved image resolution. Such methodologies have been developed using data-driven guidance, such as compressed sensing or dictionary learning, or physics-guided networks combined with artifact removal.[22] These techniques are often modified for solution-specific problems, including accelerating higher-dimensional 2D or 3D MRI scans, such as dynamic (temporal) MRI.[23] AI techniques for the joint optimization of a non-Cartesian *k*-space sampling trajectory and an image-reconstruction network have been rising in popularity. For example, one such framework, PROJECTOR,[24] proposed dubbed projection for jointly learning non-Cartesian trajectories while optimizing reconstructor trajectories. It also ensured that the learned trajectories were compatible with gradient-related hardware constraints. Previous techniques enforced these constraints via penalty terms, but PROJECTOR enforces them via embedded steps that project the learned trajectory on a feasible set.

Synthesis of images and parametric maps

Another exciting application of AI is to characterize meaningful tissue maps or images from raw data (Figures 4 and 5). Wu et al.[25] proposed CNNs for synthesizing water/fat images from only two echoes instead of multiple. The method achieved high-fidelity output images, a 10-fold acceleration in computation time, and also generalizability to unseen organ images and metal artifacts. Zou et al.[26] have also proposed reconstructing free-breathing cardiac MRI data and synthesizing cardiac cine movies from manifold learning networks. This enables a unique generation of synthetic breath-hold cine movies with data on demand: specifically, movies with different inversion contrasts. Additionally, it enables the estimation of T_1 maps with specific respiratory phases. So far, the derivation of tissue parameter maps has been achieved by repeating acquisition in steady-state conditions and longer scan times.[22] However, rapid extraction of such parameters is no longer a challenge due to AI-based solutions, such as synthetic mapping of T_1, T_{1p}, R_2*, and T_2 relaxation, chemical exchange saturation transfer proton volume fraction and exchange rate, magnetization transfer, and susceptibility. Conventional magnetic resonance fingerprinting (MRF) is regularly used for quantitative parameter estimation. However, it suffers from the computational burden of dictionary generation and pattern matching. The burden further grows exponentially with the number of fitting parameters considered. ML has also been utilized to accelerate both the acquisition and reconstruction and thus optimize MRF sequences.[22]

End-to-end design

End-to-end design of reconstruction and segmentation techniques have recently been a heavy focus in the medical imaging community. Often addressed separately, these two tasks could benefit from being handled in tandem. Tolpadi et al.[27] recently hosted and summarized a challenge entitled "K2S," hosted at the 25th International Conference on Medical Image Computing and Computer-Assisted Intervention (Singapore, 2022). Eight-times under-sampled raw MRI measurements were provided as training data with their fully sampled counterparts and segmentation masks (i.e., a unique dataset consisting of 300 knee MRI scans accompanied by radiologist-approved tissue segmentation labels). In the testing phase, the challenge participants submitted DL models that generated high-efficiency segmenta-

Table 2. Technical terms and metrics used in everyday machine learning: what do they mean?

Technical terms	
Terms	Explanations
Feature	An individual measurable property or characteristic of a phenomenon being observed, often represented as a variable in a dataset.
Label	The output or target variable in supervised learning, representing the prediction or classification to be made.
Instance	A single example or data point in a dataset, typically represented as a row in a table.
Model	A mathematical representation or algorithm that learns patterns and relationships from data to make predictions or decisions.
Training data	The data used to train a machine learning (ML) model, consisting of input features and corresponding labels.
Test data	Data used to evaluate the performance of a trained ML model, separate from the training data.
Validation data	Data used to fine-tune hyperparameters and assess model performance during the training process.
Loss function	A function that measures the difference between predicted and actual values, used to train and optimize ML models.
Optimization algorithm	An algorithm used to adjust the parameters of a model during training to minimize the loss function.
Gradient descent	An optimization algorithm that iteratively updates the parameters of a model by moving in the direction of steepest descent of the loss function.
Epoch	One complete pass through the entire training dataset during the training of an ML model.
Batch	A subset of the training data used in one iteration of training, typically chosen to improve efficiency.
Batch size	The number of training examples utilized in one iteration of training during the gradient descent process.
Learning rate	A hyperparameter that controls the step size during the optimization process, determining the rate at which the model parameters are updated.
Stop criteria	Criteria by which model stop training, such as for "x" number of epochs or until the loss stops decreasing by "x" %. Clear stop criteria and assessment of training loss allow a fairer comparison of model weights.
Regularization	Techniques used to prevent overfitting by adding a penalty term to the loss function, discouraging complex models.
Dropout	A regularization technique used in neural networks to randomly deactivate neurons during training to prevent overfitting.
Activation function	A mathematical function applied to the output of each neuron in a neural network, determining its output.
Backpropagation	An algorithm used to train neural networks by iteratively adjusting the weights of connections based on the error calculated during forward pass.
Convolutional neural network	A type of neural network designed for processing structured grids of data, commonly used in image recognition tasks.
Recurrent neural network (RNN)	A type of neural network designed to process sequences of data, with connections between units forming directed cycles, commonly used in natural language processing tasks.
Long short-term memory	A type of RNN unit capable of learning long-term dependencies, commonly used in sequence prediction tasks.
Common metrics	
Accuracy	The proportion of correctly classified instances (both true positives and true negatives) out of the total instances.
Precision	The proportion of true positive predictions out of all positive predictions made by the model.
Recall (sensitivity)	The proportion of true positive predictions out of all actual positive instances in the dataset.
F1 Score	The harmonic mean of precision and recall, providing a balance between the two metrics.
Specificity	The proportion of true negative predictions out of all actual negative instances in the dataset.
ROC area under the curve score	The area under the receiver operating characteristic (ROC) curve, representing the model's ability to discriminate between positive and negative classes across different thresholds.
Confusion matrix	A table used to evaluate the performance of a classification model, showing the counts of true positive, true negative, false positive, and false negative predictions.
Mean squared error (MSE)	The average of the squared differences between predicted and actual values, commonly used for regression tasks.
Root mean squared error	The square root of the MSE, providing a measure of the average magnitude of error in the predicted values.
Mean absolute error	The average of the absolute differences between predicted and actual values, providing a measure of average error magnitude.
Peak signal-to-noise ratio	A measure of image quality and fidelity, calculated as the ratio between the maximum power of a signal verses noise. Commonly used for reconstruction tasks.
Structural similarity index metric	A measure of preceptive similarity between two images whose formula is based on comparison of image structure, contrast, and brightness. Commonly used for reconstruction tasks.
R-squared	A statistical measure of the proportion of variance in the dependent variable that is explained by the independent variables in a regression model.
Mean average precision	A metric used to evaluate object detection models, representing the average precision over all classes at various intersection over union thresholds.
Cohen's kappa	A statistic that measures inter-rater agreement for categorical items, considering how much agreement would be expected by chance.
Mean intersection over union	A metric commonly used to evaluate semantic segmentation models, measuring the ratio of intersection to union of predicted and ground truth masks. Values range from 0 to 1, indicating no to perfect overlap, respectively.
Dice coefficient	A metric for segmentation assessment calculated by the ratio of 2 × the intersection divided by the total area of predicted and ground truth masks. This metric has good utility for small regions of interest because there is no bias from background labels. Background is often more prevalent so inclusion of these labels leading to unfavorable class imbalance.
Log loss (binary cross-entropy)	A loss function used in binary classification tasks, measuring the difference between predicted probabilities and actual binary outcomes.
Silhouette score	A measure of how similar an object is to its own cluster compared with other clusters, used to assess the quality of clustering algorithms.
Explained variance score	A metric used to evaluate the performance of regression models, measuring the proportion of variance in the target variable explained by the model.

tion maps directly from the under-sampled raw data. No correlations were found between the reconstruction and segmentation metrics (Figure 6). Some researchers suggest pre-training segmentation models on "pretext tasks". In these tasks, the model is trained to restore distorted images. Context prediction and context restoration challenges demonstrate that segmentation models can be made robust with pre-training, particularly if labeled data availability is limited.[22]

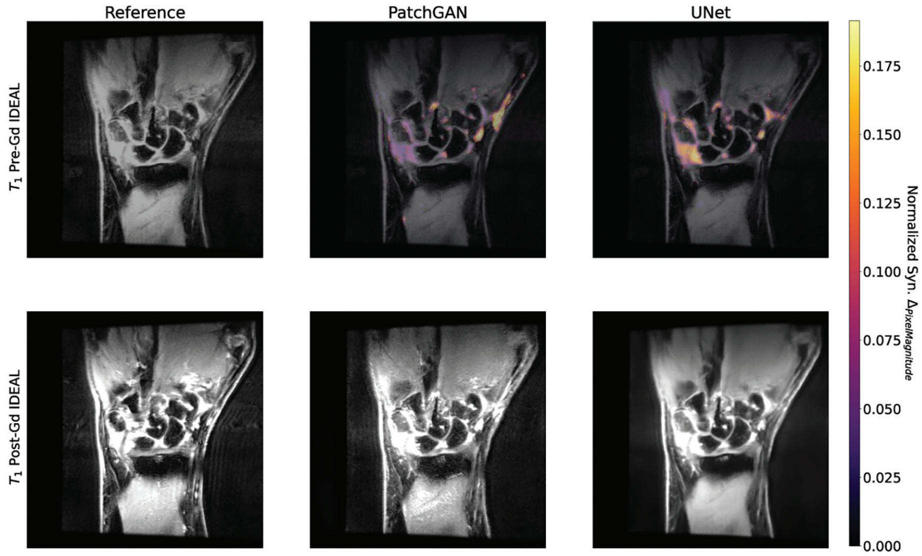


Figure 4. Occlusion maps for PatchGAN and U-Net pipelines. For U-Net and PatchGAN, hotspots primarily included intercarpal joint regions. Particularly for the U-Net, the maps also emphasized the forearm muscles. Given that the synovial joints are where an inflammatory imaging algorithm would see the most utility, the fact that both algorithms placed heavy emphasis on the intercarpal regions was promising, indicating that both focused on synovitis-relevant regions to make predictions. Reproduced via Creative Commons License from.[72]

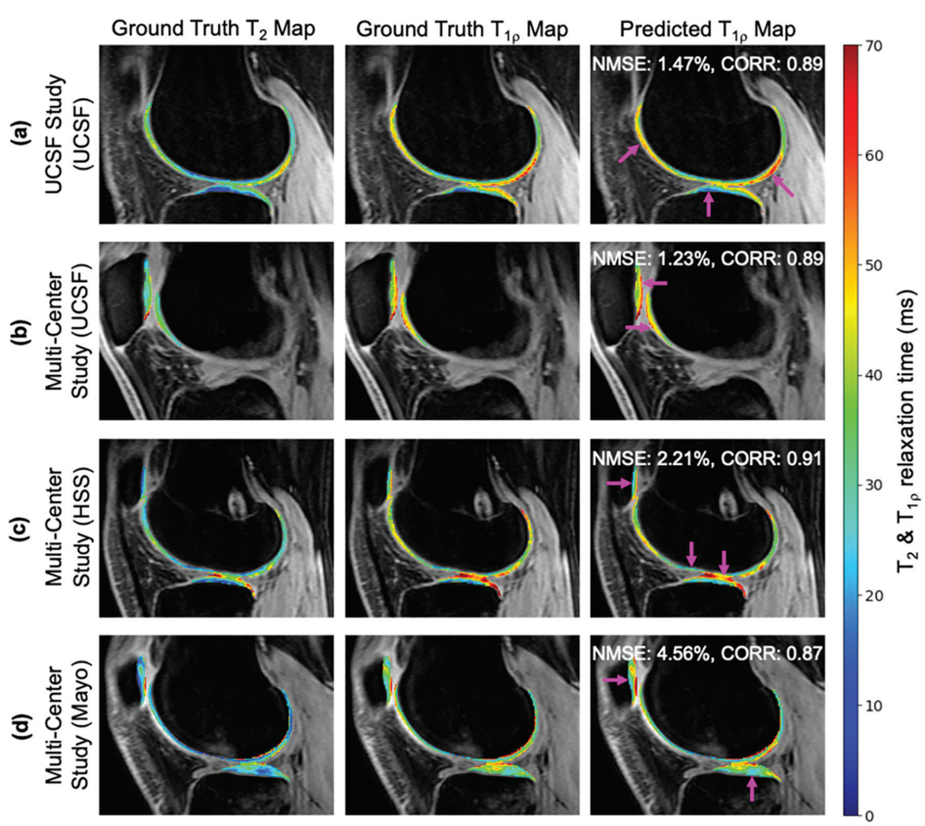


Figure 5. Four knees from patients who participated in one of two studies: **(a)** the UCSF (cohort A) study or **(b-d)** the multi-center (cohort B) study at one of three centers. Input ground truth T2 maps exhibit distinct intensity elevation and textural patterns compared with ground truth $T_{1\rho}$ maps. Nevertheless, predicted $T_{1\rho}$ maps generated by the convolutional neural network preserve these differences, as indicated by the regions marked by the arrows. Reproduced via Creative Commons License from.[74]

Image post-processing

Registration

Image registration is a critical process in imaging that focuses on the accurate alignment of images, which is necessary for the diagnosis, treatment planning, and monitoring of diseases. However, it is difficult to develop robust algorithms to register images of varying resolution and from different modalities efficiently and accurately. This is particularly challenging in the presence of significant anatomical variation in the case of MSK disease. Conventional registration methods often rely on solving pairwise optimization problems, which can be time-consuming and computationally expensive.[28] Recent literature has demonstrated the growing application of AI, in particular DL models, in image registration. CNNs, for instance, have been employed to predict the transformation required to align images. For example, a study by Sokooti et al.[29] proposed a CNN-based method for non-rigid registration on 3D chest CT follow-up data. Another novel approach involves using spatial transformer networks (STNs), a DL model that can learn spatial transformations to align images. In a study by Sokooti et al.[29] an STN was used for image registration, showing that the model could learn complex transformations from training data.[30] Models such as VoxelMorph, a CNN-based unsupervised framework for image registration,[31] have also shown promising results. Although VoxelMorph was trained on 3D brain MRI, the architecture of the models can be used to train on specific MSK datasets due to the unsupervised and generalizable nature of the models.

Segmentation

Image segmentation is a well-defined problem that involves the delineation of specific regions of interest. As manual image segmentation is both time-consuming and repetitive, the research community has explored AI to improve medical image segmentation workflows with great interest.[16] Over the years, various network architectures have been developed to segment MSK structures. One of the most popular CNN models is the U-Net, discussed earlier. It is often utilized to solve 2D or 3D segmentation tasks, such as identifying muscles, bones, cartilages, menisci, femoral and acetabular regions, and shoulder structures in knee, spine, hip, thigh, and wrist anatomy.[32,33] Usually, the performance of existing segmentation algorithms can only be fairly compared on a specific case basis, such as anatomical region, medical imaging acquisition setting, or study population.[34]

DL can establish a useful representation of any object without prior super-imposition of user-designed features. This is why the performance of a vertebral body segmentation algorithm relies on the integrity of intervertebral discs and is compromised when disc pathologies are present if not trained with enough variety of data. Identification of a thoracic vertebral body is achieved using intrinsic features and its proximity to a disc. The disc serves as an extrinsic feature for the vertebral body. In other words, it becomes the landmark that the network learns in the context of spine segmentation (Figure 7a-c). This is also the reason for failures in patch-based approaches. Only limited contextual information is passed, which limits the outcome efficiency.

On the positive side, network learning from diverse data may often learn how the images, anatomies, and pathologies are integrated beyond visual perception, suggest new biomarkers as predictors of MSK diseases through image analysis, and potentially overcome the limitations of human perception.

Anomaly detection

Anomaly detection involves identifying abnormal structures or pathologies, such as fractures, tumors, or degenerative diseases, amidst a wide range of normal anatomical variations. To accurately distinguish between benign variants and clinically significant abnormalities, DL models-particularly CNNs-have been implemented due to their ability to learn hierarchical feature representations.[35,36] Autoencoders have also been used for unsupervised anomaly detection, whereby during the training process for reconstructing input data, they learn to encode "normal" data patterns and can thus highlight deviations from the norm when encountering an anomalous data point and produce a significantly different output.[37] These models can assist in identifying subtle or complex anomalies that may be missed by the human eye while providing consistent performance, thus reducing variability between different radiologists' interpretations. Workflow efficiency can be improved by prioritizing cases with potential anomalies identified by AI. However, there is a risk of generating false positives, false negatives, or model hallucinations, leading to unnecessary interventions or missed diagnoses. Radiologists should seek AI tools that balance sensitivity and specificity to minimize false positive and negative rates.

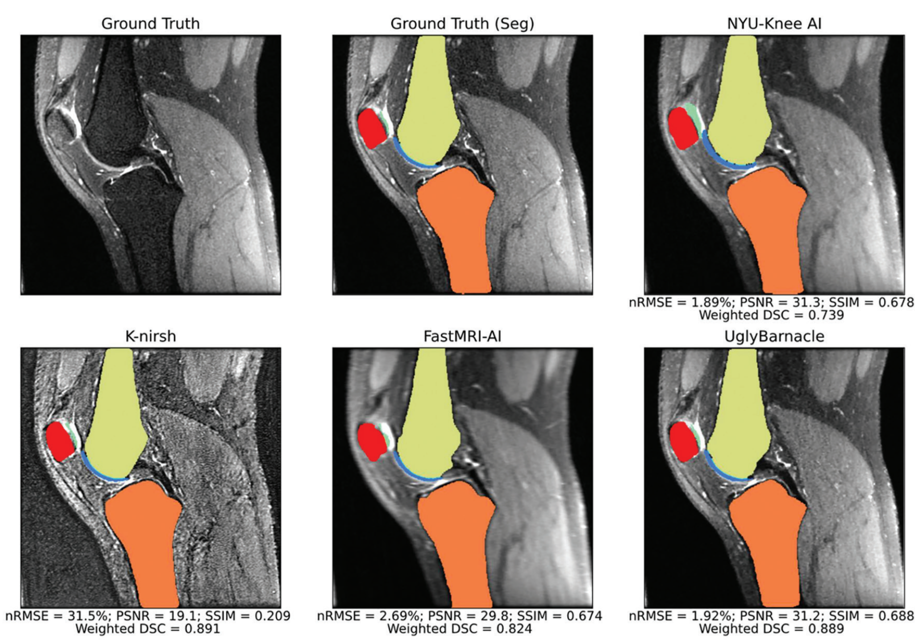


Figure 6. Miccai 2022 challenge results and submissions from the top teams. Sagittal slice segmentations are overlaid on intermediate pipeline reconstructions, displaying reconstruction and segmentation metrics for the segmented slice. Background anatomy slices were thus blurrier for some teams than for others, as different teams had different qualities of intermediate pipeline reconstruction outputs. In this example, segmentation quality was strong for all top submissions, with only some overestimation of cartilage thickness from the NYU knee artificial intelligence pipeline being apparent. K-nirsh maintains a slight edge over UglyBarnacle in reconstruction metrics for this volume. Reproduced via Creative Commons License from.[27]

Shape modeling

Shape modeling focuses on the accurate representation and analysis of the anatomical structures of the MSK system, with the challenge of capturing the complex geometry and variability of bones and soft tissues; this is essential for surgical pre-operative planning, prosthesis design, and the study of biomechanical properties. Active shape models and statistical shape modeling are common statistical methods to capture the variability of shape across a population and can be used for tasks such as segmentation.[38] However, they require a large amount of representative data for accurate modeling and can be sensitive to outliers with large shape deviations (Figure 8).

DL-based methods have been increasingly utilized for shape modeling due to their ability to learn complex, non-linear relationships. CNNs are commonly used due to their ability to process hierarchical features from image data directly. For instance, the U-Net architecture[16] and its variants have been extensively used for biomedical image segmentation tasks, providing detailed shape models of various anatomical structures. U-Net's strength lies in its symmetric expanding path, which allows precise localization, a key factor in accurate shape modeling. Another DL model, V-Net,[39] is a 3D variant of U-Net and is used for volumetric medical image segmentation, providing 3D shape models. Both U-Net and V-Net have shown competitive performance compared with traditional methods, with the added advantage of handling large datasets and capturing fine-grained details. DL models have recently been used for shape prediction and generation. For instance, GANs have been employed to generate realistic 3D shapes to synthesize anatomical structures for augmentation and analysis.[40] One hidden benefit of an AI-based shape model is the ability to predict changes in MSK structures over time, aiding in prognostic assessments.[35]

Radiomics

Radiomics, merging the word "radiology" with "-omics" to describe the high-throughput, data-driven approach to characterizing radiological images, involves computer-assisted image analysis where many quantitative "features" are extracted from images that are not readily appreciable to the human eye. Radiomic features have historically involved mathematical operations on the voxels of an image, converting morphological information about anatomical structure into quantitative values. Over time, the number of features has grown exponentially as more features have been identified, making the application of ML techniques, or classifiers, to identify radiomic features increasingly popular over the past few years.[41] Support vector machines, random forests, and neural networks have been used to identify and analyze features that are most predictive of disease presence, severity, progression, and response to treatment. CNNs are also increasingly being applied to automate feature extraction. However, the clinical utility of radiomics is still being established, and integration into clinical workflows remains a challenge.

Metal artifact reduction

AI, particularly DL algorithms, is increasingly applied to mitigate metal artifacts in MSK imaging. Metal implants or instruments introduce significant artifacts, particularly in MRI, which can impair diagnostic accuracy and limit the utility of these scans. Current literature points to the use of AI in CT and radiography, but its application in MRI is less explored.[42] In the context of MRI, the integration of AI for metal artifact reduction is still in its infancy. Existing techniques without the use of AI, such as multi-acquisition variable-resonance image combination and slice encoding for metal artifact correction (SEMAC), have limitations in their application and efficacy. Studies have used neural networks to accelerate SEMAC MRI while maintaining comparable metal artifact suppression,[43] as well as using unsupervised learning or attention maps from deep neural networks to guide correction.[44] However, most of these studies rely on phantom data or MRIs of other organs of interest. There is a need for more research and development, including robust validation studies, to explore the full potential of AI in MSK MRI specifically.

Report generation

Generating accurate and informative reports is a crucial task for radiologists to convey their findings and interpretations to the referring physician in a clear, concise, and clinically relevant manner. To reduce the reporting burden on radiologists, natural language processing (NLP) techniques, such as recurrent neural networks, long short-term memory networks, and more recently, transformer-based models, such as bidirectional encoder representations from transformers and generative pre-trained transformer, can be utilized for generating radiological reports. These are trained on a large body of annotated radiological reports to learn the language and structure of report writing, as well as the relationships between imaging findings and clinical diagnoses. An additional speech recognition step can also add to the automation of the report generation process,[45] creating a text output that can be considered a "preliminary report." As radiology reports traditionally lack standardized structure and content, NLP can then be used for the extraction of meaningful or contextual information[46] from the preliminary radiology report, whether traditional text or text from speech recognition. Applications range from the extraction of specific MSK data or follow-up recommendations[47] to the generation of a final report of classification, diagnostic criteria, disease probability, or follow-up recommendations. However, AI may not capture the subtleties of human

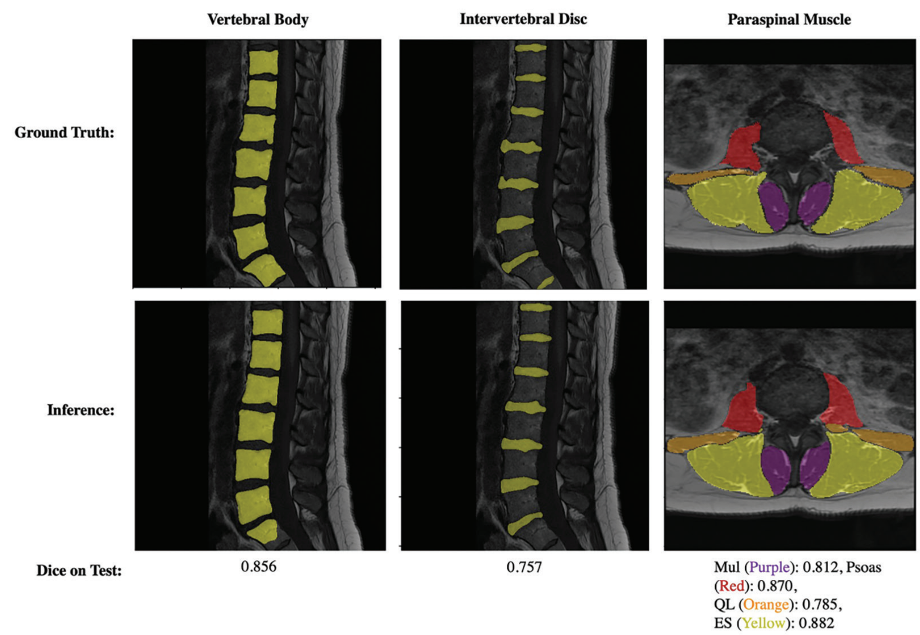


Figure 7. (a) Visualization of segmentation results from each network. The first, second, and third columns show examples of the vertebral body, intervertebral disc, and paraspinal muscle segmentation results, respectively, along with a three-dimensional Dice coefficient of each network's performance. The Dice coefficient measures the similarity between segmentation masks, where 1 indicates perfect overlap and 0 indicates no overlap. Reproduced via Creative Commons License from.[32]

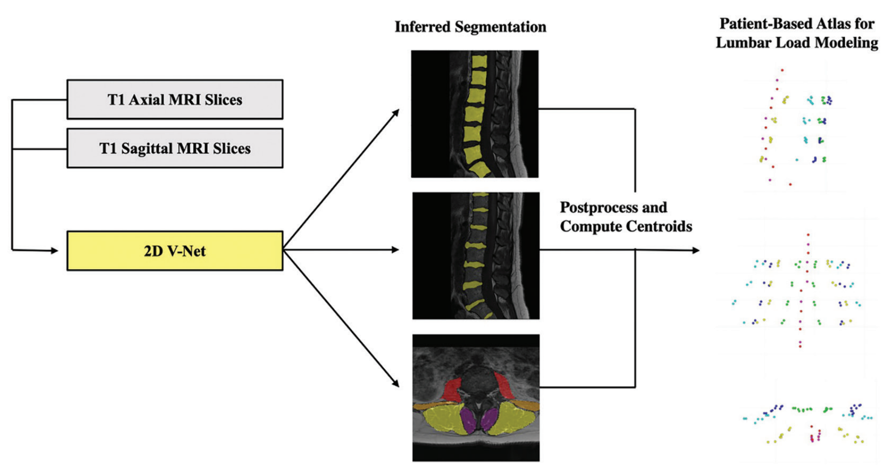


Figure 7. (b) Visualization of centroid construction. T_1 axial and T_1 sagittal MRI slices were input into their respective V-Net to generate inferred segmentation masks of the vertebral bodies, intervertebral discs, and paraspinal muscles. After postprocessing, centers of mass were computed on each segmentation mask to calculate the position of volume-wise centroids for each vertebral body and intervertebral disc and slice-wise centroids for each paraspinal muscle. These centroids were then converted to patient-based space, yielding a three-dimensional atlas of the lumbar spine for further biomechanical modeling. Reproduced via Creative Commons License from.[32] MRI, magnetic resonance imaging.

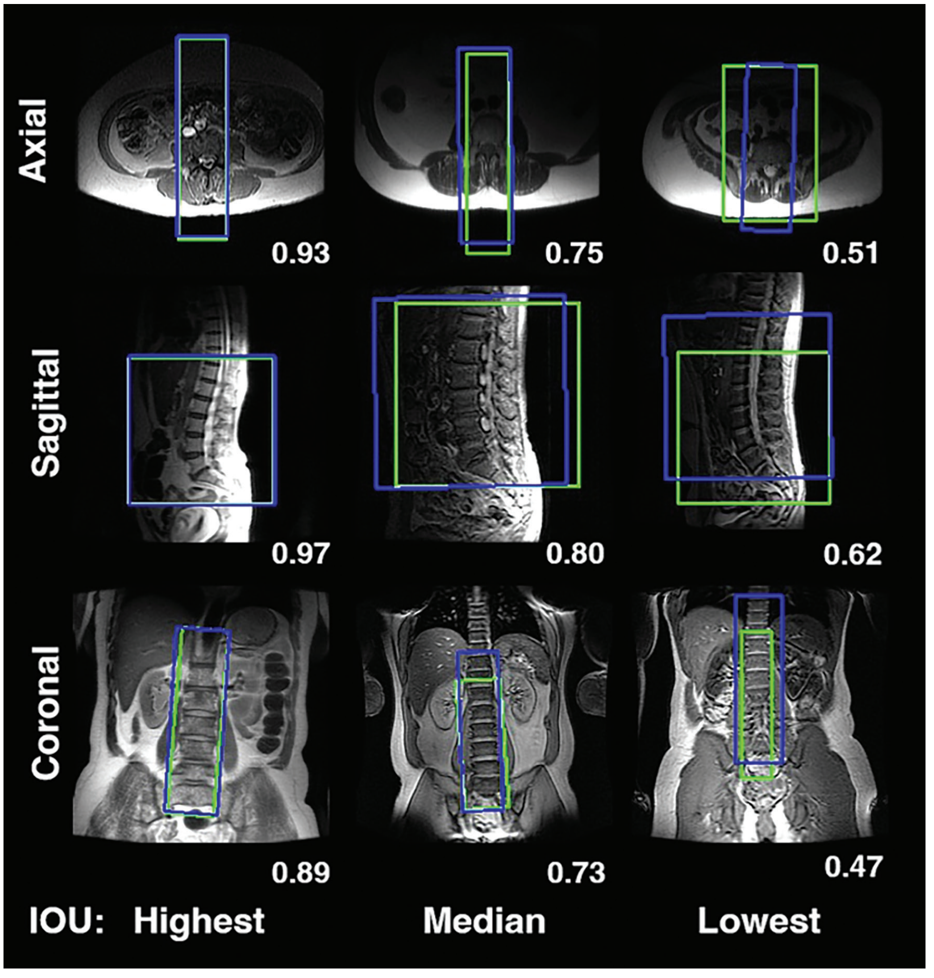


Figure 7. (c) Visualization of centroid construction. T_1 axial and T_1 sagittal magnetic resonance imaging slices were input into their respective V-Net to generate inferred segmentation masks of the vertebral bodies, intervertebral discs, and paraspinal muscles. After postprocessing, centers of mass were computed on each segmentation mask to calculate the position of volume-wise centroids for each vertebral body and intervertebral disc and slice-wise centroids for each paraspinous muscle. These centroids were then converted to patient-based space, yielding a three-dimensional atlas of the lumbar spine for further biomechanical modeling. Reproduced via Creative Commons License from.[71]

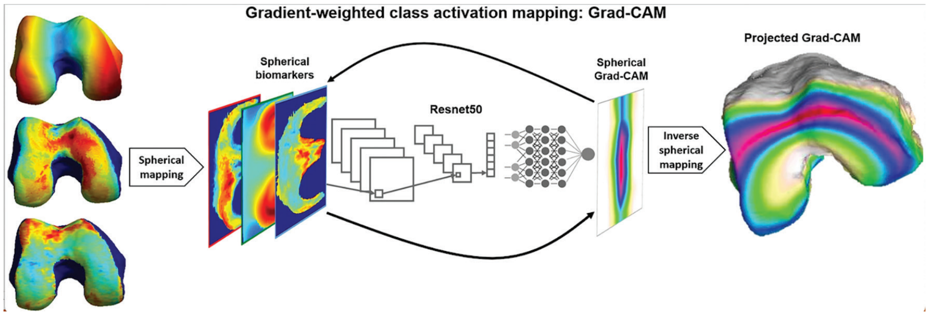


Figure 8. The authors used the Grad-CAM model interpretation technique to obtain a class discriminative localization map for each prediction. They first computed the gradient of the class of interest (before the "softmax" function) regarding feature maps of the last convolutional layer in the Resnet. These gradients flowing back were globally average-pooled to obtain the neuronal importance weights for the target class. A heat map of location importance was then up-sampled to match the image size and overlaid on the input image. The authors then leveraged the invertible property of their spherical transformation method to generate articular surface importance heat maps for model interpretation for each bone and each single biomarker. This process was performed on the first timepoint of every unique patient in the hold-out test set (n = 875) and is illustrated for the femur. Reproduced via Creative Commons License from.[73]

language, leading to reports that lack the nuanced communication often necessary between radiologists and referring physicians. Radiologists should view AI in report generation as a complementary tool that can assist with the reporting process but not as a replacement for the expert interpretation provided by a trained radiologist.

Considerations

Challenges defining ground truth data, benchmarks, and radiologists' availabilities

To achieve the highest yield from AI technologies, it is imperative to have large and reliable ground truth datasets for training, validation, and testing. Ideally, these should be from several different sources and representative of diverse communities accessible by non-radiologists, such as AI researchers, engineers, and data scientists.[48] The recent increase in the availability of such publicly available medical image banks and large-scale international AI challenges have catalyzed progress in the field, leading to the development of AI algorithms capable of handling different tasks, such as classification, detection, or segmentation, in different modalities.[49-51] The ground truth required for the current supervised AI models requires a labor- and time-intensive curation process for ideal workflow and to ensure the generalizability of a model. Moreover, this process is subject to regulatory constraints, commercial and operational pressures, as well as epistemic differences and limits of labeling.[52,53] Annotated images and their respective radiology reports are available in hospital databases but due to ethical reasons are not readily available to developers. It is important to follow the regulatory procedures and obtain approval from responsible committees to ensure an ethical approach when accessing and sharing this data between developers.[52]

Radiologists rely on visual detection, pattern recognition, memory, and cognitive reasoning to consolidate a final interpretation while making decisions.[4] Radiologists' errors have a vast impact on medical errors, which constitute the third most common cause of death in the USA, following cancer and heart disease.[54,55] The error rate is approximately 4% in clinical radiology practices, which translates into 40 million errors out of 1 billion worldwide radiographs annually.[4] Of particular importance, the distinction between an "error" and "observation variation" is highly relevant when creating such datasets. Imaging findings alone, without clinical information, are frequently not enough to definitively indicate a specific diagnosis. Consequently, interpreting radiologic studies is typically not a straightforward binary process of discriminating normal from pathologic entities. Professional acceptability lies on an arbitrary scale, between an obvious error and the unavoidable difference of opinion in interpretation.[56] This is particularly of concern given that most clinical AI applications are developed using data generated by "expert radiologists." Thus, these models are subjected to many kinds of human errors and biases and it falls on us humans to be cognizant of inequality, data availability, and privacy, ethical and medicolegal concerns with these rapidly evolving technologies.[57,58]

The top five most influential radiology societies from the USA, Canada, Europe, Australia, and New Zealand recently released a joint statement on potential practical and ethical concerns in deploying and integrating AI in radiology practices. The key take-home statements, which also apply specifically to MSK radiology, include a strong recommendation for rigorous monitoring of its uses and safety in clinical practice, close collaboration between developers, end-users, and regulators, and strict adherence to all the regulatory steps from the development to deployment and integration in the clinical workflow.[59] Radiologists in particular should be aware of automation bias as a potential source of error when working with AI tools in decision making.[60]

Model deployment

Deploying and maintaining AI models requires a robust infrastructure that addresses computational needs for both initial deployments using off-the-shelf pre-trained models and more advanced adaptations through fine-tuning. Most radiologists and clinical departments start with off-the-shelf pre-trained AI models. These models are developed on large, general datasets and can be used directly for common imaging tasks with minimal setup and without extensive customization. Standard computing hardware, including central processing units or modest graphics processing units (GPUs), can be used to run these models, making them accessible to most clinical environments.

Fine-tuning is necessary when adapting a pre-trained model to specific datasets or unique clinical scenarios in MSK radiology. This involves modifying the pre-trained model's parameters to better fit the particular characteristics of the new data, such as custom protocols for rare conditions, integrating specific patient demographics, or adapting models to unique imaging modalities or contrasts, improving the performance and relevance of the model. From a computational perspective, fine-tuning is less resource-intensive than training a model from scratch, as the model has already learned useful features from the initial large-scale dataset. This can be particularly beneficial in medical imaging, where annotated datasets are often limited and expensive to acquire. For instance, a model initially trained on a large dataset of general MRI images can be fine-tuned on a smaller dataset of specific MSK conditions. Studies using this approach have been reviewed by Cheplygina et al.[61], demonstrating improved performance on the tasks of interest. However, higher computational resources than those used for deployment are still needed for the fine-tuning process to handle the training workload. High-performance GPUs or tensor processing units are resources that can accelerate the processing of large datasets and complex model architectures during the training phase of fine-tuning. Cloud-based solutions with an environment that is secure and compliant with the Health Insurance Portability and Accountability Act also offer scalable resources that can be dynamically adjusted based on the computational load, making them ideal for training and deploying models without the need for local high-performance hardware.

Successful deployment of AI tools requires seamless integration into clinical workflows, which may involve Digital Imaging and Communications in Medicine (DICOM) standards and interoperability with various Picture Archiving and Communication System software, supported by robust infrastructure capable of handling ongoing model monitoring and updates to ensure sustained performance over time, adjust for any data shifts or incorporate new data, and maintain model relevance and performance.

Equitable medical artificial intelligence

The development and deployment of AI technologies in MSK radiology must be prioritized for fairness and justice. Algorithms should aim to mitigate biases, ensure accessibility to all demographic groups, and deliver personalized care tailored to individual needs, irrespective of socio-economic status or background. Doo and McGinty[62] argue that bias in radiology AI stems from various stages of model design encompassing the selection of training data, algorithm development, deployment, and performance assessment. These biases, in turn, have repercussions on patient care and health outcomes. Notably, there is a lack of standardized protocols for demographic labeling in AI. Existing datasets often blur distinctions between crucial identifiers, such as sex and gender, or oversimplify complex racial categories, leading to distorted outcomes and predictive inaccuracies. Consequently, AI models trained on such biased datasets tend to reinforce preexisting biases, contributing to unintended consequences.

When contemplating advanced healthcare imaging within the AI landscape, a fundamental query arises: Is it possible to completely anonymize (deidentifying without any possibility of reidentification) data?[63] At first glance, the task appears simple: selectively erase or encode identifiers within the metadata headers of images. Despite the widespread use of the DICOM standard for radiologic data, an increasing number of exceptions complicate efforts to establish standardized procedures. Recently, the progress in facial recognition technology has raised concerns about the potential for matching images from CT or MRI scans with individuals' photographs. Consequently, it has become standard practice in medical imaging research to alter images using defacing or skull-stripping algorithms to eliminate facial features. Unfortunately, such alterations can undermine the generalizability of ML models developed using such data.[64] The topic is extremely complicated in terms of types of biases and there are several remedies, which are almost impossible to comprehensively cover in the scope of the article. However, it is important to introduce the concepts of bias and equitable medical AI in MSK radiology and something to be conscious of while utilizing the AI tools.[64] Some of the most common issues with MSK imaging in AI and potential solutions to those are listed in Table 3.

Conclusion: current trends and future directions

Integration of AI with other emerging technologies, such as augmented reality and virtual reality is enabling more immersive and interactive visualization of medical images. New tools may facilitate better surgical planning, training, and intraoperative guidance. Additionally, AI-assisted tools have a niche role in aiding radiologists who are training and provide an avenue for additional diagnostic opinion where multiple

Table 3. Common problems and potential solutions

Common problems	Potential solutions
Data quality issues	- Preprocessing techniques, such as denoising and image enhancement. - Augmentation methods to increase dataset diversity and robustness.
Limited annotated data	- Semi-supervised or weakly supervised learning approaches. - Active learning strategies to prioritize data labeling efforts. - Transfer learning from pre-trained models on larger datasets.
Class imbalance	- Data resampling techniques, such as oversampling or undersampling. - Class-weighted loss functions to penalize errors on minority classes. - Synthetic data generation to balance class distribution.
Interpretability and explainability	- Model visualization techniques, such as saliency maps and activation maximization. - Explainable artificial intelligence methods, such as local interpretable model-agnostic explanations or Shapley additive explanations. - Incorporating attention mechanisms to highlight important image regions.
Overfitting and generalization	- Regularization techniques, such as dropout and weight decay. - Cross-validation and validation set monitoring to detect overfitting. - Domain adaptation methods to improve model robustness across different datasets.
Computational resource constraints	- Model compression techniques, such as pruning and quantization. - Distributed training frameworks for parallel processing. - Cloud-based solutions for scalable compute resources.
Ethical and legal considerations	- Adherence to data protection regulations, such as the Health Insurance Portability and Accountability Act. - Bias detection and mitigation strategies during model development. - Transparent reporting of model performance and limitations.

radiologists reading images is not feasible. Protocolling, which involves choosing the right imaging protocol to obtain the most diagnostic images for each patient, is supervised by a radiologist and is particularly important in MSK MRI applications where imaging protocols frequently require patient-specific tailoring. The limited number of research reports, using CNN and natural language classifier-based algorithms, have demonstrated encouraging outcomes.[65-67] Nevertheless, it is important to acknowledge the diversity of MSK imaging protocols for a wide spectrum of clinical scenarios, where these tools should be fine-tuned and advanced by taking medical history, prior imaging studies, scanner-specific data, contrast information, and radiation exposure dose into account.[68] AI can also offer dual working solutions for scheduling, by reducing both MRI times and waiting times by identifying no-shows or canceled appointments ahead of time.[69] Finally, radiology reports are the final product of radiologists and are the means of communication of findings between physicians. ML can help generate decision-making algorithms as a support system based on the available information on the patient's medical background.[68,70] Conversely, ML-based NLP can be a powerful tool to harness data from radiology reports and is currently being investigated.[9]

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. World Health Organization. Musculoskeletal health. [Crossref]
2. Harkey P, Duszak R Jr, Gyftopoulos S, Rosenkrantz AB. Who refers musculoskeletal extremity imaging examinations to radiologists? *AJR Am J Roentgenol.* 2018;210:834-841. [Crossref]
3. Reiner BI, Krupinski E. The insidious problem of fatigue in medical imaging practice. *J Digit Imaging.* 2012;25:3-6. [Crossref]
4. Waite S, Scott J, Gale B, Fuchs T, Kolla S, Reede D. Interpretive error in radiology. *Am J Roentgenol.* 2017;208:739-749. [Crossref]
5. Turing AM. Computing machinery and intelligence. *Mind New Ser.* 1950;59:433-460. [Crossref]
6. Driver CN, Bowles BS, Bartholmai BJ, Greenberg-Worisek AJ. Artificial intelligence in radiology: a call for thoughtful application. *Clin Transl Sci.* 2020;13:216-218. [Crossref]
7. Health C for D and R: Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. FDA 2023. [Crossref]
8. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur Radiol Exp.* 2018;2:35. [Crossref]
9. Debs P, Fayad LM. The promise and limitations of artificial intelligence in musculoskeletal imaging. *Front Radiol.* 2023;3:1242902. [Crossref]
10. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* 2018;18:500-510. [Crossref]
11. Richardson ML, Garwood ER, Lee Y, et al. Noninterpretive uses of artificial intelligence in radiology. *Acad Radiol.* 2021;28:1225-1235. [Crossref]
12. Glover M, Almeida RR, Schaefer PW, Lev MH, Mehan WA Jr. Quantifying the impact of noninterpretive tasks on radiology report turn-around times. *J Am Coll Radiol.* 2017;14:1498-1503. [Crossref]
13. Kelly S, Kaye SA, Oviedo-Trespalacios O. What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telemat Inform.* 2023;77:101925. [Crossref]
14. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics.* 1970;12:55-67. [Crossref]
15. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521:436-444. [Crossref]
16. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. 2015;234-241. [Crossref]
17. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep.* 2020;10:12598. [Crossref]
18. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In *Adv Neural Inf Process Syst*. 2014;27:1-9. [Crossref]
19. Nie D, Trullo R, Lian J, et al. Medical image synthesis with deep convolutional adversarial networks. *IEEE Trans Biomed Eng.* 2018;65:2720-2730. Erratum in: *IEEE Trans Biomed Eng*. 2020;67(9):2706. [Crossref]
20. Dosovitskiy A, Beyer L, Kolesnikov A, et al. an image is worth 16x16 words: transformers for image recognition at scale. 2021. [Crossref]

21. Vision Transformer and Language Model Based Radiology Report Generation | IEEE Journals & Magazine | IEEE Xplore. [Crossref]

22. Shimron E, Perlman O. AI in MRI: computational frameworks for a faster, optimized, and automated imaging workflow. *Bioengineering (Basel).* 2023;10:492. [Crossref]

23. Oscanoa JA, Middione MJ, Alkan C, et al. Deep Learning-based reconstruction for cardiac MRI: a review. *Bioengineering (Basel).* 2023;10:334. [Crossref]

24. Improving spreading projection algorithm for rapid k-space sampling trajectories through minimized off-resonance effects and gridding of low frequencies - Giliyar Radhakrishna - 2023 - Magnetic Resonance in Medicine - Wiley Online Library. [Crossref]

25. Wu Y, Alley M, Li Z, et al. Deep learning-based water-fat separation from dual-echo chemical shift-encoded imaging. *Bioengineering (Basel).* 2022;9:579. [Crossref]

26. Zou Q, Priya S, Nagpal P, Jacob M. Joint cardiac T1 mapping and cardiac cine using manifold modeling. *Bioengineering (Basel).* 2023;10:345. [Crossref]

27. Tolpadi AA, Bharadwaj U, Gao KT, et al. K2S Challenge: From Undersampled K-space to automatic segmentation. *Bioengineering (Basel).* 2023;10:267. [Crossref]

28. Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: a survey. *IEEE Trans Med Imaging.* 2013;32:1153-1190. [Crossref]

29. Sokooti H, de Vos B, Berendsen F, et al. 3D convolutional neural networks image registration based on efficient supervised learning from artificial deformations. 2019. [Crossref]

30. Lee MCH, Oktay O, Schuh A, Schaap M, Glocker B. Image-and-spatial transformer networks for structure-guided image registration. 2019. [Crossref]

31. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Trans Med Imaging.* 2019;38:1788-1800. [Crossref]

32. Hess M, Allaire B, Gao KT, et al. Deep learning for multi-tissue segmentation and fully automatic personalized biomechanical models from BACPAC clinical lumbar spine MRI. *Pain Med.* 2023;24(Suppl 1):139-148. [Crossref]

33. Caliva F, Iriondo C, Martinez AM, Majumdar S, Pedoia V. Distance Map Loss Penalty Term for Semantic Segmentation. 2019. [Crossref]

34. Bonaldi L, Pretto A, Pirri C, Uccheddu F, Fontanella CG, Stecco C. Deep learning-based medical images segmentation of musculoskeletal anatomical structures: a survey of bottlenecks and strategies. *Bioengineering (Basel).* 2023;10:137. [Crossref]

35. Hirvasniemi J, Runhaar J, van der Heijden RA, et al. The KNee OsteoArthritis Prediction (KNOAP2020) challenge: an image analysis challenge to predict incident symptomatic radiographic knee osteoarthritis from MRI and X-ray images. Osteoarthritis Cartilage. 2023;31:115-125. [Crossref]

36. Kijowski R, Liu F, Caliva F, Pedoia V. Deep learning for lesion detection, progression, and prediction of musculoskeletal disease. *J Magn Reson Imaging.* 2020;52:1607-1619. [Crossref]

37. Jamaludin A, Lootus M, Kadir T, et al. ISSLS prize in bioengineering science 2017: automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. *Eur Spine J.* 2017;26:1374-1383. [Crossref]

38. Heimann T, Meinzer HP. Statistical shape models for 3D medical image segmentation: a review. *Med Image Anal.* 2009;13:543-563. [Crossref]

39. Milletari F, Navab N, Ahmadi SA. V-Net: Fully Convolutional neural networks for volumetric medical image segmentation. 2016. [Crossref]

40. Wu J, Zhang C, Xue T, Freeman WT, Tenenbaum JB. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. 2017. [Crossref]

41. Fritz B, Yi PH, Kijowski R, Fritz J. Radiomics and deep learning for disease detection in musculoskeletal radiology: an overview of novel MRI- and CT-based approaches. *Invest Radiol.* 2023;58:3-13. [Crossref]

42. Selles M, Wellenberg RHH, Slotman DJ, et al. Image quality and metal artifact reduction in total hip arthroplasty CT: deep learning-based algorithm versus virtual monoenergetic imaging and orthopedic metal artifact reduction. *Eur Radiol Exp.* 2024;8:31. [Crossref]

43. Seo S, Do WJ, Luu HM, Kim KH, Choi SH, Park SH. Artificial neural network for slice encoding for metal artifact correction (SEMAC) MRI. *Magn Reson Med.* 2020;84:263-276. [Crossref]

44. Kim JW, Kwon K, Kim B, Park H. Attention Guided Metal Artifact Correction in MRI using deep neural networks. 2019. [Crossref]

45. Hannun A, Case C, Casper J, et al. Deep speech: scaling up end-to-end speech recognition. 2014. [Crossref]

46. Do BH, Wu AS, Maley J, Biswal S. Automatic retrieval of bone fracture knowledge using natural language processing. *J Digit Imaging.* 2013;26:709-713. [Crossref]

47. Carrodeguas E, Lacson R, Swanson W, Khorasani R. Use of machine learning to identify follow-up recommendations in radiology reports. *J Am Coll Radiol.* 2019;16:336-343. [Crossref]

48. Liew C. The future of radiology augmented with artificial intelligence: a strategy for success. *Eur J Radiol.* 2018;102:152-156. [Crossref]

49. Syeda-Mahmood T. Role of big data and machine learning in diagnostic decision support in radiology. *J Am Coll Radiol.* 2018;15:569-576. [Crossref]

50. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60-88. [Crossref]

51. Lassau N, Bousaid I, Chouzenoux E, et al. Three artificial intelligence data challenges based on CT and MRI. *Diagn Interv Imaging.* 2020;101:783-788. [Crossref]

52. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. *Radiology.* 2020;295:4-15. [Crossref]

53. Zając HD, Avlona NR, Kensing F, Andersen TO, Shklovski I. Ground Truth Or Dare: factors affecting the creation of medical datasets for training AI. In *Proc 2023 AAAIACM Conf AI Ethics Soc*. 2023:351-362. [Crossref]

54. Makary MA, Daniel M. Medical error-the third leading cause of death in the US. *BMJ.* 2016;353:i2139. [Crossref]

55. Sabih DE, Sabih A, Sabih Q, Khan AN. Image perception and interpretation of abnormalities; can we believe our eyes? Can we do something about it? *Insights Imaging.* 2011;2:47-55. [Crossref]

56. Brady AP, Neri E. Artificial intelligence in radiology-ethical considerations. *Diagnostics (Basel).* 2020;10:231. [Crossref]

57. Link TM, Pedoia V. Using AI to improve radiographic fracture detection. *Radiology.* 2022;302:637-638. [Crossref]

58. 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;30:80-90. [Crossref]

59. 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. *Can Assoc Radiol J.* 2024;75:226-244. [Crossref]

60. 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;293:436-440. [Crossref]

61. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med Image Anal.* 2019;54:280-296. [Crossref]

62. Doo FX, McGinty GB. Building Diversity, Equity, and Inclusion within radiology artificial intelligence: representation matters, from data to the workforce. *J Am Coll Radiol.* 2023;20:852-856. [Crossref]

63. Lotan E, Tschider C, Sodickson DK, et al. Medical imaging and privacy in the era of artificial intelligence: myth, fallacy, and the

future. *J Am Coll Radiol*. 2020;17:1159-1162. [Crossref]

64. 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]
65. Lee YH. Efficiency improvement in a busy radiology practice: determination of musculoskeletal magnetic resonance imaging protocol using deep-learning convolutional neural networks. *J Digit Imaging*. 2018;31:604-610. [Crossref]
66. Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH. Automatic determination of the need for intravenous contrast in musculoskeletal MRI examinations using IBM Watson's Natural Language Processing Algorithm. *J Digit Imaging*. 2018;31:245-251. [Crossref]
67. Richardson ML. MR protocol optimization with deep learning: a proof of concept. *Curr Probl Diagn Radiol*. 2021;50:168-174. [Crossref]
68. Gyftopoulos S, Lin D, Knoll F, Doshi AM, Rodrigues TC, Recht MP. Artificial intelligence in musculoskeletal imaging: current status and future directions. *Am J Roentgenol*. 2019;213:506-513. [Crossref]
69. Kurasawa H, Hayashi K, Fujino A, et al. Machine-learning-based prediction of a missed scheduled clinical appointment by patients with diabetes. *J Diabetes Sci Technol*. 2016;10:730-736. [Crossref]
70. Gorelik N, Gyftopoulos S. Applications of artificial intelligence in musculoskeletal imaging: from the request to the report. *Can Assoc Radiol J*. 2021;72:45-59. [Crossref]
71. Ozhinsky E, Liu F, Pedoia V, Majumdar S. Machine learning-based automated scan prescription of lumbar spine MRI acquisitions. *Magn Reson Imaging*. 2024;110:29-34. [Crossref]
72. Tolpadi AA, Luitjens J, Gassert FG, et al. Synthetic inflammation imaging with PatchGAN deep learning networks. *Bioenginering (Basel)*. 2023;10:516. [Crossref]
73. Morales AG, Lee JJ, Caliva F, et al. Uncovering associations between data-driven learned qMRI biomarkers and chronic pain. *Sci Rep*. 2021;11:21989. [Crossref]
74. Bioengineering | Free Full-Text | Synthetic Knee MRI T1p Maps as an Avenue for Clinical Translation of Quantitative Osteoarthritis Biomarkers. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242835

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Artificial intelligence system for identification of overlooked lung metastasis in abdominopelvic computed tomography scans of patients with malignancy

Hye Soo Cho[1]
Eui Jin Hwang[1,2]
Jaeyoun Yi[3]
Boorym Choi[3]
Chang Min Park[1,2]

[1]Seoul National University Hospital, Seoul National University College of Medicine, Department of Radiology, Seoul, Republic of Korea

[2]Seoul National University College of Medicine, Department of Radiology, Seoul, Republic of Korea

[3]Coreline Soft Inc., Seoul, Republic of Korea

PURPOSE

This study aimed to evaluate whether an artificial intelligence (AI) system can identify basal lung metastatic nodules examined using abdominopelvic computed tomography (CT) that were initially overlooked by radiologists.

METHODS

We retrospectively included abdominopelvic CT images with the following inclusion criteria: a) CT images from patients with solid organ malignancies between March 1 and March 31, 2019, in a single institution; and b) abdominal CT images interpreted as negative for basal lung metastases. Reference standards for diagnosis of lung metastases were confirmed by reviewing medical records and subsequent CT images. An AI system that could automatically detect lung nodules on CT images was applied retrospectively. A radiologist reviewed the AI detection results to classify them as lesions with the possibility of metastasis or clearly benign. The performance of the initial AI results and the radiologist's review of the AI results were evaluated using patient-level and lesion-level sensitivities, false-positive rates, and the number of false-positive lesions per patient.

RESULTS

A total of 878 patients (580 men; mean age, 63 years) were included, with overlooked basal lung metastases confirmed in 13 patients (1.5%). The AI exhibited an area under the receiver operating characteristic curve value of 0.911 for the identification of overlooked basal lung metastases. Patient- and lesion-level sensitivities of the AI system ranged from 69.2% to 92.3% and 46.2% to 92.3%, respectively. After a radiologist reviewed the AI results, the sensitivity remained unchanged. The false-positive rate and number of false-positive lesions per patient ranged from 5.8% to 27.6% and 0.1% to 0.5%, respectively. Radiologist reviews significantly reduced the false-positive rate (2.4%–12.6%; all *P* values < 0.001) and the number of false-positive lesions detected per patient (0.03–0.20, respectively).

CONCLUSION

The AI system could accurately identify basal lung metastases detected in abdominopelvic CT images that were overlooked by radiologists, suggesting its potential as a tool for radiologist interpretation.

CLINICAL SIGNIFICANCE

The AI system can identify missed basal lung lesions in abdominopelvic CT scans in patients with malignancy, providing feedback to radiologists, which can reduce the risk of missing basal lung metastasis.

KEYWORDS

Artificial intelligence, computed tomography, image analysis, metastasis, radiology

Corresponding author: Eui Jin Hwang

E-mail: ken921004@hotmail.com

Received 14 May 2024; revision requested 15 June 2024; accepted 01 August 2024.

Epub: 09.09.2024

Publication date: 03.03.2025

DOI: 10.4274/dir.2024.242835

You may cite this article as: Cho HS, Hwang EJ, Yi J, Choi B, Park CM. Artificial intelligence system for identification of overlooked lung metastasis in abdominopelvic computed tomography scans of patients with malignancy. *Diagn Interv Radiol.* 2025;31(2):102-110.

Abdominopelvic computed tomography (CT) is frequently performed in patients with cancer to evaluate various cancers of the abdominopelvic or extra-abdominopelvic organs. Lung metastasis frequently occurs in the advanced stages of various solid organ cancers, and abdominopelvic CT images inevitably capture the base of the lungs. Therefore, evaluating the presence of nodules suggestive of metastasis to the lung base is an important component in the interpretation of abdominopelvic CT scans in patients.[1-4] However, in a busy clinical environment, a radiologist may pay relatively less attention to the basal lungs compared with the abdominal organs, which are the main targets of evaluation.[4] Therefore, metastatic nodules in the basal lungs can be overlooked by interpreting radiologists, which may adversely affect a patient's treatment policy decisions or prognosis, leading to a medicolegal dispute.

The automatic detection of pulmonary nodules on chest CT images is one of the most widely investigated topics in artificial intelligence (AI)-based medical image analysis. Various studies have reported the radiologist-level performance of AI and the enhanced performance of radiologists using AI for lung nodule detection on CT scans.[5,6] Based on these impressive results, commercial AI-based software medical devices have begun to be utilized in daily clinical practice as computer-aided detection (CAD) tools.[7-10]

In addition to its use as a CAD tool, AI's utilization as a second reader–that is, for use in analyzing images after the radiologist's interpretation and providing feedback to the radiologist in case of suspected interpretation errors–can be another attractive scenario for applying AI in daily practice.[11-15] AI, as a second reader, can provide a safety net for radiologists against the risk of interpretation errors or medicolegal disputes without requiring the rigorous effort of scrutinizing the AI's results following every examination. The detection of pulmonary nodules in the basal lungs, as acquired using abdominopelvic CT, can serve as a compelling scenario for employing an AI second reader.[15] This is because it is beyond the primary focus of examination, yet carries a relatively high risk of interpretation errors, which could result in critical outcomes.

In consideration of the above, we aim to evaluate whether an AI system could detect metastatic pulmonary nodules in the basal lungs that have been overlooked by radiologists on abdominopelvic CT images.

Main points

- An artificial intelligence (AI) system for pulmonary nodule detection on computed tomography (CT) images can be utilized as a second reader after the radiologist's interpretation, to identify overlooked pulmonary nodules.
- As a second reader, the AI may analyze images after the radiologist's interpretation and provide feedback to the radiologist only when the AI suspects that the radiologist has overlooked a pulmonary nodule. In this scenario, the oversight of significant pulmonary nodules can be prevented without the need to review the AI results of all the examinations. In our study, the applied AI system could accurately identify basal lung metastases captured in abdominopelvic CT images that were overlooked by radiologists, suggesting its potential as a second reader after the radiologist's interpretation.
- We believe that our study contributes significantly to the literature by highlighting the effectiveness of AI in improving the accuracy of interpreting abdominopelvic CT images in patients with malignancies. Additionally, it underscores the importance of AI as a second reader to reduce interpretation errors.

Methods

This single-center, retrospective, diagnostic cohort study was approved by the Seoul National University Hospital Institutional Review Board on January 5, 2022 (approval number: 2112-142-1284). During the approved research period, patient data required for this study were accessed for research purposes. The requirement for informed consent was waived by the institutional review board.

Patients

Patients were consecutively included in a single tertiary referral institution in South Korea with the following criteria: a) patients diagnosed with solid organ cancers (International Statistical Classification of Diseases and Related Health Problems, 10th revision, C00 to C75); b) patients who underwent abdominopelvic CT between March 1 and March 31, 2019; and c) abdominopelvic CT scans interpreted as negative for basal lung metastasis in the formal reports of radiologists, based on a manual review of unstructured radiological reports by a thoracic radiologist. Patients who underwent chest CT on the same day as abdominopelvic CT and those lost to follow-up within 3 years without a clinical diagnosis of lung metastasis were excluded (Figure 1).

The first CT examination was performed on patients who underwent CT more than once. For multiphase CT examinations, images that captured the largest portion of the basal lungs were included in the analyses.

Diagnosis of pulmonary metastasis

To confirm the clinical diagnosis of pulmonary metastasis in patients, one thoracic radiologist (E.J.H., with 5 years of experience as a faculty thoracic radiologist) reviewed the medical records and CT images (including the index abdominopelvic CT and follow-up chest and abdominopelvic CT images). Pulmonary lesions that were pathologically confirmed as metastases, as well as lesions with persistent growth on follow-up CT images and a clinical impression of metastasis, were regarded as pulmonary metastases. Pulmo-

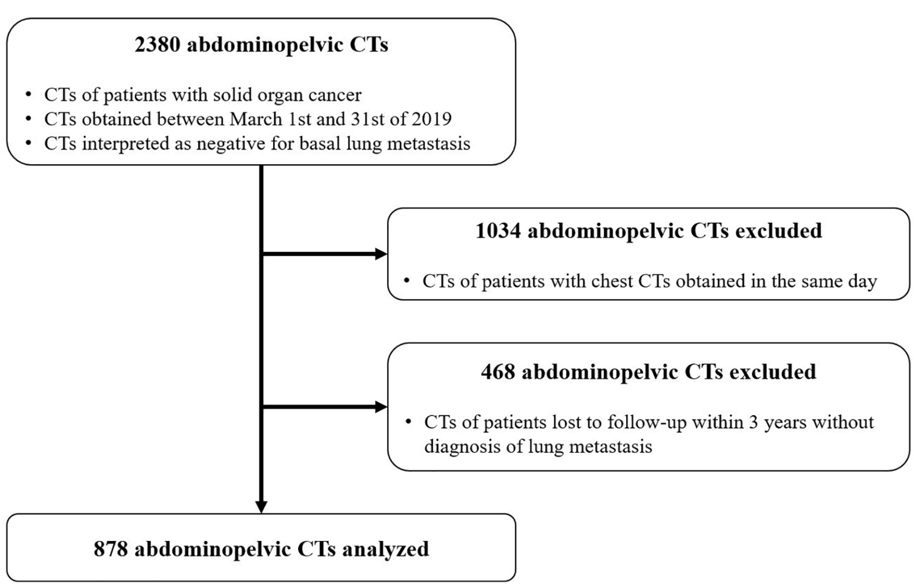


Figure 1. Flow diagram of the study. CT, computed tomography.

nary lesions that were stable for >3 years were considered benign. All individual pulmonary metastases present on the index abdominopelvic CT images but not documented in the radiologist's report were recorded as "overlooked metastases."

Artificial intelligence system

To detect pulmonary metastases in the basal lungs captured by abdominopelvic CT, an AI model based on a commercialized deep-learning-based CAD system (AVIEW Lung Nodule CAD, Coreline Soft, Seoul, Korea) was used. The CAD system was designed to detect pulmonary nodules in chest CT images and was approved for clinical use in Korea as an assistant tool for physicians in interpreting chest CT scans.

Since the original CAD system was optimized for low-dose chest CT images for lung cancer screening, the performance of the AI model may degrade when used for detecting pulmonary metastasis. Therefore, additional training of the AI model was conducted to optimize its performance in detecting small metastatic pulmonary nodules. A total of 3,558 CT scans were conducted, with 21,469 clinically diagnosed pulmonary metastases from a single institution (the same institution as where the present study was conducted). All of the abdominopelvic CT images were analyzed using an additionally trained AI model. Each pulmonary nodule was annotated by drawing three-dimensional bounding boxes on the CT images, along with a probability score (between 0 and 1) for the presence of a lesion (Figures 2-5). Then, these annotated CT images were used for the existing AI model for the original CAD system.

Radiologist's evaluation of artificial intelligence findings

All abdominopelvic CT images with corresponding AI results were reviewed by a fellowship trainee in thoracic radiology (H.S.C., 1st year of fellowship training) who was blinded to the diagnosis of pulmonary metastasis. The radiologist classified all lesions identified by the AI into three groups: those with the potential for pulmonary metastasis, clearly benign lesions, and pseudo-lesions. Subsequently, the radiologist checked the diagnoses of pulmonary metastasis to confirm that the lesions detected by AI were overlooked pulmonary metastases and classified the individual AI-detected lesions as either true-positives or false-positives.

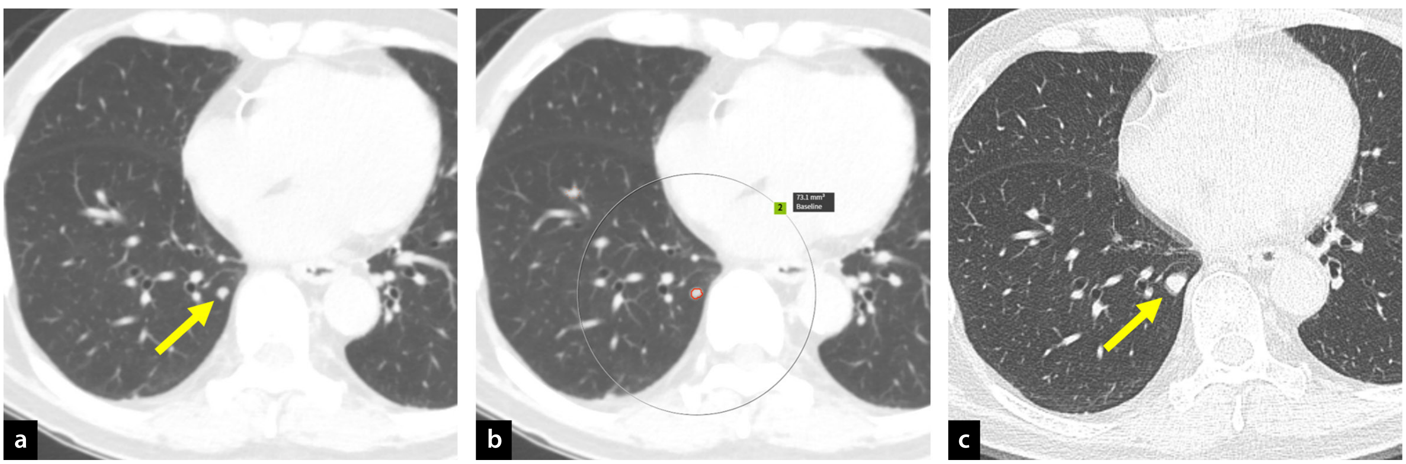


Figure 2. (a) This abdominal computed tomography (CT) image of a 72-year-old male patient with hepatocellular carcinoma shows a small nodule in the right lower lobe (arrow), a feature that was overlooked in the initial interpretation. **(b)** The artificial intelligence system detected the nodule with a probability score of 0.79. **(c)** A chest CT image obtained 167 days later shows growth of the nodule (arrow), which was clinically diagnosed as metastasis.

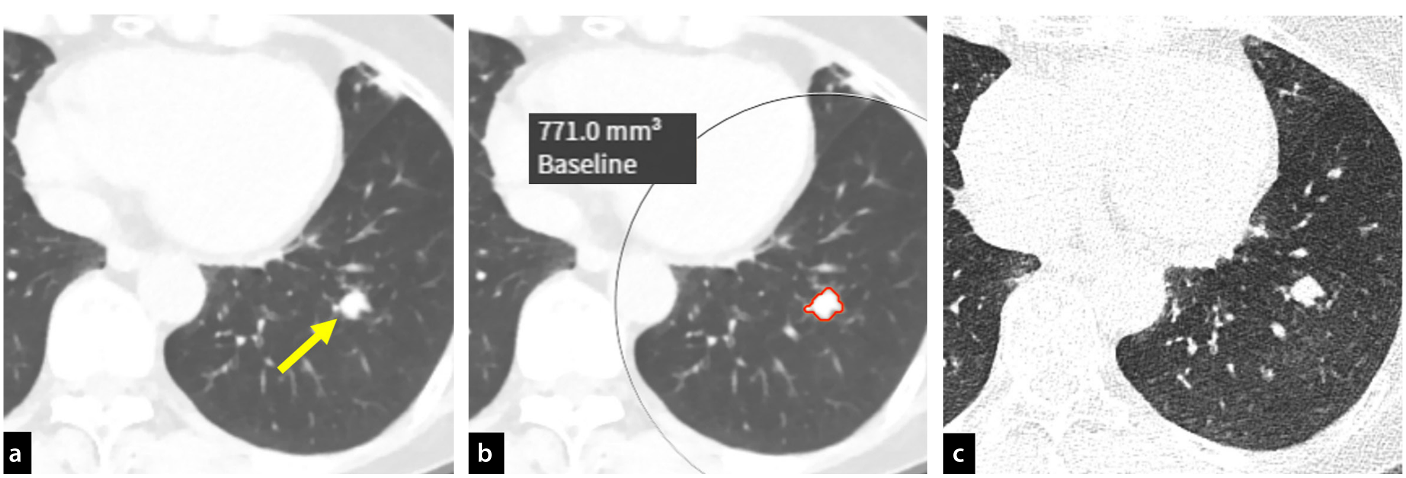


Figure 3. (a) This abdominopelvic computed tomography (CT) image of a 65-year-old male patient with colon cancer shows a nodule in the left lower lobe (arrow), a feature that was overlooked in the initial interpretation. **(b)** The artificial intelligence system detected the nodule with a probability score of 0.87. The radiologist who reviewed the AI results interpreted the lesion as a true nodule with the possibility of metastasis. **(c)** A chest CT image obtained 28 months later shows that the lesion remains unchanged, suggesting benignancy. AI, artificial intelligence.

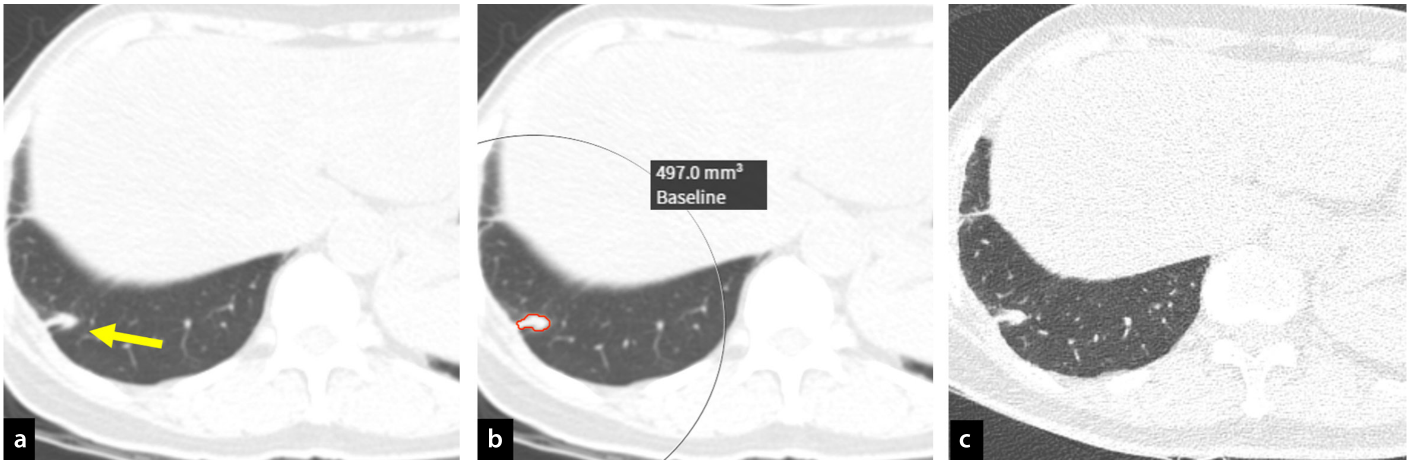


Figure 4. (a) This abdominopelvic computed tomography (CT) image of a 58-year-old female patient with colon cancer shows a nodular lesion in the right lower lobe (arrow), an observation that was not described in the initial interpretation. **(b)** The artificial intelligence (AI) system identified the lesion with a probability score of 0.50. The radiologist who reviewed the AI results interpreted the lesion as focal atelectasis rather than a true nodule. **(c)** A chest CT image obtained 44 months later shows that the lesion remained stable, suggesting benignancy.

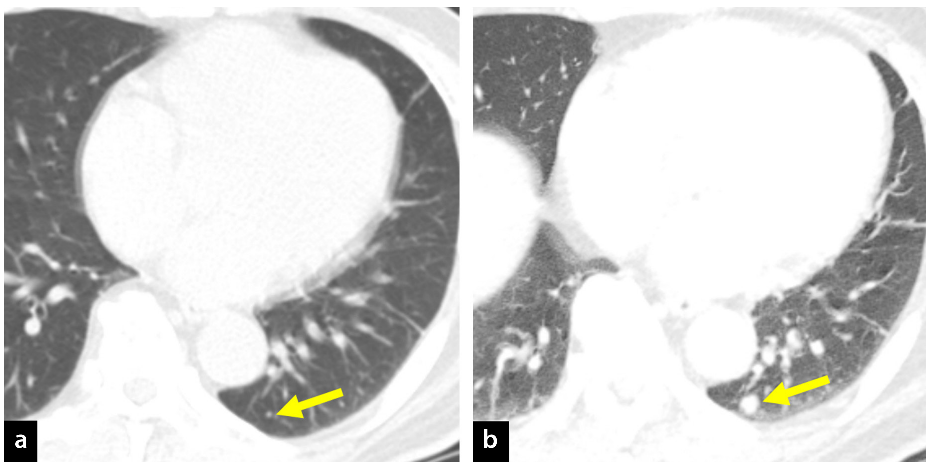


Figure 5. (a) This abdominopelvic computed tomography (CT) image of a 72-year-old male patient with colon cancer shows a tiny nodule in the left lower lobe (arrow), an observation that was overlooked in the initial interpretation. The artificial intelligence system did not detect the lesion. **(b)** A chest CT image obtained 204 days later shows the growth of the nodule, suggesting a diagnosis of lung metastasis (arrow).

Performance metrics and statistical analysis

First, the discriminative performance of the AI model in identifying patients with overlooked metastases was evaluated using an area under the receiver operating characteristic curve (AUC-ROC) analysis. Subsequently, the performance and efficacy of the AI model were evaluated using metrics at threshold probability scores of 0.4, 0.5, 0.6, and 0.7.

- Patient-level sensitivity = number of patients with true-positive detection of overlooked metastases/number of patients with overlooked metastases.

- Patient-level false-positive rate = number of patients with false-positive detection of overlooked metastases/number of patients without overlooked metastases.

- Patient-level positive predictive value (PPV) = number of patients with true-positive detection of overlooked metastases/number of patients with positive AI results.

- Lesion-level sensitivity = number of true-positive detections of overlooked metastases/number of all overlooked metastases.

- Number of false-positive lesions per patient = number of false-positive detections of overlooked metastases/number of patients.

- Lesion-level PPV = number of true-positive detections of overlooked metastases/number of all lesions detected by AI.

All metrics were obtained for both the AI results and the radiologist's review of the AI results (following the exclusion of clearly benign lesions or pseudo-lesions). The performance metrics of the AI results and the radiologist's review of the AI results were compared using McNemar's tests, chi-squared tests, and paired t-tests.

Decision curve analysis was conducted to evaluate the net benefit of using the AI tool as a second reader for detecting overlooked pulmonary metastasis, considering the benefit of true-positive results and the cost of false-positive results.

Statistical analysis

All statistical analyses were performed using MedCalc statistical software (MedCalc Software Ltd, Ostend, Belgium, 22.006 version). Statistical significance was set at $P < 0.05$.

Results

Patient characteristics

A total of 878 abdominopelvic CT images from 878 patients (580 men; mean age ± standard deviation: 62 ± 11 years) were included in the study (Figure 1). The most common primary malignancy was hepatocellular carcinoma (411, 47%), followed by stomach cancer (169, 19%) and colorectal cancer (96, 11%). A total of 707 CT examinations (81%) were obtained after the administration of intravenous contrast media. Table 1 presents the demographic information of the patients and their CT imaging characteristics.

Sixty-nine (7.8%) patients were diagnosed with lung metastases within 3 years of an abdominopelvic CT, including 5 patients who had already been diagnosed with lung metastases at the time of the CT. In a retrospective evaluation of abdominopelvic CT

Table 1. Patients and computed tomography characteristics

Variables	All patients (n = 878)	Patients with overlooked lung metastases (n = 13)	Patients without overlooked lung metastases (n = 865)
Age, mean ± SD (years)	62 ± 11	65 ± 13	62 ± 11
Male-to-female patient ratio	580:298	7:6	573:292
Primary malignancy			
Hepatocellular carcinoma	411 (47%)	3 (25%)	408 (47%)
Stomach cancer	169 (19%)	1 (8%)	168 (19%)
Colorectal cancer	96 (11%)	2 (15%)	94 (11%)
Biliary tree or pancreatic cancer	46 (5%)	5 (42%)	41 (5%)
Uterus or ovary cancer	25 (3%)	0	25 (3%)
Urinary tract cancer	16 (2%)	1 (8%)	15 (2%)
Breast cancer	15 (2%)	1 (8%)	14 (2%)
Prostate cancer	9 (1%)	0	9 (1%)
Others	91 (10%)	0	91 (11%)
CT examination with intravenous contrast media	844 (96%)	10 (83%)	834 (96%)
Multiphase CT examination	707 (81%)	9 (75%)	699 (81%)

Numbers in parentheses indicate proportions among all patients. CT, computed tomography; SD, standard deviation.

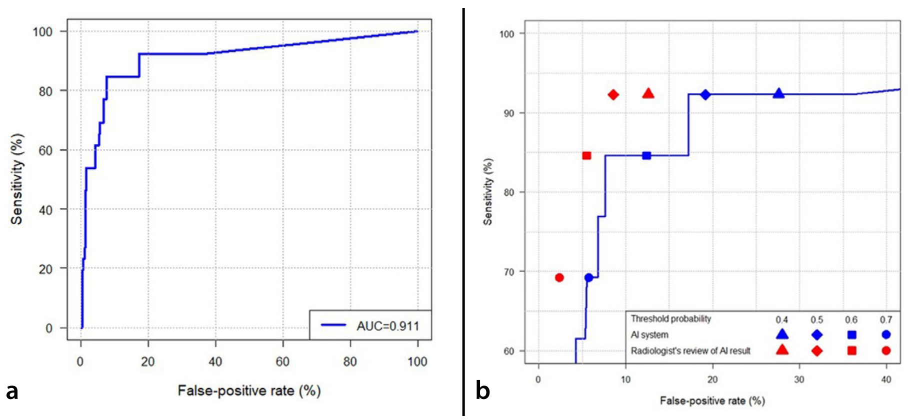


Figure 6. Receiver operating characteristic (ROC) curve for the identification of abdominopelvic computed tomography (CT) scans with overlooked basal lung metastases. **(a)** A ROC curve show that the artificial intelligence (AI) system identified abdominopelvic CT scans with overlooked basal lung metastases with an area under the ROC curve of 0.911. **(b)** The modified ROC curve shows that the sensitivity and false-positive rate of the AI system ranged from 69.2–92.3% and 46.2–92.3%, respectively at thresholds between 0.4 and 0.7. The radiologist's review significantly reduced the false-positive rate (2.4–12.6%) while preserving the sensitivity. AUC, area under the curve.

images, 13 (1.5%) patients had pulmonary metastases that were overlooked during interpretation. Of these 13 patients, 3 had already been diagnosed with lung metastases at the time of the CT. For the other 10 patients, the time interval between the abdominopelvic CT with overlooked lung metastases and the clinical diagnosis of lung metastasis was 141 days (interquartile range, 78–195 days).

Performance of the artificial intelligence system

For the discrimination of CT examinations with and without overlooked pulmonary metastases, the AI system exhibited an AUC-ROC value of 0.911 [95% confidence interval (CI), 0.890–0.929; Figure 6]. The results of the AI analyses and their performances for different thresholds are listed in Table 2 and Table 3. At the lowest threshold (0.4), the AI system detected 475 lesions (0.54 per examination) in 251 patients (positive rate, 28.7%). In contrast, it detected 100 lesions (0.11 per examination) in 59 (positive rate, 6.7%) patients at the highest threshold (0.7). The sensitivities of the AI system for the identification of patients with overlooked metastases were 92.3% (12/13; 95% CI, 64.0%–99.8%) at the lowest threshold and 69.2% (9/13; 95% CI, 38.6%–90.9%) at the highest threshold. Correspondingly, the patient-level false-positive rates ranged from 5.8% (50/865; 95% CI, 4.3%–7.6%) to 27.6% (239/865; 95% CI, 24.7%–30.7%), and the PPVs ranged from 4.8% (12/251; 95% CI, 2.5%–8.2%) to 15.3% (9/59; 95% CI, 7.2%–27.0%). The accuracy of the AI system ranged from 72.7% (638/878; 95% CI, 69.6%–75.6%) to 93.8% (824/878; 95% CI, 92.1%–95.4%).

Among 26 overlooked pulmonary metastases in eight patients, the sensitivities of the AI system were 92.3% (24/26; 95% CI, 74.5%–99.1%) at the lowest threshold and 46.2% (12/26; 95% CI, 26.6%–66.6%) at the highest. Correspondingly, the number of false-positive detections per examination ranged from 0.10 (88/878; 95% CI, 0.03–0.17) to 0.51 (451/878; 95% CI, 0.35–0.69), and the PPVs ranged from 5.1% (24/475; 95% CI, 3.3%–7.4%) to 12.0% (12/100; 95% CI, 6.4%–20.0%).

In the decision curve analysis, using the AI system as a second reader for detecting overlooked pulmonary metastases exhibited a higher net benefit than the default scenario without AI when the risk threshold was ≤3.7% (Figure 7). In other words, using the AI would be beneficial if the ratio of the cost from false-positive results to the benefit from true-positive results is ≤3.7:96.3 (1:26).

Review of the artificial intelligence results by the radiologist

Following the review of the AI results by the radiologist, 57.9% (275/475) of the

Table 2. Patient-level performance of the artificial intelligence system and the radiologist's review of the artificial intelligence results

Performance metric	Threshold probability 0.4	Threshold probability 0.5	Threshold probability 0.6	Threshold probability 0.7
AI system				
Number of true-positive results	12	12	11	9
Number of false-positive results	239	164	107	50
Number of true-negative results	1	1	2	4
Number of false-negative results	626	701	758	815
Positive rate	28.7% (251/878; 25.6%, 31.7%)	20.0% (176/878; 17.4, 22.9)	13.4% (118/878; 11.3, 15.9)	6.7% (59/878; 5.2, 8.6)
Number of detections per patient	0.54 (475/878; 0.37, 0.71)	0.36 (319/878; 0.22, 0.50)	0.23 (205/878; 0.12, 0.35)	0.11 (100/878; 0.05, 0.18)
Sensitivity	92.3% (12/13; 64.0, 99.8)	92.3% (12/13; 64.0, 99.8)	84.6% (11/13; 54.6, 98.1)	69.2% (9/13; 38.6, 90.9)
False-positive rate	27.6% (239/865; 24.7, 30.7)	19.2% (164/865; 16.4, 21.7)	12.4% (107/865; 10.3, 14.8)	5.8% (50/865; 4.3, 7.6)
PPV	4.8% (12/251; 2.5, 8.2)	6.8% (12/176; 3.6, 11.6)	9.3% (11/118; 4.8, 16.1)	15.3% (9/59; 7.2, 27.0)
Accuracy	72.7% (638/878; 69.6, 75.6)	81.2% (713/878; 78.5, 83.7)	87.6% (769/878; 85.2, 89.7)	93.8% (824/878; 92.1, 95.4)
Radiologist's review of the AI results				
Number of true-positive results	12	12	11	9
Number of false-positive results	109	74	48	21
Number of true-negative results	1	1	2	4
Number of false-negative results	756	791	817	844
Positive rate	13.8% (121/878; 11.6, 16.2)	9.8% (86/878; 7.9, 12.0)	6.7% (59/878; 5.2, 8.6)	3.4% (30/878; 2.3, 4.8)
P value	<0.001	<0.001	<0.001	<0.001
Number of detections per patient	0.23 (200/878; 0.18, 0.28)	0.15 (128/878; 0.11, 0.18)	0.09 (78/878; 0.06, 0.11)	0.04 (35/878; 0.02, 0.06)
P value	<0.001	<0.001	<0.001	<0.001
Sensitivity	92.3% (12/13; 64.0, 99.8)	92.3% (12/13; 64.0, 99.8)	84.6% (11/13; 54.6, 98.1)	69.2% (9/13; 38.6, 90.9)
P value	NA	NA	NA	NA
False-positive rate	12.6% (109/865; 10.5, 15.0)	8.6% (74/865; 6.8, 10.6)	5.5% (48/865; 4.1, 7.3)	2.4% (21/865; 1.5, 3.7)
P value	<0.001	<0.001	<0.001	<0.001
PPV	9.9% (12/121; 5.2, 16.7)	14.0% (12/86; 7.4, 23.1)	18.6% (11/59; 9.7, 30.9)	30.0% (9/30; 14.7, 49.4)
P value	0.059	0.061	0.077	0.104
Accuracy	87.5% (768/878; 85.1, 89.6)	91.5% (803/878; 89.4, 93.2)	94.3% (828/878; 92.6, 95.7)	97.2% (853/878; 95.8, 98.2)
P value	<0.001	<0.001	<0.001	<0.001

Numbers in parentheses indicate numerators/denominators and 95% confidence intervals. *P* values indicate a comparison between the AI system and the radiologist's review of the AI's result. AI, artificial intelligence; PPV, positive predictive value; NA, not applicable.

lesions detected by the AI were regarded as false-positive detections at the lowest threshold, while 65.0% (65/100) were regarded as false-positive detections at the highest threshold. As a result, the positivity rate after the radiologist's review was 13.8% (121/878) at the lowest threshold and 3.4% (30/878) at the highest threshold.

The sensitivities in the identification of patients with overlooked metastases were 92.3% (12/13; 95% CI, 64.0%–99.8%) at the lowest threshold and 69.2% (9/13; 95% CI, 38.6%–90.9%) at the highest threshold, consistent with the initial analyses by the AI. Meanwhile, the patient-level false-positive rates ranged from 2.4% (21/865; 95% CI, 1.5%–3.7%) to 12.6% (109/865; 95% CI, 10.5%–15.0%), representing a significant reduction compared with the initial analyses by the AI (all $p < 0.001$). Additionally, the patient-level PPVs ranged from 9.9% (12/121; 95% CI, 5.2%–16.7%) to 30.0% (9/30; 95% CI, 14.7%–49.4%) (Table 2) and were increased from the initial analyses by the AI, although the difference was not statistically significant. The accuracy ranged from 87.5% (768/878;

Table 3. Lesion-level performance of the artificial intelligence system and the radiologist's review of the artificial intelligence results

Performance metric	Threshold probability 0.4	Threshold probability 0.5	Threshold probability 0.6	Threshold probability 0.7
AI system				
Number of true-positive results	24	18	16	12
Number of false-positive results	451	301	189	88
Number of true-negative results	401	547	663	764
Number of false-negative results	2	12	10	14
Number of detection per patient	0.54 (475/878; 0.37, 0.71)	0.36 (319/878; 0.22, 0.50)	0.23 (205/878; 0.12, 0.35)	0.11 (100/878; 0.05, 0.18)
Sensitivity	92.3% (24/26; 74.5, 99.1)	69.2% (18/26; 48.2, 85.7)	61.5% (16/26; 40.6, 79.8)	46.2% (12/26; 26.6, 66.6)
Number of false-positive lesions per patient	0.51 (451/878; 0.35, 0.69)	0.34 (301/878; 0.20, 0.48)	0.22 (189/878; 0.10, 0.33)	0.10 (88/878; 0.03, 0.17)
PPV	5.1% (24/475; 3.3, 7.4)	5.6% (18/319; 3.4, 8.8)	7.8% (16/205; 4.5, 12.4)	12.0% (12/100; 6.4, 20.0)
Radiologist's review of the AI results				
Number of true-positive results	24	18	16	12
Number of false-positive results	176	110	62	23
Number of true-negative results	676	738	790	829
Number of false-negative results	2	12	10	14
Number of detection per patient	0.23 (200/878; 0.18, 0.28)	0.15 (128/878; 0.11, 0.18)	0.09 (78/878; 0.06, 0.11)	0.04 (35/878; 0.02, 0.06)
Sensitivity	92.3% (24/26; 74.5, 99.1)	69.2% (18/26; 48.2, 85.7)	61.5% (16/26; 40.6, 79.8)	46.2% (12/26; 26.6, 66.6)
P value	NA	NA	NA	NA
Number of false-positive lesions per patient	0.20 (176/878; 0.17, 0.25)	0.12 (110/878; 0.10, 0.16)	0.07 (62/878; 0.05, 0.09)	0.03 (23/878; 0.02, 0.04)
P value	<0.001	<0.001	<0.001	<0.001
PPV	12.0% (24/200; 7.8, 17.3)	14.1% (18/128; 0.09, 0.21)	20.5% (16/78; 0.12, 0.31)	34.3% (12/35; 19.1, 52.2)
P value	0.001	0.003	0.003	0.003

Numbers in parentheses indicate numerators/denominators and 95% confidence intervals. *P* values indicate a comparison between the AI system and the radiologist's review of the AI's result. AI, artificial intelligence; PPV, positive predictive value; NA, not applicable.

95% CI, 85.1%–89.6%) to 97.2% (853/878; 95% CI, 95.8%–98.2%). The accuracies exhibited significant improvement compared with the initial analyses by the AI (all $P < 0.001$).

The lesion-level sensitivities after the radiologist's review also remained similar to those following the initial analyses by the AI [92.3% (24/26; 95% CI, 74.5%–99.1%) at the lowest threshold; 46.2% (12/26; 95% CI, 26.6%–66.6%) at the highest threshold]. Meanwhile, the number of false-positive detections per examination ranged from 0.03 (23/878; 95% CI, 0.02–0.04) to 0.20 (176/878; 95% CI, 0.17–0.25), representing a significant reduction compared with the initial analyses by the AI (all $P < 0.001$). In addition, the lesion-level PPVs exhibited a significant increase compared with the initial analyses by the AI [$P \leq 0.001$; 12.0% (24/200; 95% CI, 7.8%–17.3%) at the lowest threshold; 34.3% (12/35; 95% CI, 19.1%–52.2%) at the highest threshold].

Table 4 displays the patterns of false-positive detections by the AI system. The most common cause of false-positive detection was pulmonary nodules with the possibility of metastasis, based on the radiologist's review. Among clearly benign lesions that were regarded as false-positive detections by the radiologist's review, findings of infection or inflammation were the most common causes of false-positive detections, followed by calcified nodules.

Clinical significance

The AI system may identify missed basal lung lesions in abdominopelvic CT scans in patients with malignancy, providing feedback to radiologists, which can reduce the risk of missing basal lung metastasis.

Discussion

An AI system for pulmonary nodule detection on CT images can be utilized as a second reader after the radiologist's interpretation to prevent radiologists from overlooking clinically relevant pulmonary nodules. In the present study, we used an AI system to detect metastatic pulmonary nodules in the basal lungs captured by abdominopelvic CT images that were overlooked by radiologists. The results showed that the AI system could identify CT images with overlooked pulmonary metastases, with an AUC-ROC value of 0.911 and maximum patient-level and lesion-level sensitivity of 92.3%, respectively. Although the AI generated several false-positive detections (maximum false-positive rate of 27.6%, 0.51 false-positive detections per patient), the radiologist's review of the AI results could effectively reduce the rate and number of false-positive detections (maximum false-positive rate of 12.6%, 0.20 false-positive detections per patient; $P < 0.001$, respectively).

Multiple studies have reported good performance of AI in the detection of pulmonary nodules on chest CT images.[16-19] In this study, the performance of the AI reached a level similar to that of radiologists. However, considering that AI cannot replace a radiologist's interpretation, its efficacy needs to be investigated based on its method of utilization. Since the most widely accepted methods of utilization involve CAD tools,[20-25] many studies have reported that AI can improve the performance of radiologists in lung nodule detection.[5,6,15-17] In addition to the use of AI as a CAD tool, several other utilization methods may also be feasible.[15] For instance, one promising method is its use as a second reader. In this context, the AI may analyze images after the radiologist's interpretation and provide feedback to the radiologist only when the AI suspects that the radiologist has overlooked a pulmonary nodule. In this scenario, the oversight of significant pulmonary nodules can be prevented without the need to review the AI results of all the examinations.

We performed decision curve analyses to evaluate the net benefit of applying the AI system for true-positive and false-positive identifications. The scenario with AI as a second reader showed a higher net benefit than the scenario without AI when the ratio between the harm of false-positive interpretations to the benefit of true-positive interpretations is ≤1:26. In most clinical situations, overlooking pulmonary metastases could have significant consequences, potentially depriving the patient of timely systemic treatment. Meanwhile, false-positive detections by AI may lead to a review by the radiologist, and the associated costs would be relatively much smaller compared with the risks of overlooking pulmonary metastases. Therefore, we believe that using the AI as a second reader would be a reasonable scenario.

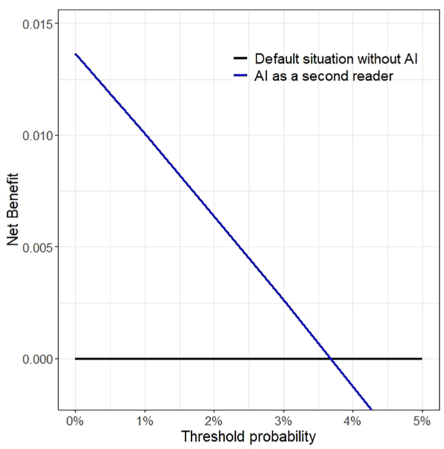


Figure 7. Decision curve for the identification of overlooked basal lung metastases in abdominopelvic computed tomography scans. The artificial intelligence (AI) as a second reader scenario (blue line) exhibited a higher net benefit than the default scenario without AI (black line) when the risk threshold is 3.7% or smaller. In other words, using the AI tool would be beneficial if the ratio of cost from false-positive results to benefit from true-positive results is 3.7:96.3 (1:26) or smaller.

In our study, an AI system was applied to the abdominopelvic CT scans of patients with cancer who were interpreted as negative for basal lung metastasis. In a retrospective evaluation of available follow-up examinations, overlooked pulmonary metastases were identified in 1.5% of patients, a frequency that should not be ignored. In this context, the AI could accurately discriminate between CT images with and without overlooked pulmonary metastases (AUC-ROC, 0.911). Furthermore, at a sensitive operating threshold, the AI could identify most CT scans with overlooked metastases (sensitivity: 92.3%). Notably, the identification of false-negative interpretations by radiologists using AI has been investigated in the field of chest radiography. Specifically, Nam et al.[26] and Jang et al.[27] reported that AI can identify lung cancers overlooked by radiologists on chest X-rays. In addition, Hwang et al.[28] reported that AI can identify various clinically relevant abnormalities on chest radiographs that were previously interpreted as normal.

Because benign pulmonary nodules and pulmonary metastases are often difficult to differentiate, false-positive detection by AI is inevitable. When used as a second reader,[15] false-positive detection may lead to unnecessary feedback to the radiologist, followed by reinterpretation by the radiologist. In our study, the maximum false-positive rate was 27.6%, indicating that the AI may generate false-positive feedback in 27.6% of CT images without overlooking metastases. Based on the review of the AI results by a radiologist, more than half of the AI detections were regarded as clearly benign nodules (findings of pulmonary infection and calcified nodules). Notably, the radiologist's review was effective because it significantly reduced the rate of false positives while maintaining a similar sensitivity for metastasis. The results also suggest that further improvements in AI performance may reduce the false-positive rate and the frequency of unnecessary reinterpretation by radiologists.

Pulmonary metastases and benign pulmonary nodules are often indistinguishable, even when evaluated by a radiologist. Therefore, as expected, there were considerable false-positive detections even after the radiologist's review (maximum false-positive rate: 12.6%). Moreover, the identification of benign nodules may lead to the requirement of chest CT examinations for further evaluation or follow-up of the pulmonary nodules. Considering that all patients were under follow-up for malignancies, we believe that additional chest CT scans may not significantly harm the patients.

Our study has several limitations. First, since our study was conducted at a single tertiary medical institution, the reproducibility of our results remains uncertain. Future studies may be required to confirm the reproducibility of our results in other clinical situations. Second, although we consecutively included 878 abdominopelvic CT scans, the absolute number of overlooked pulmonary metastases is quite small (n = 13), limiting the statis-

Table 4. Detection patterns of the artificial intelligence system

Variable	Threshold probability 0.4	Threshold probability 0.5	Threshold probability 0.6	Threshold probability 0.7
Lesions with the possibility of metastasis based on the radiologist's review				
Metastasis	24 (5.1%)	18 (5.6%)	16 (7.8%)	12 (12%)
Benign lung nodules	176 (37.1%)	110 (34.5%)	62 (30.2%)	23 (23%)
Clearly benign lesions based on the radiologist's review				
Findings of infection/inflammation	116 (24.4%)	88 (28%)	66 (32.2%)	40 (40%)
Calcified nodules	90 (18.9%)	72 (27.6%)	49 (24.0%)	23 (23%)
Ground-glass nodules	33 (6.9%)	21 (6.6%)	9 (4.4%)	1 (1%)
Pulmonary vessels	24 (5.1%)	3 (0.9%)	1 (0.5%)	0 (0%)
Atelectasis	7 (1.5%)	4 (1.3%)	1 (0.5%)	1 (1%)
Others	5 (1.1%)	3 (0.9%)	1 (0.5%)	0 (0%)
Total	475 (100%)	319 (100%)	205 (100%)	100 (100%)

Numbers in parentheses indicate the proportions among the total detections by the artificial intelligence system.

tical power. A multicenter study with a larger sample size may be required to confirm the efficacy of AI as a second reader. Third, in this study, AI was retrospectively applied to abdominopelvic CT scans. Therefore, the practical efficacy of AI systems remains unknown. A prospective study following the integration of AI into the workflow may be required to investigate its real-world efficacy. Finally, the effect of AI beyond the detection of overlooked metastases, including its effects on patient outcomes and changes in treatment decision-making, remains unknown.

In conclusion, the applied AI system could accurately identify basal lung metastases captured in abdominopelvic CT images that were overlooked by radiologists, suggesting its potential as a second reader after the radiologist's interpretation. Further prospective studies are warranted to investigate the real-world efficacy of AI as a second reader as well as the impact of AI beyond the detection of metastases.

Conflict of interest

Eui Jin Hwang reports a research grant from Coreline Soft outside the present work; Jaeyoun Yi and Boorym Choi are employees of Coreline Soft; Chang Min Park reports a research grant from Coreline Soft outside the present work and stock options of Coreline Soft. The other author declared no conflict of interest.

Funding

This work was supported by the Technology Development Program (S3141595) funded by the Ministry of SMEs and Startups (MSS, Korea).

References

1. Tan JR, Cheong EHT, Chan LP, Tham WP. Implementation of an artificial intelligence-based double read system in capturing pulmonary nodule discrepancy in CT studies. *Curr Probl Diagn Radiol.* 2021;50(2):119-122. Erratum in: *Curr Probl Diagn Radiol*. 2022;51(4):673. [Crossref]
2. Li F, Armato SG, Giger ML, MacMahon H. Clinical significance of noncalcified lung nodules in patients with breast cancer. *Breast Cancer Res Treat.* 2016;159(2):265-271. [Crossref]
3. Rubin GD. Lung nodule and cancer detection in computed tomography screening. *J Thorac Imaging.* 2015;30(2):130-138. [Crossref]
4. Lee B, Lim A, Lalvani A, et al. The clinical significance of radiologically detected silent pulmonary nodules in early breast cancer. *Ann Oncol.* 2008;19(12):2001-2006. [Crossref]
5. Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open*. 2019;2(3):e191095. [Crossref]
6. Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning-based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. *Clin Infect Dis.* 2019;69(5):739-747. [Crossref]
7. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med.* 2019;25(6):954-961. Erratum in: *Nat Med*. 2019;25(8):1319. [Crossref]
8. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410. [Crossref]
9. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318(22):2211-2223. [Crossref]
10. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199-2210. [Crossref]
11. Hinton G. Deep learning-a technology with the potential to transform health care. *JAMA*. 2018;320(11):1101-1102. [Crossref]
12. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. [Crossref]
13. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ*. 2006;332(7554):1089-1092. Erratum in: *BMJ*. 2006;332(7554):1368. [Crossref]
14. Tang A, Tam R, Cadrin-Chênevert A, et al. Canadian association of radiologists white paper on artificial intelligence in radiology. *Can Assoc Radiol J*. 2018;69:120-135. [Crossref]
15. Hwang EJ, Park CM. Clinical Implementation of deep learning in thoracic radiology: potential applications and challenges. *Korean J Radiol.* 2020;21(5):511-525. [Crossref]
16. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology*. 2019;290:218-228. [Crossref]
17. Harris M, Qi A, Jeagal L, et al. A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest X-rays for pulmonary tuberculosis. *PloS One.* 2019;14:e0221339. [Crossref]
18. Liu Z, Li Z, Qu J, et al. Radiomics of multiparametric MRI for pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study. *Clin Cancer Res.* 2019;25:3538-3547. [Crossref]
19. Sim Y, Chung MJ, Kotter E, et al. Deep convolutional neural network-based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology.* 2020;294(1):199-209. [Crossref]
20. Cha MJ, Chung MJ, Lee JH, Lee KS. Performance of deep learning model in detecting operable lung cancer with chest radiographs. *J Thorac Imaging*.2019;34(2):86-91. [Crossref]
21. Park S, Lee SM, Lee KH, et al. Deep learning-based detection system for multiclass lesions on chest radiographs: comparison with observer readings. *Eur Radiol*. 2020;30(3):1359-1368. [Crossref]
22. Majkowska A, Mittal S, Steiner DF, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology.* 2020;294(2):421-431. [Crossref]
23. Pande T, Cohen C, Pai M, Ahmad Khan F. Computer-aided detection of pulmonary tuberculosis on digital chest radiographs: a systematic review. *Int J Tuberc Lung Dis.* 2016;20:1226-1230. [Crossref]
24. Schalekamp S, van Ginneken B, Karssemeijer N, Schaefer-Prokop CM. Chest radiography: new technological developments and their applications. *Semin Respir Crit Care Med.* 2014;35:3-16. [Crossref]
25. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol.* 2019;20:405-410. [Crossref]
26. Nam JG, Hwang EJ, Kim DS, et al. Undetected lung cancer at posteroanterior chest radiography: potential role of a deep learning-based detection algorithm. *Radiol Cardiothorac Imaging.* 2020;2(6):e190222. [Crossref]
27. Jang S, Song H, Shin YJ, et al. Deep Learning-based automatic detection algorithm for reducing overlooked lung cancers on chest radiographs. *Radiology.* 2020;296(3):652-661. Erratum in: *Radiology*. 2022;302(2):E10. [Crossref]
28. Hwang EJ, Park J, Hong W, et al. Artificial intelligence system for identification of false-negative interpretations in chest radiographs. *Eur Radiol.* 2022;32(7):4468-4478. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242876

BREAST IMAGING

ORIGINAL ARTICLE

Evaluating text and visual diagnostic capabilities of large language models on questions related to the Breast Imaging Reporting and Data System Atlas 5th edition

Yasin Celal Güneş[1]
Turay Cesur[2]
Eren Çamur[3]
Leman Günbey Karabekmez[4]

[1]Kırıkkale Yüksek İhtisas Hospital, Clinic of Radiology, Kırıkkale, Türkiye

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

[3]Ankara 29 Mayıs State Hospital, Clinic of Radiology, Ankara, Türkiye

[4]Ankara Yıldırım Beyazıt University Faculty of Medicine, Department of Radiology, Ankara, Türkiye

PURPOSE

This study aimed to evaluate the performance of large language models (LLMs) and multimodal LLMs in interpreting the Breast Imaging Reporting and Data System (BI-RADS) categories and providing clinical management recommendations for breast radiology in text-based and visual questions.

METHODS

This cross-sectional observational study involved two steps. In the first step, we compared ten LLMs (namely ChatGPT 4o, ChatGPT 4, ChatGPT 3.5, Google Gemini 1.5 Pro, Google Gemini 1.0, Microsoft Copilot, Perplexity, Claude 3.5 Sonnet, Claude 3 Opus, and Claude 3 Opus 200K), general radiologists, and a breast radiologist using 100 text-based multiple-choice questions (MCQs) related to the BI-RADS Atlas 5th edition. In the second step, we assessed the performance of five multimodal LLMs (ChatGPT 4o, ChatGPT 4V, Claude 3.5 Sonnet, Claude 3 Opus, and Google Gemini 1.5 Pro) in assigning BI-RADS categories and providing clinical management recommendations on 100 breast ultrasound images. The comparison of correct answers and accuracy by question types was analyzed using McNemar's and chi-squared tests. Management scores were analyzed using the Kruskal–Wallis and Wilcoxon tests.

RESULTS

Claude 3.5 Sonnet achieved the highest accuracy in text-based MCQs (90%), followed by ChatGPT 4o (89%), outperforming all other LLMs and general radiologists (78% and 76%) ($P < 0.05$), except for the Claude 3 Opus models and the breast radiologist (82%) ($P > 0.05$). Lower-performing LLMs included Google Gemini 1.0 (61%) and ChatGPT 3.5 (60%). Performance across different categories of showed no significant variation among LLMs or radiologists ($P > 0.05$). For breast ultrasound images, Claude 3.5 Sonnet achieved 59% accuracy, significantly higher than other multimodal LLMs ($P < 0.05$). Management recommendations were evaluated using a 3-point Likert scale, with Claude 3.5 Sonnet scoring the highest (mean: 2.12 ± 0.97) ($P < 0.05$). Accuracy varied significantly across BI-RADS categories, except Claude 3 Opus ($P < 0.05$). Gemini 1.5 Pro failed to answer any BI-RADS 5 questions correctly. Similarly, ChatGPT 4V failed to answer any BI-RADS 1 questions correctly, making them the least accurate in these categories ($P < 0.05$).

CONCLUSION

Although LLMs such as Claude 3.5 Sonnet and ChatGPT 4o show promise in text-based BI-RADS assessments, their limitations in visual diagnostics suggest they should be used cautiously and under radiologists' supervision to avoid misdiagnoses.

CLINICAL SIGNIFICANCE

This study demonstrates that while LLMs exhibit strong capabilities in text-based BI-RADS assessments, their visual diagnostic abilities are currently limited, necessitating further development and cautious application in clinical practice.

KEYWORDS

BI-RADS, breast radiology, ChatGPT 4o, Claude 3.5 Sonnet, large language models

Corresponding author: Yasin Celal Güneş

E-mail: gunesyasincelal@gmail.com

Received 27 May 2024; revision requested 23 June 2024; last revision received 07 August 2024; accepted 24 August 2024.

Epub: 09.09.2024

Publication date: 03.03.2025

DOI: 10.4274/dir.2024.242876

You may cite this article as: 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.

The emergence of large language models (LLMs) marks a transformative milestone in the development of artificial intelligence (AI). These models offer unprecedented potential for understanding and generating human-like text by leveraging extensive datasets. This technological advancement holds significant promise for application in medicine.[1,2] As radiology increasingly relies on the interpretation of complex imaging data, the integration of advanced AI tools, such as LLMs, becomes crucial to enhance diagnostic accuracy and streamline workflows. LLMs have demonstrated remarkable performance in various realms of radiology, including testing radiological knowledge in different board-style examinations, simplifying radiology reports, and providing patient information.[3-7]

Recent studies have also explored the potential of LLMs specifically in breast imaging, where their capabilities show particular promise.[8-10] For instance, Rao et al.[9] evaluated the performance of two well-known LLMs, ChatGPT 3.5 and ChatGPT 4, in adhering to the American College of Radiology (ACR) eligibility criteria for breast pain and breast cancer screening, revealing impressive accuracy rates of 88.9% and 98.4%, respectively. These findings highlight the potential of LLMs as supportive tools in breast imaging, which is especially relevant given the ongoing radiologist shortages and the increasing volume of imaging studies.[11,12] Despite these advancements, it is crucial to acknowledge the limitations and challenges associated with LLMs, including their susceptibility to generating plausible-sounding but incorrect answers (hallucinations).[13]

The Breast Imaging Reporting and Data System (BI-RADS) Atlas, released in its latest edition in 2013, has provided standardized nomenclature, report organization, assessment structure, and a classification system for mammography, ultrasound, and magnetic resonance imaging (MRI) of the breast.[14] The BI-RADS Atlas is crucial for radiologists as it standardizes breast imaging terminology and reporting, ensuring clear communication and consistent, accurate patient management.[15]

While the BI-RADS Atlas offers a standardized approach to breast imaging, recent research has begun exploring how LLMs can further enhance radiological assessment and reporting accuracy. Haver et al.[16] demonstrated that ChatGPT 4 accurately predicted the BI-RADS category in 73.6% of 250 fictitious breast imaging reports. Cozzi et al.[17] evaluated the concordance between different LLMs (ChatGPT 3.5, ChatGPT 4, and Google Bard) and radiologists across 2,400 reports in three different languages, revealing a moderate agreement (Gwet's agreement coefficient: 0.52–0.42). Despite the growing emphasis on the importance of LLMs in breast imaging, there is a significant gap in the literature regarding the evaluation of multimodal LLMs' performance on breast ultrasound images. Additionally, no studies compare LLMs' knowledge of BI-RADS Atlas with that of radiologists. Hence, the first aim of this study is to evaluate the performance of nine large LLMs compared with breast and general radiologists on text-based multiple-choice questions (MCQs) related to the BI-RADS Atlas, 5th edition. The second aim is to assess the capability of five multimodal LLMs in assigning BI-RADS categories and providing clinical management recommendations for breast ultrasound images.

Main points

- This study evaluated the performance of large language models (LLMs) and multimodal LLMs in interpreting the Breast Imaging Reporting and Data System categories and providing clinical management recommendations. The evaluation involved two steps: assessing LLMs on text-based multiple-choice questions (MCQs) and evaluating multimodal LLMs on breast ultrasound images.
- Claude 3.5 Sonnet and ChatGPT 4o achieved high accuracy rates of 90% and 89%, respectively, in text-based MCQs, outperforming general radiologists, who had accuracy rates of 78% and 76%. This demonstrates the strong potential of these advanced LLMs in supporting and enhancing the diagnostic accuracy of radiologists in text-based assessments.
- Multimodal LLMs showed lower accuracy in evaluating breast ultrasound images, with Claude 3.5 Sonnet achieving only 59% accuracy. This highlights a critical limitation in their current ability to handle visual diagnostic tasks effectively compared with text-based assessments.
- The study underscores the necessity for further development of multimodal LLMs to improve their visual diagnostic capabilities. Until these improvements are realized, the use of multimodal LLMs in clinical practice should be closely supervised by experienced radiologists to prevent potential misdiagnoses and ensure patient safety.

Methods

Study design

This cross-sectional observational study had two steps. In the first step, it compared different LLMs, namely ChatGPT 4o, ChatGPT 4, ChatGPT 3.5, Google Gemini 1.5 Pro, Google Gemini 1.0, Microsoft Copilot, Perplexity, Claude 3.5 Sonnet, Claude 3 Opus, and Claude 3 Opus 200K, along with the responses of two general radiologists and a breast radiologist in answering MCQs regarding the 5th edition of the BI-RADS Atlas.

In the second step, the study compared different multimodal LLMs, namely ChatGPT 4o, ChatGPT 4V, Claude 3.5 Sonnet, Claude 3 Opus, and Google Gemini 1.5 Pro. This step focused on determining the correct BI-RADS category and clinical management by evaluating breast ultrasound images. An overview of the workflow is shown in Figure 1.

The study did not require ethics committee approval as it relied solely on fictional MCQs and a publicly available breast ultrasound dataset that had no identifiable patient information. Its design conformed to the principles articulated in the Standards for Reporting Diagnostic Accuracy Studies statement.[18]

Data collection for breast multiple-choice questions

The ACR published the 5th edition of the BI-RADS Atlas in 2013 to standardize terminology and reporting organization in breast radiology.[14] A total of 100 MCQs were prepared and categorized using the information in this atlas related to ultrasound, mammography, MRI, and general BI-RADS knowledge by general radiologist 3 (Y.C.G.). Each question had four choices, with only one correct answer and three distractors. The distractors were carefully chosen to be reasonable and related to the question. Each question was formulated to be clear and focused on a single concept to assess breast radiology knowledge. The questions were categorized according to the BI-RADS Atlas sections as follows: 16 on breast ultrasound, 39 on mammography, 22 on breast MRI, and 23 on general BI-RADS knowledge. All created MCQs are listed in Supplementary Material 1.

Design of input–output procedures and performance evaluation for large language models

The input prompt was initiated as follows: "I am working on a breast radiology quiz and will provide you MCQs. Act like a radiology professor with 30 years of expertise in breast imaging. Please indicate the correct answer. There is only one correct answer." This prompt was presented in April 2024 on eight distinct platforms with default param-

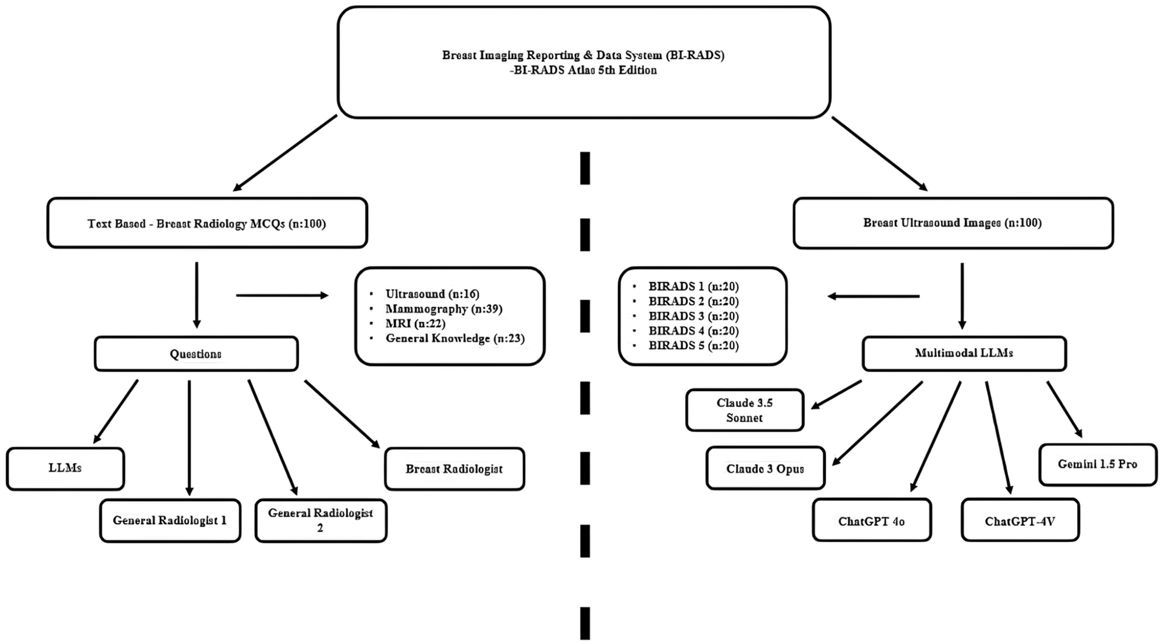


Figure 1. The workflow of the study. MCQs, multiple-choice questions; LLMs, large language models; MRI, magnetic resonance imaging.

eters: OpenAI's ChatGPT 4 and 3.5 (https://chat.openai.com), Google Gemini 1.5 Pro and 1.0 (https://gemini.google.com/), Microsoft Copilot (https://copilot.microsoft.com) (Balanced), Perplexity (https://perplexity.ai), Claude 3 Opus (https://claude.ai), and Claude 3 Opus 200K (https://poe.com). The same prompt was presented to OpenAI's ChatGPT 4o (https://chat.openai.com) in May 2024 and Claude 3.5 Sonnet (https://claude.ai) in July 2024 (Figure 2). Specific settings, such as temperature and randomness, were left at their default values unless specified otherwise by the platform.

The MCQs were sequentially added to the same chat session by copying and pasting from the MCQs list. Each LLM was presented with 100 questions by general radiologist 3, and the responses were evaluated. It is crucial to note that the employed LLMs were not pre-trained with a specific prompt or question set for this study. Each question was asked in a single chat session, without opening a new chat tab for individual inquiries.

Radiologist 3 evaluated LLMs' answers according to the correct answer list, marking them either correct (1) or incorrect (0).

Radiologists performance evaluation for breast multiple-choice questions

Two European Board of Radiology-certified junior general radiologists–radiologist 1 (T.C.) with 6 years of experience, and radiologist 2 (E.Ç.) with 6 years of experience–and a breast radiologist (L.G.K.) with 13 years of experience, independently assessed the MCQs blindly using their computers. All three answered questions in different sessions. Upon completion of all questions, radiologist 3 evaluated each other's answers according to the correct answer list, marking them either correct (1) or incorrect (0).

Multimodal large language models and visual breast ultrasound questions

The publicly available Breast Ultrasound Images dataset was utilized to assess the performance of multimodal LLMs with breast ultrasound images.[19] This dataset comprises 780 images classified as normal, benign, and malignant, sourced from 600 women aged 25–75 years. The images were acquired using the LOGIQ E9 ultrasound system [General Electric (GE) Healthcare, Wauwatosa, WI, USA] and the LOGIQ E9 Agile ultrasound system [General Electric (GE) Healthcare, Wauwatosa, WI, USA] at Baheya Hospital in Cairo, Egypt. The transducers used were 1–5 MHz on an ML6-15-D Matrix linear probe, and the images were stored in PNG format with dimensions of 500 × 500 pixels.[19]

The breast radiologist selected 20 images for each BI-RADS category from 1 to 5, resulting in a total of 100 images. These BI-RADS categories served as a reference standard. These images were presented to five different multimodal LLMs: Claude 3.5 Sonnet, Claude 3 Opus, Google Gemini 1.5 Vision Pro, ChatGPT 4o, and ChatGPT 4V.

For each image, the multimodal LLMs received the following prompt: "I am working on a breast radiology quiz and will provide you with breast ultrasound images. Please act as a radiology professor with 30 years of expertise in breast imaging. Evaluate the images and assign only one correct BI-RADS category from BI-RADS 1 to BI-RADS 5 according to the 5th edition of the BI-RADS Atlas. Lastly, provide clinical management recommendations for each category according to the same Atlas" (Figure 3).

This evaluation was conducted in July 2024, with each image presented along with the prompt, using the multimodal LLMs' default parameters.

The management recommendations provided by the multimodal LLMs, based on the BI-RADS categories, were evaluated using a 3-point Likert scale defined as the Management Score:

- 3 points: Correct management recommendations according to the BI-RADS category
- 2 points: Partially correct management recommendations according to the BI-RADS category
- 1 point: Completely incorrect management recommendations according to the BI-RADS category

Radiologist 3 provided the images and prompts to the multimodal LLMs and record-

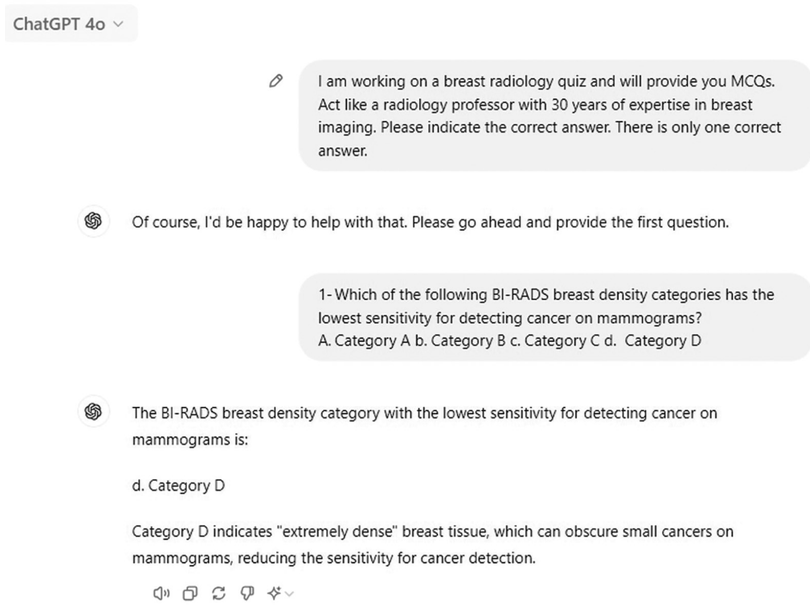


Figure 2. Demonstration of prompts provided to large language models and their responses. MCQs, multiple-choice questions; BI-RADS, Breast Imaging Reporting and Data System.

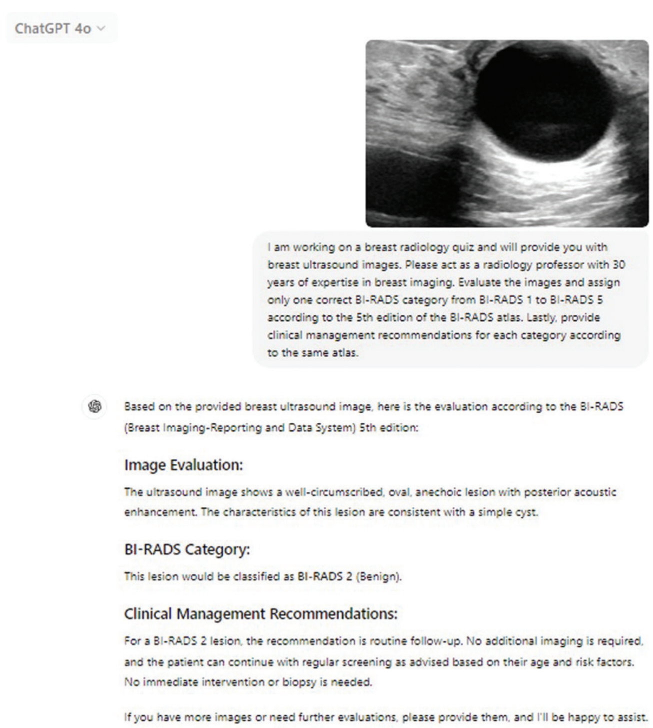


Figure 3. Illustrations of prompts and breast ultrasound images provided to multimodal large language models and their responses. MCQs, multiple-choice questions; BI-RADS, Breast Imaging Reporting and Data System.

ed their responses. The accuracy of these responses for BI-RADS categories was classified as correct (1) or incorrect (0), and clinical management recommendations were scored using the Management Score by the breast radiologist.

Statistical analysis

The distribution of variables was assessed using the Kolmogorov–Smirnov test. Descriptive statistics were represented using percentages. Non-parametric tests were employed to compare quantitative data due to the nature of the data distribution. The Kruskal–Wallis test was used to compare quantitative data, and Tamhane's T2 test was employed for multiple post-hoc comparisons following the initial Kruskal–Wallis test. McNemar's test was used to compare the proportion of correct responses between different questions. The chi-squared test was used to compare the correct answers by question types. The Wilcoxon test was used to compare the Management Scores of multimodal LLMs. The SPSS 26.0 (IBM, USA) package was used for statistical analyses, and statistical significance was set at $P < 0.05$.

Results

Accuracy of large language models on text-based breast multiple-choice questions

The highest success among the LLMs was achieved by Claude 3.5 Sonnet with an accuracy rate of 90%. ChatGPT 4o ranked second with an accuracy rate of 89%, followed by Claude 3 Opus 200K with an accuracy rate of 84%. Subsequently, Claude 3 Opus had an accuracy rate of 82%, and ChatGPT 4 had an accuracy rate of 79%. The diagnostic accuracy of the breast radiologist was 82%, radiologist 1 was 78%, radiologist 2 was 76%. Google Gemini 1.5 Pro had a 67% accuracy rate, and Microsoft Copilot with a 65% accuracy rate, while both Google Gemini 1.0 and Perplexity scored 61%, and ChatGPT 3.5 scored 60% accuracy (Figure 4).

Claude 3.5 Sonnet achieved the highest accuracy rate among the evaluated LLMs, outperforming most models with a statistically significant difference ($P < 0.05$), except when compared with ChatGPT 4o and Claude 3 Opus. Both Claude 3.5 Sonnet and ChatGPT 4o also surpassed the accuracy of the general radiologists ($P < 0.05$), although their performance was comparable with that of the breast radiologists ($P > 0.05$). Additionally, no significant differences were observed between the breast radiologist and general radiologists ($P > 0.05$).

When comparing the LLMs Claude 3 Opus 200K, Claude 3 Opus, and ChatGPT 4 with the radiologists, there were no statistically significant differences ($P > 0.05$); however, these models showed significant superiority over lower-performing LLMs, namely Google Gemini 1.5 Pro, Microsoft Copilot, and ChatGPT 3.5 ($P < 0.001$). No significant differences were found between the performances of the LLMs and radiologists across different question categories ($P > 0.05$). Detailed comparisons of the performance between radiologists and LLMs are shown in Table 1, while the performance across question categories is illustrated in Figure 5 and Table 2.

Accuracy of multimodal large language models on visual breast ultrasound questions

In a visual test consisting of 100 questions on breast ultrasound images, Claude 3.5 Sonnet achieved an accuracy rate of 59%, ChatGPT 4o 39%, Google Gemini 1.5 Pro 31%, ChatGPT 4V 20%, and Claude 3 Opus 19% (Figure 6). The performance of Claude 3.5 Sonnet was significantly higher than that of the other multimodal LLMs ($P < 0.05$). While there was no significant difference in performance between ChatGPT 4o and Google Gemini 1.5 Pro ($P = 0.067$), Claude 3 Opus and ChatGPT 4V had significantly lower performance ($P < 0.05$) (Table 3).

The accuracy rates of each model by BI-RADS categories were analyzed using the chi-squared test. The statistical analysis revealed that only Claude 3 Opus's accuracy rate did not vary by BI-RADS categories ($P = 0.992$); for other models, accuracy rates showed significant variation by category ($P < 0.05$) (Table 4).

In post-hoc tests:

• Claude 3.5 Sonnet had a higher accuracy rate for BI-RADS 5 questions (85%) compared with other categories ($P = 0.001$), while its accuracy rate for BI-RADS 1 questions (35%) was lower compared with other categories ($P = 0.001$).

• Google Gemini 1.5 Pro's accuracy rate for BI-RADS 5 questions (0%) was lower compared with other categories ($P < 0.001$).

• ChatGPT 4V had a higher accuracy rate for BI-RADS 5 questions (45%) compared with other categories ($P = 0.001$), but a lower accuracy rate for BI-RADS 1 questions (0%) ($P = 0.012$).

• ChatGPT 4o had a higher accuracy rate for BI-RADS 2 questions (65%) compared with other categories ($P = 0.007$) (Figure 7).

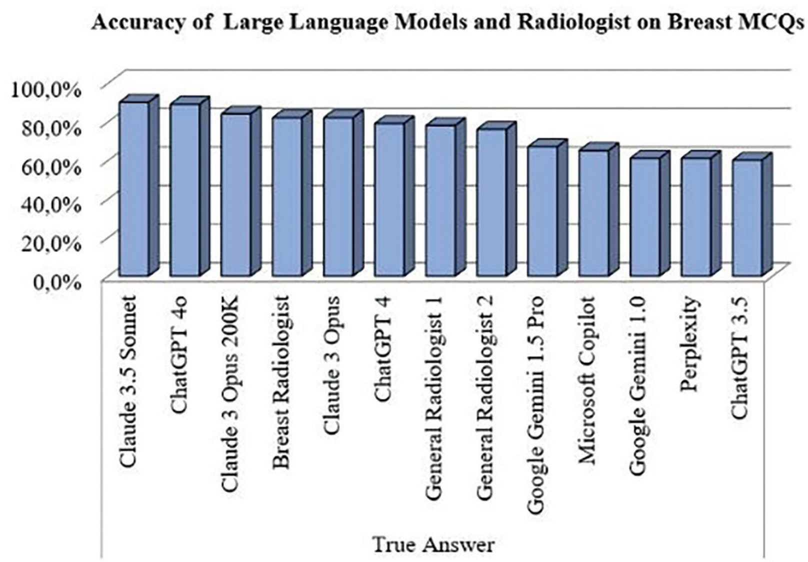


Figure 4. Accuracy of large language models and radiologists on breast multiple-choice questions. MCQs, multiple-choice questions.

Accuracy of multimodal large language models on clinical management recommendations

The mean Management Score of Claude 3.5 Sonnet (mean: 2.12 ± 0.97) was significantly superior to that of all other multimodal LLMs ($P < 0.05$). The mean Management Score of ChatGPT 4o (mean: 1.78 ± 0.98) was not significantly different from Google Gemini 1.5 Pro (mean: 1.64 ± 0.93), but it outperformed ChatGPT 4V (mean: 1.40 ± 0.80) and Claude 3 Opus (mean: 1.42 ± 0.81) ($P < 0.05$). The details of the Management Score are given in Supplementary Material 2.

Discussion

This study aimed to evaluate the performance of LLMs and multimodal LLMs in breast radiology knowledge. The most striking finding of our study is that although LLMs excel at text-based questions, their per-

Table 1. Comparison of the accuracy of LLMs and radiologists with *P* values obtained from McNemar's test

	Claude 3.5 Sonnet	Claude 3 Opus 200k	Claude 3 Opus	ChatGPT 4o	ChatGPT 4	ChatGPT 3.5	BR	R-1	R-2	Google Gemini 1.5 Pro	Google Gemini 1.0	Perplexity
Claude 3.5 Sonnet	-	0.210	0.096	1	0.019	<0.001	0.077	0.004	<0.001	<0.001	<0.001	<0.001
Claude 3 Opus 200k	0.210	-	0.774	0.302	0.359	0.001	0.832	0.327	0.152	<0.001	<0.001	<0.001
Claude 3 Opus	0.096	0.774	-	0.189	0.648	0.002	1	0.584	0.361	0.007	<0.001	<0.001
ChatGPT 4o	1	0.302	0.189	-	0.041	<0.001	0.210	0.035	0.004	<0.001	<0.001	<0.001
ChatGPT 4	0.019	0.359	0.648	0.041	-	0.004	0.710	1	0.700	0.038	0.002	0.002
ChatGPT 3.5	<0.001	0.001	0.002	<0.001	0.004	-	0.003	0.005	0.012	0.337	1	1
BR	0.077	0.832	1	0.210	0.710	0.003	-	0.208	0.327	0.029	0.002	0.002
R-1	0.017	0.327	0.584	0.035	1	0.005	0.208	-	0.805	0.091	0.005	0.005
R-2	0.004	0.152	0.361	0.004	0.700	0.012	0.327	0.805	-	0.176	0.018	0.018
Google Gemini 1.5 Pro	<0.001	<0.001	0.007	<0.001	0.038	0.337	0.029	0.091	0.176	-	0.263	0.263
Google Gemini 1.0	<0.001	<0.001	0.001	<0.001	0.002	1	0.002	0.005	0.018	0.263	-	1
Perplexity	<0.001	<0.001	0.001	<0.001	0.002	1	0.002	0.005	0.018	0.263	1	-
Microsoft Copilot	<0.001	0.002	<0.001	<0.001	0.035	0.522	0.007	0.037	0.100	0.860	0.607	0.607

LLMs, large language model; BR, breast radiologist; R-1, general radiologist 1; R-2, general radiologist.

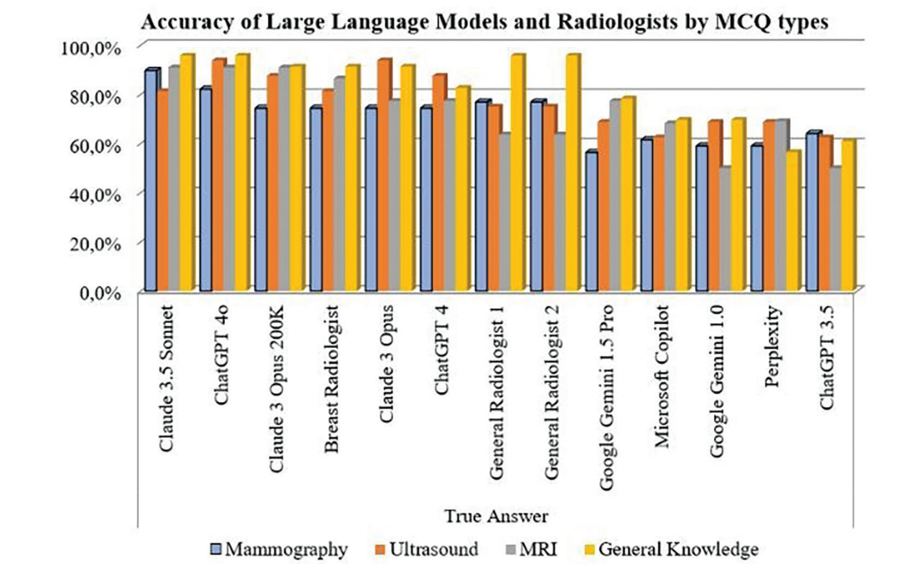


Figure 5. Accuracy of large language models and radiologists by multiple-choice question types. MCQs, multiple-choice questions; MRI, magnetic resonance imaging.

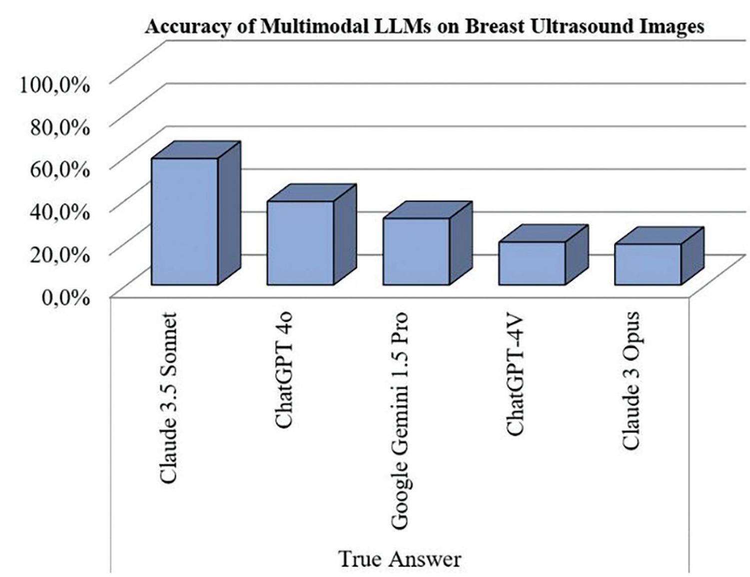


Figure 6. Accuracy of multimodal large language models on breast ultrasound images. LLMs, large language models.

formance in evaluating real-life case images is not as successful. Multimodal LLMs fall short compared with their text-based counterparts. Considering that real clinical cases are often complex and diagnoses are made through visual assessment by physicians, multimodal LLMs have not yet demonstrated sufficient performance to be used as clinical decision support systems in real-world settings.

Claude 3.5 Sonnet demonstrated the highest accuracy rate, achieving 90% in answering BI-RADS Atlas 5th edition questions. Following closely were ChatGPT 4o and Claude 3 Opus 200k with accuracy rates of 89% and 84%, respectively, while ChatGPT 4 achieved an accuracy rate of 79%. Among the radiologists, the breast radiologist exhibited the best performance with an accuracy rate of 82%, followed by general radiologist 1 with 78%, and general radiologist 2 with 76%. Claude 3.5 Sonnet demonstrated superior performance compared with all other LLMs, except for ChatGPT 4o and Claude 3 Opus models ($P < 0.05$). The performance of Claude 3.5 Sonnet and ChatGPT 4o did not show a significant difference from that of the breast radiologist ($P > 0.05$), but it notably outperformed both general radiologists ($P < 0.05$).

No statistically significant difference was found between ChatGPT 4o, Claude 3 Opus 200k, Claude 3 Opus, and ChatGPT 4 ($P > 0.05$). These LLMs, along with both the breast and general radiologists, performed significantly better than ChatGPT 3.5, Google Gemini 1.5 Pro, Google Gemini 1.0, and Perplexity ($P < 0.05$).

While interpreting real-life breast ultrasound images, Claude 3.5 Sonnet achieved an accuracy rate of 59%, ChatGPT 4o 39%, Google Gemini 1.5 Pro 31%, ChatGPT 4V 20%, and Claude 3 Opus 19%. Claude 3.5 Sonnet outperforms all the other multimodal LLMs ($P < 0.05$). The diagnostic performance of multimodal LLMs significantly differs with the BI-RADS category, except Claude 3 Opus. Claude 3.5 Sonnet (85%) and Chat GPT 4V (45%) showed superior performance in the BI-RADS 5 category ($P = 0.001$), while Google Gemini 1.5 Pro showed a higher accuracy rate (65%) for BI-RADS 2 questions ($P = 0.007$). Gemini 1.5 Pro did not correctly answer any questions in the BI-RADS 5 category, and ChatGPT 4V did not correctly answer any questions in the BI-RADS 1 category, making them the least accurate in these respective categories ($P < 0.05$).

In the Management Score, which compares the recommendations of multimodal LLMs according to BI-RADS categories, Claude 3.5 Sonnet (mean: 2.12 ± 0.97) outperformed all other multimodal LLMs ($P < 0.05$).

Notably, our study is the first to evaluate the diagnostic performance of multimodal LLMs breast radiology visual cases. Moreover, this study is the first to demonstrate the performance of the newly released Claude 3.5 Sonnet and ChatGPT 4o in breast radiology. Furthermore, there are currently no other studies that have evaluated the proficiency of different LLMs in breast radiology MCQs, both in internal comparisons and when compared with radiologists.

Multimodal LLMs, such as Claude 3.5 Sonnet and ChatGPT 4o, may perform better than a breast radiologist on text-based questions, but they can make critical errors when questions involve images that impact clinical management. For example, Gemini 1.5 Pro failed to recognize any cases in the BI-RADS 5 category, and Claude 3 Opus could not identify any normal images in the BI-RADS 0 category. This finding suggests that using multimodal LLMs without an experienced radiologist in clinical practice could lead to misdiagnoses, either missing critical conditions or misinterpreting normal findings as pathological.

On the other hand, the superior performance of LLMs on text-based questions compared with general radiologists suggests that they could serve as a supportive tool, especially for junior radiologists. They can aid in the correct use of BI-RADS nomenclature and proper classification.

Table 2. Accuracy rates of radiologists and LLMs by categories

			Mammography	MRI	General knowledge	*P*	
ChatGPT 4o	False True	n n	7 (17.9%) 32 (82.1%)	2 (9.1%) 20 (90.9%)	1 (4.3%) 22 (95.7%)	0.332	X^2
ChatGPT 4	False True	n n	10 (25.6%) 29 (74.4%)	5 (22.7%) 17 (77.3%)	4 (17.4%) 19 (82.6%)	0.700	X^2
ChatGPT 3.5	False True	n n	14 (35.9%) 25 (64.1%)	11 (50.0%) 11 (50.0%)	9 (39.1%) 14 (60.9%)	0.744	X^2
Claude 3.5 Sonnet	False True	n n	4 (10.3%) 35 (89.7%)	2 (9.1%) 20 (90.9%)	1 (4.3%) 22 (95.7%)	0.542	X^2
Claude Opus 3 200k	False True	n n	10 (25.6%) 29 (74.4%)	2 (9.1%) 20 (90.9%)	2 (8.7%) 21 (91.3%)	0.209	X^2
Claude Opus 3	False True	n n	10 (25.6%) 29 (74.4%)	5 (22.7%) 17 (77.3%)	2 (8.7%) 21 (91.3%)	0.193	X^2
Breast radiologist	False True	n n	10 (25.6%) 29 (74.4%)	3 (13.6%) 19 (86.4%)	2 (8.7%) 21 (91.3%)	0.364	X^2
General radiologist I	False True	n n	9 (23.1%) 30 (76.9%)	8 (36.4%) 14 (63.6%)	1 (4.3%) 22 (95.7%)	0.074	X^2
General radiologist II	False True	n n	9 (23.1%) 30 (76.9%)	5 (22.7%) 17 (77.3%)	5 (21.7%) 18 (78.3%)	0.905	X^2
Google Gemini Pro 1.5	False True	n n	17 (43.6%) 22 (56.4%)	4 (18.2%) 18 (81.8%)	7 (30.4%) 16 (69.6%)	0.235	X^2
Google Gemini 1.0	False True	n n	16 (41.0%) 23 (59.0%)	11 (50.0%) 11 (50.0%)	7 (30.4%) 16 (69.6%)	0.513	X^2
Microsoft Copilot	False True	n n	15 (38.5%) 24 (61.5%)	7 (31.8%) 15 (68.2%)	7 (30.4%) 16 (69.6%)	0.906	X^2
Perplexity	False True	n n	16 (41.0%) 23 (59.0%)	9 (40.9%) 13 (59.1%)	10 (43.5%) 13 (56.5%)	0.885	X^2

X^2, Chi-squared; LLM, large language model; MRI, magnetic resonance imaging.

Table 3. Comparison of accuracy of multimodal large language models with *P* values obtained from McNemar's test

	Claude 3.5 Sonnet	Claude 3 Opus	ChatGPT 4o	ChatGPT 4V	Google Gemini 1.5 Pro
Claude 3.5 Sonnet	-	<0.001	0.006	<0.001	<0.001
Claude 3 Opus	<0.001	-	0.003	1	0.067
ChatGPT 4o	0.006	0.003	-		0.302
ChatGPT 4V	<0.001	1	0.003	-	0.109
Google Gemini 1.5 Pro	<0.001	0.067	0.302	0.109	-

When multimodal LLMs correctly identify an image and assign an appropriate BI-RADS score, their management recommendations for patients closely align with the BI-RADS categories. Therefore, their success with text-based questions indicates that if they can visually determine the correct BI-RADS category, they are likely to provide accurate clinical management advice.

The variability in LLM text-based performance may be due to differences in training designs, such as different datasets, model architectures, and fine-tuning techniques.[20] LLMs such as Microsoft Copilot, Google Gemini 1.0, and Perplexity, which have internet access, sometimes provide arbitrary answers based on non-scientific information they reference.[21] This could explain their lower performance compared with other LLMs. ChatGPT and Claude 3 Opus models are trained on closed datasets, and it is unclear whether the BI-RADS Atlas was used in their training. Memorization may contribute to their high performance.

Several studies have explored the performance of LLMs on text-based radiology questions.[22,23] For instance, Almeida et al.[22] found that ChatGPT 4 achieved a 76% accuracy rate on mammography questions during the Brazilian radiology board examination, compared with 65% for ChatGPT 3.5. Our study showed higher accuracy rates, with ChatGPT 4 at 79% and ChatGPT 4o at 89%, suggesting that the difference in question difficulty may account for this variance. Furthermore, ChatGPT 4 demonstrated a general accuracy rate of 58.5%, surpassing that of 2nd-year radiology residents (52.8%) but falling short of 3rd-year residents (61.9%) in the ACR Diagnostic Radiology In-Training (DXIT) examination.[23] However, with only 10 breast radiology questions, the DXIT exam may not fully capture overall performance in this specialty. In contrast, our study's focus on a comprehensive set of BI-RADS Atlas

Table 4. Accuracy rates of multimodal large language models by categories

			BI-RADS-1	BI-RADS-2	BI-RADS-3	BI-RADS-4	BI-RADS-5	*P*	
Claude 3.5 Sonnet	False True	n n	13 (65.0%) 7 (35.0%)	5 (25.0%) 15 (75.0%)	12 (60.0%) 8 (40.0%)	8 (40.0%) 12 (60.0%)	3 (15.0%) 17 (85.0%)	0.004	X^2
ChatGPT 4o	False True	n n	13 (65.0%) 7 (35.0%)	7 (35.0%) 13 (65.0%)	17 (85.0%) 3 (15.0%)	14 (70.0%) 6 (30.0%)	10 (50.0%) 10 (50.0%)	0.015	X^2
ChatGPT 4V	False True	n n	20 (100.0%) 0 (0.0%)	14 (70%) 6 (30%)	19 (95.0%) 1 (5.0%)	16 (80.0%) 4 (20.0%)	11 (55.0%) 9 (45.0%)	0.002	X^2
Claude Opus 3	False True	n n	16 (80.0%) 4 (20.0%)	17 (85%) 3 (15%)	16 (80.0%) 4 (20.0%)	16 (80.0%) 4 (20.0%)	16 (80.0%) 4 (20.0%)	0.992	X^2
Google Gemini 1.5 Pro	False True	n n	10 (50.0%) 10 (50.0%)	13 (65.0%) 7 (35.0%)	14 (70.0%) 6 (30.0%)	12 (60.0%) 8 (40.0%)	20 (100%) 0 (0.0%)	0.010	X^2

BI-RADS, Breast Imaging Reporting and Data System.

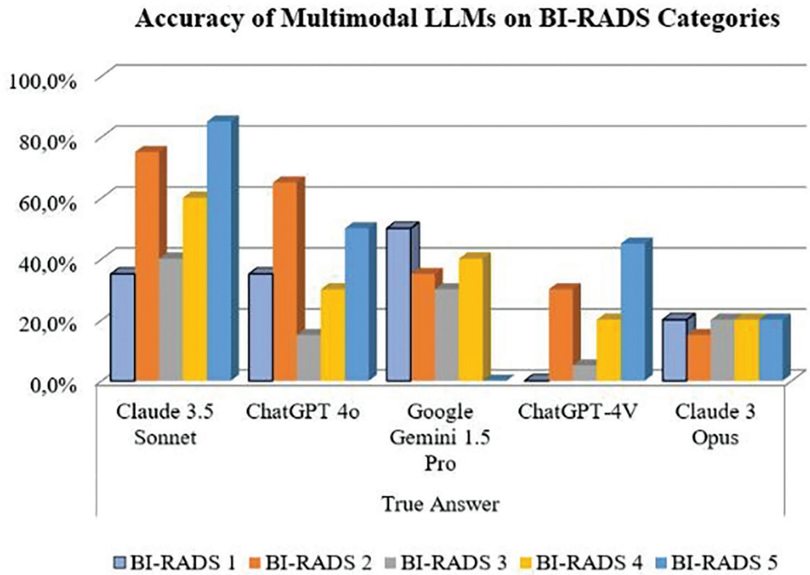


Figure 7. Accuracy of multimodal large language models on Breast Imaging Reporting and Data System (BI-RADS) categories. LLMs, large language models.

questions resulted in higher accuracy rates, underscoring that LLM performance is greatly influenced by both the specificity and quantity of the questions.

Rao et al.[9] observed that ChatGPT 4 outperformed ChatGPT 3.5 on select-all-that-apply questions related to breast pain and cancer screening, with both models performing better on these MCQs than on open-ended ones. This aligns with our findings, where the use of MCQs with a single correct answer likely contributed to the elevated success rates of LLMs. In a different context, Haver et al.[24] demonstrated ChatGPT's ability to simplify responses to frequently asked questions about breast cancer prevention and screening, achieving a 92% simplification rate. Our study, which focused on more technical and specific questions, showed that ChatGPT 4 had an accuracy rate of 79%, while ChatGPT 4o performed even better, with an accuracy rate of 89%.

When comparing the performance and readability of different LLMs, Tepe and Emekli[25] found that responses generated by Gemini 1.0 and Microsoft Copilot achieved higher readability scores ($P < 0.001$), whereas ChatGPT 4 demonstrated superior accuracy ($P < 0.001$). Our study confirmed these results, showing that ChatGPT 4 outperformed both Gemini 1.0 and Microsoft Copilot in terms of accuracy. Similarly, Griewing et al.[26] noted a 58.8% concordance between breast tumor board decisions and those generated by ChatGPT 3.5 and 4, with Sorin et al.[27] reporting a 70% agreement for ChatGPT 3.5. These findings suggest a partial alignment between LLMs and radiologists in clinical decision-making, though the variations in performance are likely due to differences in study designs and the prompts used. These studies collectively suggest that although LLMs show promise, their current performance may not yet be adequate for seamless integration into clinical decision support systems.

The challenges LLMs face in interpreting visual questions are evident in several studies.[28-30] Horiuchi et al.[30] conducted a study involving 106 musculoskeletal radiology cases, comparing the performance of ChatGPT 4 on text-based questions with ChatGPT 4V on visual questions. ChatGPT 4 correctly answered 46 out of 106 questions, significantly outperforming ChatGPT 4V, which correctly answered only 9 out of 106 ($P < 0.001$). Similarly, Dehdab et al.[28] evaluated ChatGPT 4V's performance on chest computed tomography slices across 60 different cases, including coronavirus disease-2019, non-small cell lung cancer, and control cases, finding an overall diagnostic accuracy of 56.76%, with variability depending on the case type.

In breast radiology, Haver et al.[29] compared ChatGPT 4V's performance on 151 mammography images from the ACR BI-RADS Atlas, reporting an accuracy rate of 28.5% (43/151). Although ChatGPT 4V correctly identified more than 50% of cases involving mass shape, architectural distortion, and associated features, it performed poorly on calcifications, intramammary lymph nodes, skin lesions, and solitary dilated ducts, with less than 15% correct responses.[29] In our study, ChatGPT 4V similarly showed low performance, correctly answering only 20% of breast ultrasound questions. Notably, it had an accuracy rate of 45% (9/20) for BI-RADS 5 lesions but failed to correctly identify any BI-RADS 0 lesions (0/20), indicating a tendency to misinterpret normal parenchymal tissue as pathology.

Nonetheless, as LLMs and multimodal LLMs continue to rapidly evolve and newer, more advanced models emerge, they are poised to become supportive tools for radiologists in the future. However, ethical considerations, such as ensuring patient privacy and obtaining informed consent from patients involved in the integration of LLMs into clinical decision support systems, are paramount.[31] Moreover, the lack of transparency in the decision-making mechanisms of LLMs during the diagnostic process is a significant concern.[32] Therefore, it is imperative that LLMs and multimodal LLMs are utilized under the supervision of a responsible radiologist to ensure their contribution to the diagnostic process aligns with the highest standards of patient care and safety.

An intriguing finding of our study is the notable performance of the recently introduced Claude 3.5 Sonnet, which closely rivals ChatGPT 4o. This suggests that the Claude models hold promise in the medical domain as well. Furthermore, our study contributes significantly to the existing literature by evaluating the performance of various LLMs, including both free and paid versions, alongside radiologists in the realm of breast radiology.

While our study offers valuable insights into LLMs' and multimodal LLMs' understanding of the BI-RADS Atlas, it does have limitations. First, the number of text-based questions was limited and presented in an MCQ format. Considering LLMs capacity to handle open-ended questions in real clinical scenarios, their performance may better reflect real-world situations with such questions. Further research comparing LLM performance on both open-ended and MCQs is warranted. Second, in our study evaluating multimodal LLMs' performance, we only used breast ultrasound images. Further research should include ultrasound, mammography, and MRI images to better understand the comprehensive capabilities of multimodal LLMs across different imaging modalities. Last, our study employed a single prompt to assess the performances, highlighting the need for research into the impact of different prompts and various prompt settings on LLMs' performance in breast radiology.

In conclusion, although LLMs such as Claude 3.5 Sonnet and ChatGPT 4o show potential in supporting radiologists with text-based BI-RADS assessments, their current limitations in visual diagnostics suggest that these tools should be used with caution and under the supervision of experienced radiologists to avoid misdiagnoses.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. 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]
2. 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]
3. 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):e230582. [Crossref]
4. Horiuchi D, Tatekawa H, Oura T, et al. Comparison of the diagnostic performance from patient's medical history and imaging findings between GPT-4 based ChatGPT and radiologists in challenging neuroradiology cases. *medRxiv*. 2023. [Crossref]
5. Richardson ML, Garwood ER, Lee Y, et al. Noninterpretive uses of artificial intelligence in radiology. *Acad Radiol*. 2021;28(9):1225-1235. [Crossref]
6. Suthar PP, Kounsal 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):e43958. [Crossref]
7. Güneş YC, Cesur T. Diagnostic Accuracy of Large Language Models in the European Board of Interventional Radiology Examination (EBIR) sample questions. *Cardiovasc Intervent Radiol*. 2024;47:836-837. [Crossref]
8. Diab KM, Deng J, Wu Y, Yesha Y, Collado-Mesa F, Nguyen P. Natural language processing for breast imaging: a systematic review. *Diagnostics (Basel)*. 2023;13(8):1420. [Crossref]
9. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. *J Am Coll Radiol*. 2023;20(10):990-997. [Crossref]
10. Sorin V, Glicksberg BS, Artsi Y, et al. Utilizing large language models in breast cancer management: systematic review. *J Cancer Res Clin Oncol*. 2024;150(3):140. [Crossref]
11. 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;42:685-696. [Crossref]
12. Warner-Smith M, Ren K, Mistry C, et al. Protocol for evaluating the fitness for purpose of an artificial intelligence product for radiology reporting in the BreastScreen New South Wales breast cancer screening programme. *BMJ Open*. 2024;14(5):e082350. [Crossref]
13. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*. 2023;15(2):e35179. [Crossref]
14. Sickles EA. ACR BI-RADS® Atlas, Breast imaging reporting and data system. *American College of Radiology*. 2013:39. [Crossref]
15. Ekpo EU, Ujong UP, Mello-Thoms C, McEntee MF. Assessment of interradiologist agreement regarding mammographic breast density classification using the fifth edition of the BI-RADS Atlas. *AJR Am J Roentgenol*. 2016;206(5):1119-1123. [Crossref]
16. Haver HL, Yi PH, Jeudy J, Bahl M. Use of ChatGPT to assign BI-RADS assessment categories to breast imaging reports. *AJR Am J Roentgenol*. 2024. [Crossref]
17. Cozzi A, Pinker K, Hidber A, et al. BI-RADS category assignments by GPT-3.5, GPT-4, and Google Bard: a multilanguage study. *Radiology*. 2024;311(1):e232133. [Crossref]
18. 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):826-832. [Crossref]
19. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. *Data Brief*. 2020;28:104863. [Crossref]
20. Yao Y, Duan J, Xu K, et al. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. *High-Confidence Computing*. 2024;4:100211. [Crossref]
21. Aydın Ö, Karaarslan E. Is ChatGPT leading generative AI? What is beyond expectations? *Academic Platform Journal of Engineering and Smart Systems*. 2023;11(3):118-134. [Crossref]
22. 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*. 2023;6(1):e230103. [Crossref]
23. Payne DL, Purohit K, Borrero WM, et al. Performance of GPT-4 on the American College of Radiology in-training examination: evaluating accuracy, model drift, and fine-tuning. *Acad Radiol*. 2024;31(7):3046-3054. [Crossref]
24. Haver HL, Gupta AK, Ambinder EB, et al. Evaluating the use of ChatGPT to accurately simplify patient-centered information about breast cancer prevention and screening. *Radiology Imaging Cancer*. 2024;6(2):e230086. [Crossref]
25. Tepe M, Emekli E. Assessing the responses of large language models (ChatGPT-4, Gemini, and Microsoft Copilot) to frequently asked questions in breast imaging: a study on readability and accuracy. *Cureus*. 2024;16(5):e59960. [Crossref]
26. Griewing S, Gremke N, Wagner U, Lingenfelder M, Kuhn S, Boekhoff J. Challenging ChatGPT 3.5 in senology-an assessment of concordance with breast cancer tumor board decision making. *J Pers Med*. 2023;13(10):1502. [Crossref]
27. Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer*. 2023;9(1):44. [Crossref]
28. Dehdab R, Brendlin A, Werner S, et al. Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment. *Jpn J Radiol*. 2024. [Crossref]
29. Haver HL, Bahl M, Doo FX, et al. Evaluation of multimodal ChatGPT (GPT-4V) in describing mammography image features. *Can Assoc Radiol J*. 2024:8465371241247043. [Crossref]
30. 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*. 2024. [Crossref]
31. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*. 2023;90:104512. [Crossref]
32. Tokayev KJ. Ethical implications of large language models a multidimensional exploration of societal, economic, and technical concerns. *International Journal of Social Analytics*. 2023;8(9):17-33. [Crossref]

Supplementary Material 1. MCQs

1. Which of the following BI-RADS breast density categories has the lowest sensitivity for detecting cancer on mammograms?

A) Category A

B) Category B

C) Category C

D) Category D

2. According to the BI-RADS breast density categories, in which of the following categories is mammography the most sensitive for detecting cancer?

A) Category A

B) Category B

C) Category C

D) Category D

3. Which of the following statements about BI-RADS breast density categories is true?

A) Mammography is equally sensitive for detecting cancer in all breast density categories.

B) Mammography is more sensitive for detecting cancer in women with dense breasts than in women with fatty breasts.

C) Mammography is less sensitive for detecting cancer in women with dense breasts than in women with fatty breasts.

D) There is no relationship between breast density and the sensitivity of mammography.

4. According to the BI-RADS assessment categories, which of the following categories has the highest likelihood of malignancy?

A) Category 1

B) Category 2

C) Category 3

D) Category 4

5. Which of the following BI-RADS assessment categories has a likelihood of malignancy of > 2% but <95%?

A) Category 1

B) Category 2

C) Category 3

D) Category 4

6. Which of the following management recommendations is associated with BI-RADS assessment category 3?

A) Routine mammography screening

B) Short-interval follow-up or continued surveillance mammography

C) Tissue diagnosis

D) Surgical excision

7. When is it appropriate to use BI-RADS assessment category 6?

A) When a mammographic examination is incomplete

B) When a finding is probably benign

C) When a malignancy has been biopsy-proven

D) When a finding is highly suggestive of malignancy

8. What is the management recommendation for a BI-RADS category 4 assessment?

A) Routine mammography screening

B) Short-interval follow-up or continued surveillance mammography

C) Tissue diagnosis

D) Surgical excision

9. Which of the following findings is NOT typically assessed as BI-RADS category 3?

A) Non-calcified circumscribed solid mass

B) Palpable lesion

C) Focal asymmetry

D) Solitary group of punctate calcifications

10. What is the likelihood of malignancy for a finding assessed as BI-RADS category 3?

A) Essentially 0%

B) > 0% but ≤2%

C) >2% but <95%

D) ≥ 95%

11. Which of the following is NOT a characteristically benign finding that may be assessed as BI-RADS category 2?

A) Involuting calcified fibroadenoma

B) Skin calcifications

C) Non-calcified circumscribed solid mass

D) Oil cyst

12. Which of the following findings may be described in a BI-RADS category 2 assessment?

A) Non-calcified circumscribed solid mass

B) Skin calcifications

C) Architectural distortion

D) Solitary group of punctate calcifications

13. Which of the following findings is NOT validated as being probably benign (BI-RADS category 3)?

A) Non-calcified circumscribed solid mass

B) Focal asymmetry

C) Solitary group of punctate calcifications

D) Palpable lesion

14. A screening mammogram shows unilateral axillary adenopathy with no suspicious findings in the breasts. The patient has no known infectious or inflammatory cause for the adenopathy. What should the BI-RADS® final assessment be?

A) Negative (BI-RADS® category 1)

B) Benign (BI-RADS® category 2)

C) Probably benign (BI-RADS® category 3)

D) Suspicious (BI-RADS® category 4)

15. Which of the following US descriptors for tissue composition corresponds most closely to the BI-RADS® breast density category "heterogeneously dense"?

A) Homogeneous background echotexture-fat

B) Homogeneous background echotexture-fibroglandular

C) Heterogeneous background echotexture

D) Not given in the provided text

16. Which of the following is NOT a finding that may be described in a BI-RADS® category 2 US assessment?

A) Simple cyst

B) Intramammary lymph node

C) Non-palpable solid mass

D) Postsurgical fluid collection

17. What is the recommended follow-up interval for a stable probably benign (BI-RADS® category 3) finding on US after the initial 6-month follow-up examination?

A) 3 months

B) 6 months

C) 1 year

D) 2 years

18. A US examination reveals a large axillary mass in a patient with known metastatic melanoma. The mass was previously biopsied and confirmed to be an axillary lymph node with metastatic melanoma. Except for the axillary mass, the US examination shows no abnormalities in the breast. What is the appropriate BI-RADS® assessment for this examination?

A) BIRADS 1

B) BIRADS 2

C) BIRADS 3

D) BIRADS 4

19. Which of the following is NOT a category of background parenchymal enhancement (BPE) on breast MRI?

A) Minima

B) Mild

C) Moderate

D) Severe

20. Which of the following is NOT a descriptor for the margin of a mass on breast MRI?

A) Circumscribed

B) Not circumscribed

C) Irregular

D) Rounded

21. Which of the following is NOT a modifier describing non-mass enhancement distribution?

A) Focal

B) Linear

C) Granular

D) Segmental

22. Which of the following is NOT an internal enhancement pattern for non-mass enhancement?

A) Homogeneous

B) Heterogeneous

C) Focal

D) Clumped

23. Which of the following is NOT an intracapsular silicone rupture finding on MRI?

A) Linguine sign

B) Subcapsular line

C) Keyhole sign

D) Peri-implant fluid

24. Which BI-RADS® assessment category is not recommended for screening for mammography?

A) BIRADS 1

B) BIRADS 2

C) BIRADS 3

D) BIRADS 4

25. According to BI-RADS classification, which of the following is NOT a type of asymmetry?

A) Asymmetry

B) Global asymmetry

C) Focal asymmetry

D) Diffuse asymmetry

26. Which of the following calcification morphologies should be assigned to BI-RADS® category 4C?

A) Amorphous

B) Coarse heterogeneous

C) Fine pleomorphic

D) Fine linear or fine-linear branching

27. Which of the following statements is true regarding the margin of a mass in mammography?

A) The margin must be completely well-defined for the mass to be classified as circumscribed.

B) At least 75% of the margin must be well-defined for the mass to qualify as circumscribed.

C) If any portion of the margin is indistinct, the mass should be classified as such.

D) Spiculated margins are less suspicious than microlobulated margins.

28. What is an obscured margin in mammography?

A) A margin that is completely hidden by other tissue

B) A margin that is mostly well-defined, but part of it is hidden

C) A margin that is indistinct and irregular

D) A margin that is spiculated and jagged

29. Which of the following statements about density in mammography is true?

A) Breast cancers are always lower in density than normal breast tissue.

B) Breast density is the most reliable mammographic feature of masses.

C) Breast cancers can be fat-containing.

D) The likelihood of malignancy for a high-density mass is significantly greater than that for equal- and low-density masses.

30. Which of the following is a characteristic of a fat-containing mass in mammography?

A) It is always malignant.

B) It is almost always benign.

C) It is a mixed-density mass.

D) It is associated with a high risk of breast cancer.

31. Which of the following types of calcifications is typically benign in mammography?

A) Fine and linear

B) Pleomorphic

C) Coarse or "popcorn-like"

D) Punctate

32. Which of the following statements about coarse or "popcorn-like" calcifications in mammography is true?

A) They are typically associated with breast cancer.

B) They are small and difficult to see on mammograms.

C) They are a sign of a benign breast lesion.

D) They are more common in younger women

33. Which of the following is a characteristic of large rod-like calcifications in mammography?

A) They are typically associated with breast cancer.

B) They are small and difficult to see on mammograms.

C) They are more common in younger women.

D) They follow a ductal distribution.

34. Which of the following is a characteristic of benign round calcifications in mammography?

A) They are typically clustered together.

B) They are always larger than 1 mm in size.

C) They are more common in younger women.

D) They are often formed in the acini of lobules.

35. Which of the following statements about dystrophic calcifications in mammography is true?

A) They are typically associated with breast cancer.

B) They are always smaller than 1 mm in size.

C) They are more common in younger women.

D) They are caused by radiation therapy or trauma.

36. Which of the following is a characteristic of milk of calcium calcifications in mammography?

A) They are typically associated with breast cancer.

B) They always appear as round, smudgy deposits on all mammographic projections.

C) They are more common in younger women.

D) They change shape on different mammographic projections.

37. Which of the following distributions of amorphous calcifications is suspicious and generally warrants biopsy?

A) Bilateral, diffuse

B) Grouped, linear, or segmental

C) Punctate

D) Coarse

38. What is the BI-RADS® assessment category for a single group of coarse heterogeneous calcifications, which has a positive predictive value of slightly less than 15%?

A) 4A

B) 4B

C) 4C

D) 5

39. What is the BI-RADS® assessment category for fine pleomorphic calcifications, which have a positive predictive value (PPV) of 29%?

A) 4A

B) 4B

C) 4C

D) 5

40. What is the BI-RADS® assessment category for fine linear and fine-linear branching calcifications, which have the highest PPV (70%) among suspicious calcifications?

A) 4A

B) 4B

C) 4C

D) 5

41. Which distribution of calcifications is of concern because it suggests deposits in a duct or ducts and their branches, raising the possibility of extensive or multifocal breast cancer?

A) Clustered

B) Grouped

C) Linear

D) Segmental

42. What is a possible cause of asymmetry that is visible on only one mammographic projection?

A) Summation artifacts

B) Real lesions

C) Cancer

D) Calcification

43. Where are intramammary lymph nodes frequently located?

A) Medial and lower portions of the breast

B) Lateral and usually upper portions of the breast closer to the axilla

C) Central portion of the breast

D) Posterior portion of the breast

44. Which mammographic finding is a rare finding and has been reported to be associated with non-calcified DCIS?

A) Intramammary lymph node

B) Solitary dilated duct

C) Asymmetric breast tissue

D) Architectural distortion

45. Which of the following is a sign of malignancy?

A) Bilateral nipple inversion

B) New nipple retraction

C) Stable nipple inversion for a long period of time

D) Nipple eversion

46. Which of the following is NOT a concern for skin thickening?

A) Focal or diffuse skin thickening greater than 2 mm in thickness

B) Skin thickening that represents a change from previous mammograms

C) Unilateral skin thickening after radiation therapy

D) Diffuse skin thickening with no other suspicious findings

47. Why did the fifth edition of BI-RADS® remove ranges of percentage dense tissue for the four density categories?

A) To emphasize the text descriptions of breast density

B) To indicate that percentage breast density is not associated with changes in mammographic sensitivity

C) To indicate that percentage breast density is more important than text descriptions of breast density for breast cancer risk assessment

D) To simplify the BI-RADS® reporting system

48. What is the key difference between an asymmetry and a mass on a mammogram?

A) Asymmetry has concave-outward borders, while a mass has convex-outward borders.

B) Asymmetry is unilateral, while a mass can be bilateral.

C) Asymmetry is interspersed with fat, while a mass is not.

D) Asymmetry is less conspicuous than a mass.

49. Why was the shape descriptor "lobular" eliminated in the 2013 edition of BI-RADS®?

A) Because it was redundant with the margin descriptor "microlobulated"

B) Because it was always associated with benign masses

C) Because it was difficult to distinguish from other shape descriptors

D) Because it was not a reliable indicator of malignancy

50. Which of the following is a key difference between "round" and "punctate" calcifications in the 2013 edition of BI-RADS®?

A) Round calcifications are typically benign, while punctate calcifications may be associated with malignancy

B) Punctate calcifications are defined as particles <0.5 mm, while round calcifications are defined as particles ≥0.5 mm

C) Round calcifications are typically isolated, while punctate calcifications are typically grouped

D) Punctate calcifications are more common in younger women, while round calcifications are more common in older women

51. Which of the following statements about coarse heterogeneous calcifications is true?

A) They are always associated with malignancy.

B) They are typically benign when present as multiple bilateral groupings.

C) They are larger than dystrophic calcifications.

D) They are more likely to be malignant when they occur together with fine pleomorphic calcifications.

52. What percentage of cases of developing asymmetry are found to be malignant?

A) 5%

B) 15%

C) 25%

D) 35%

53. What is the range of likelihood of malignancy for findings placed in BI-RADS® category 4A?

A) >2% TO ≤10%

B) >10% TO ≤20%

C) >20% TO ≤30%

D) >30% TO ≤50%

54. Which of the following findings is an example of a category 4A finding in BI-RADS®?

A) A circumscribed solid mass with smooth margins

B) A partially (<75%) circumscribed solid mass with US features suggestive of a fibroadenoma

C) A mass with spiculated margins and heterogeneous internal echogenicity

D) A cluster of irregular microcalcifications

55. Which of the following findings is an example of a category 4C finding in BI-RADS®?

A) A circumscribed solid mass with smooth margins

B) A partially circumscribed solid mass with US features suggestive of a fibroadenoma

C) A new indistinct, irregular solid mass

D) A cluster of punctate microcalcifications

56. When is BI-RADS® category 6 used?

A) When a tissue diagnosis of malignancy has been made but prior to complete surgical excision

B) When a biopsy is recommended for a suspicious lesion

C) When a benign lesion is found on imaging

D) When a patient has a history of breast cancer

57. What is the primary use of BI-RADS® category 0?

A) To indicate a finding that is highly suggestive of malignancy

B) To indicate the recommendation for additional imaging evaluation

C) To indicate the presence of a benign lesion

D) To indicate the need for a biopsy

58. Which of the following is a common mammographic finding associated with gynecomastia?

A) A circumscribed solid mass with smooth margins

B) A cluster of round calcifications

C) A “flame-shaped” area of increased density extending postero-laterally from the nipple

D) A spiculated mass with heterogeneous internal echogenicity

59. Which of the following is recommended by the ACR Practice Guideline for the Performance of a Breast Ultrasound Examination (2011) for optimal US image quality?

(A) Use of a low-frequency linear array transducer

B) Use of a broad bandwidth linear array transducer with a center frequency of at least 10 MHz

C) Use of a handheld, high-frequency breast US system

D) Use of a system with a low-resolution imaging capability

60. What is an important consideration when setting the field of view (FOV) on a breast ultrasound examination?

A) The FOV should be deep enough to include the pleura and lung.

B) The FOV should be set to a shallower depth when a lesion is found.

C) The FOV should be set deeply enough to include breast tissue and the pectoralis muscle posterior to it.

D) The FOV should be set to a very narrow depth to improve image resolution.

61. Which of the following is the correct method for taking measurements of a breast lesion on ultrasound?

A) Take two measurements from the same view, and take the third measurement from a view that is perpendicular to the first two.

B) Take three measurements from the same view, with each measurement representing a different plane.

C) Take two measurements from the same view, and take the third measurement from a view that is parallel to the first two.

D) Take three measurements from different views, with each measurement representing the longest axis of the lesion.

62. Which of the following is NOT a type of margin that can be used to characterize a mass on ultrasound?

A) Circumscribed

B) Indistinct

C) Spiculated

D) Irregular

63. What is the key feature of an indistinct margin on ultrasound?

A) The margin is clearly demarcated from the surrounding tissue.

B) The margin is poorly defined and blends into the surrounding tissue.

C) The margin is spiculated and irregular.

D) The margin is angular and has sharp corners.

64. What is the normal skin thickness in the periareolar area and inframammary folds on ultrasound?

A) Up to 2 mm

B) Up to 3 mm

C) Up to 4 mm

D) Up to 5 mm

65. Which of the following is a characteristic of edema on ultrasound?

A) Increased echogenicity of the surrounding tissue

B) Decreased echogenicity of the surrounding tissue

C) A mass-like appearance

D) Calcifications

66. Which of the following is a standardized descriptor for lesion stiffness on ultrasound elastography?

A) Soft

B) Intermediate

C) Hard

D) All of the above

67. What is the key difference between a "complicated cyst" and a "complex cystic and solid" mass on ultrasound?

A) The presence of internal echoes

B) The presence of septations

C) The presence of a discrete solid component

D) The size of the mass

68. Which of the following is a common benign mass that can be found in or on the skin on ultrasound?

A) Sebaceous cyst

B) Metastasis

C) Cancer

D) Abscess

69. What is the characteristic ultrasound appearance of extravasated silicone or silicone gel bleed?

A) A well-defined mass with posterior acoustic shadowing

B) A cystic mass with internal echoes

C) An echogenic mass with a "snowstorm" appearance

D) A hypoechoic mass with indistinct margins

70. Which of the following is NOT a US descriptor for tissue composition?

A) Homogeneous background echotexture-fat

B) Homogeneous background echotexture-fibroglandular

C) Heterogeneous background echotexture

D) Coarse background echotexture

71. According to the BI-RADS fifth edition, what is the correct term for a mass that contains solid and cystic components on ultrasound?

A) Complex mass

B) Complicated mass

C) Complex cystic and solid mass

D) Cystic mass

72. Which type of calcification is typically associated with an involuting fibroadenoma?

A) Fine linear and branching

B) Round

C) Coarse or "Popcorn-Like"

D) Amorphous

73. What is the recommended time point for assessing breast parenchymal enhancement (BPE) on breast MRI?

A) 2 minutes

B) 5 minutes

C) 90 seconds

D) 15 minutes

74. Which of the following is NOT a characteristic of breast parenchymal enhancement (BPE) on breast MRI?

A) Occurs regardless of menstrual cycle or menopausal status

B) Directly related to the amount of fibroglandular tissue

C) Evaluated with respect to the amount of fibroglandular tissue

D) May demonstrate progressive enhancement over time

75. Which of the following is a consideration when scheduling a breast MRI for elective examinations?

A) Scheduling the patient early in her menstrual cycle to minimize background enhancement

B) Scheduling the patient late in her menstrual cycle to maximize breast enhancement

C) Avoiding the use of contrast agents in pre-menopausal women

D) Performing the MRI regardless of the menstrual cycle or menstrual status

76. Which of the following features of a focus on breast MRI is suggestive of malignancy?

A) Not unique compared to the BPE

B) Bright on bright-fluid imaging

C) Washout kinetics

D) Persistent kinetics

77. Which of the following is a suggestive feature of a fibroadenoma on breast MRI?

A) Enhancing internal septations

B) Non-enhancing dark internal septations

C) Washout kinetics

D) Irregular shape

78. Which of the following is a cause of a false-positive interpretation of a rim-enhancing lesion on contrast-enhanced ultrasound?

A) Galactocele

B) Fat necrosis

C) Fibroadenoma

D) Malignant tumor

79. Which of the following internal enhancement patterns of non-mass enhancement (NME) is suggestive of malignancy?

A) Homogeneous

B) Heterogeneous

C) Clumped

D) Clustered ring

80. What is the primary factor used to determine the second phase of a contrast-enhanced lesion on MRI?

A) Initial-phase enhancement pattern

B) Delayed-phase enhancement pattern

C) Lesion morphology

D) Lesion size

81. Which of the following delayed-phase enhancement patterns is most commonly associated with malignant lesions?

A) Persistent

B) Plateau

C) Washout

D) Mixed

82. Which of the following is a potential cause of asymmetric breast parenchymal enhancement (BPE) on contrast-enhanced MRI?

A) Radiation therapy

B) Menstrual cycle

C) Age

D) Menopausal status

83. What is the criterion for classifying the initial phase of enhancement on contrast-enhanced MRI?

A) Percent increase in signal intensity compared to precontrast image

B) Time to peak enhancement

C) Shape of the enhancement curve

D) Type of contrast agent used

84. Which of the following is NOT a characteristic of intracapsular silicone rupture?

A) Linguine sign

B) Intraparenchymal oil cyst

C) Subcapsular line

D) Keyhole sign

85. How can you differentiate between a focal bulge in an intact breast implant and extruded silicone from an extracapsular rupture on MRI?

A) The focal bulge will have signs of intracarpsular rupture on MRI, while the extracapsular rupture will not.

B) The extracapsular rupture will have signs of intracapsular rupture inside the implant on MRI, while the focal bulge will not.

C) The focal bulge will be located on the outer edge of the implant, while the extracapsular rupture will be located in the center of the implant.

D) The extracapsular rupture will be larger than the focal bulge.

86. What is the appearance of a subcapsular line on MRI in an intracapsular silicone implant rupture?

A) A dark line paralleling the implant edge

B) A white line paralleling the implant edge

C) A dark line perpendicular to the implant edge

D) A white line perpendicular to the implant edge

87. What are the four categories used to describe the amount of background enhancement on contrast-enhanced MRI?

A) None, minimal, moderate, marked

B) Minimal, mild, moderate, marked

C) Mild, moderate, marked, severe

D) None, mild, moderate, severe

88. Which of the following statements about breast parenchymal enhancement (BPE) on contrast-enhanced MRI is true?

A) BPE is only seen in patients with dense breasts.

B) BPE can occur regardless of the menstrual cycle or menopausal status of the patient.

C) BPE is always related to the amount of fibroglandular parenchyma present.

D) Younger patients with dense breasts are less likely to demonstrate BPE than older patients with dense breasts.

89. What is the key distinguishing feature of a focus on contrast-enhanced breast MRI?

A) It is a small, punctate enhancing dot that is non-specific

B) It is a small, punctate enhancing dot that shows washout kinetics

C) It is a non-enhancing dot that corresponds to a precontrast finding

D) It is a small, punctate enhancing dot that is separated by intervening normal breast parenchyma

90. What type of enhancement pattern on MRI most closely resembles the pleomorphic pattern on mammography?

A) Punctate

B) Linear

C) Clumped

D) Regional

91. What is the appropriate BI-RADS® assessment for isolated unilateral axillary adenopathy in the absence of a known infectious or inflammatory cause?

A) Benign (category 2)

B) Probably benign (category 3)

C) Suspicious (category 4)

D) Malignant (category 5)

92. What is the standard term for a mammographic view that is angled toward the axilla?

A) Craniocaudal view

B) Mediolateral oblique view

C) Lateral view

D) Axillary view

93. What is the standard abbreviation for a tangential mammographic view?

A) TAN

B) CV

C) XCCM

D) XCCL

94. What is not the standard abbreviation for mammography views?

A) MLO75

B) SIO

C) FL

D) XCCL

95. What is not the standard abbreviation for a step-oblique view?

A) MLO15

B) MLO45

C) MLO75

D) MLO90

96. Which of the following is NOT a type of fat-containing lesion that can be seen on a mammogram?

A) Oil cyst

B) Lipoma

C) Galactocele

D) Fibroadenoma

97. What is the standard abbreviation for a superolateral-to-inferomedial oblique view?

A) IOS

B) SOI

C) SIO

D) ISO

98. What is the recomended term for ultrasound regarding special cases?

A) Clustered microcyst

B) Fibrocystic changes

C) Clustered fibrocyst

D) Microcyst

99. What is the recomended term for implants regarding location?

A) Postglandular

B) Postpectoral

C) Retroglandular

D) Glandular

100. Which year was the fifth edition of the BI-RADS® Atlas released?

A) 2013

B) 2010

C) 2008

D) 2007

Supplementary Material 2. Accuracy rates of multimodal LLMs on BI-RADS categories

			BIRADS-1	BIRADS-2	BIRADS-3	BIRADS-4	BIRADS-5	*P*	
Claude 3.5 Sonnet	False	n	13 (65.0%)	5 (25.0%)	12 (60.0%)	8 (40.0%)	3 (15.0%)	0.004	X^2
	True	n	7 (35.0%)	15 (75.0%)	8 (40.0%)	12 (60.0%)	17 (85.0%)		
ChatGPT 4o	False	n	13 (65.0%)	7 (35.0%)	17 (85.0%)	14 (70.0%)	10 (50.0%)	0.015	X^2
	True	n	7 (35.0%)	13 (65.0%)	3 (15.0%)	6 (30.0%)	10 (50.0%)		
ChatGPT 4V	False	n	20 (100.0%)	14 (70%)	19 (95.0%)	16 (80.0%)	11 (55.0%)	0.002	X^2
	True	n	0 (0.0%)	6 (30%)	1 (5.0%)	4 (20.0%)	9 (45.0%)		
Claude Opus 3	False	n	16 (80.0%)	17 (85%)	16 (80.0%)	16 (80.0%)	16 (80.0%)	0.992	X^2
	True	n	4 (20.0%)	3 (15%)	4 (20.0%)	4 (20.0%)	4 (20.0%)		
Gemini 1.5 Pro	False	n	10 (50.0%)	13 (65.0%)	14 (70.0%)	12 (60.0%)	20 (100%)	0.010	X^2
	True	n	10 (50.0%)	7 (35.0%)	6 (30.0%)	8 (40.0%)	0 (0.0%)		

X^2, chi-square; LLMs, large language models; BI-RADS, Breast Imaging Reporting and Data System.

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242972

CHEST IMAGING

ORIGINAL ARTICLE

Automatic machine learning accurately predicts the efficacy of immunotherapy for patients with inoperable advanced non-small cell lung cancer using a computed tomography-based radiomics model

Siyun Lin[1,2*]
Zhuangxuan Ma[3*]
Yuanshan Yao[4]
Hou Huang[2]
Wufei Chen[3]
Dongfang Tang[1]
Wen Gao[1]

[1]Huadong Hospital, Fudan University, Department of Thoracic Surgery, Shanghai, China

[2]Shanghai Key Laboratory of Clinical Geriatric Medicine, Shanghai, China

[3]Huadong Hospital, Fudan University, Department of Radiology, Shanghai, China

[4]Shanghai Chest Hospital, Shanghai JiaoTong University School of Medicine, Department of Thoracic Surgery, Shanghai, China

***These authors are joint first authors.**

Corresponding authors: Wufei Chen, Dongfang Tang, Wen Gao

E-mails: chenwufei_2008@163.com, tangdongfangchest@163.com, gaowenchest@163.com

Received 08 August 2024; revision requested 13 September 2024; last revision received 12 October 2024; accepted 18 November 2024.

Epub: 16.01.2025

Publication date: 03.03.2025

DOI: 10.4274/dir.2024.242972

PURPOSE

Patients with advanced non-small cell lung cancer (NSCLC) have varying responses to immunotherapy, but there are no reliable, accepted biomarkers to accurately predict its therapeutic efficacy. The present study aimed to construct individualized models through automatic machine learning (autoML) to predict the efficacy of immunotherapy in patients with inoperable advanced NSCLC.

METHODS

A total of 63 eligible participants were included and randomized into training and validation groups. Radiomics features were extracted from the volumes of interest of the tumor circled in the preprocessed computed tomography (CT) images. Golden feature, clinical, radiomics, and fusion models were generated using a combination of various algorithms through autoML. The models were evaluated using a multi-class receiver operating characteristic curve.

RESULTS

In total, 1,219 radiomics features were extracted from regions of interest. The ensemble algorithm demonstrated superior performance in model construction. In the training cohort, the fusion model exhibited the highest accuracy at 0.84, with an area under the curve (AUC) of 0.89–0.98. In the validation cohort, the radiomics model had the highest accuracy at 0.89, with an AUC of 0.98–1.00; its prediction performance in the partial response subgroup outperformed that in both the clinical and radiomics models. Patients with low rad scores achieved improved progression-free survival (PFS); (median PFS 16.2 vs. 13.4, $P = 0.009$).

CONCLUSION

autoML accurately and robustly predicted the short-term outcomes of patients with inoperable NSCLC treated with immune checkpoint inhibitor immunotherapy by constructing CT-based radiomics models, confirming it as a powerful tool to assist in the individualized management of patients with advanced NSCLC.

CLINICAL SIGNIFICANCE

This article highlights that autoML promotes the accuracy and efficiency of feature selection and model construction. The radiomics model generated by autoML predicted the efficacy of immunotherapy in patients with advanced NSCLC effectively. This may provide a rapid and non-invasive method for making personalized clinical decisions.

KEYWORDS

Advanced non-small cell lung cancer, immunotherapy, radiomics, automatic machine learning, models

You may cite this article as: Lin S, Ma Z, Yao Y, et al. Automatic machine learning accurately predicts the efficacy of immunotherapy for patients with inoperable advanced non-small cell lung cancer using a computed tomography-based radiomics model. *Diagn Interv Radiol.* 2025;31(2):130-140.

Non-small cell lung cancer (NSCLC) is a prevalent and malignant tumor with high incidence and mortality rates globally.[1] Over 30% of new NSCLC cases are diagnosed at locally advanced stages [tumor–node–metastasis (TNM) stage III]. The absence of notable early symptoms often leads to diagnoses at advanced stages or after local metastasis has occurred, which frequently delays surgical treatment.

The current standard treatment for patients with advanced NSCLC involves concurrent chemoradiotherapy followed by immunotherapy.[2] Definitive efficacy and improved prognoses have been achieved in all stages of NSCLC with the use of immune checkpoint inhibitors (ICIs), either alone or in combination with chemotherapy.[3,4] In the CHECKMATE-816 clinical trial, nivolumab combined with chemotherapy extended event-free survival (EFS) by 10.8 months and decreased the risk by 37% compared with the control group [hazard ratio (HR) 0.63, confidence interval (CI): 0.43–0.91, P = 0.0052].[5] Furthermore, the recent NEOTORCH trial reported a similar extension in EFS and a significantly higher pathological complete response (CR) rate (24.8% vs. 1.0%, $P < 0.0001$) in the group receiving combined immune-chemotherapy.[6] However, in the Pacific trial (NCT02125461), only one-third of patients who received adjuvant therapy with durvalumab remained disease-free after 5 years,[7,8] indicating that immunotherapy may not be suitable for all patients due to factors such as the specific tumor immune microenvironment, residual toxicity, and societal expense. Effective immunotherapy is often positively correlated with high programmed death-ligand 1 (PD-L1) expression and the tumor mutation burden (TMB), but these require tissue from biopsies for detection. The challenge of not being able to perform repeated biopsies after developing chemo-resistance complicates treatment options for patients at an advanced stage. Therefore, there is an urgent need to develop non-invasive methods to accurately predict the efficacy of immunotherapy, which could benefit a broader group of patients.

In recent years, thin-slice computed tomography (CT) scans have become integral in diagnosing and staging NSCLC.[9,10] With advancements in medical imaging, there has been a transition from traditional qualitative diagnosis to the extraction of multimodal image data for quantitative analysis. Radiomics, a promising tool in image analysis, allows for the extraction of high-throughput features from imaging data. These features, combined with specific modeling techniques, can enhance the accuracy of disease diagnosis, differentiation, and prognosis evaluation.[11] Previously, we developed and implemented delta radiomics diagnostic features to refine and personalize the diagnosis of invasive adenocarcinoma in lung partial solid nodules.[12]

Automatic machine learning (autoML) algorithms have facilitated the analysis of complex, large-sample data into predictive models and automated classifications. By integrating substantial amounts of data from radiology, pathology, genomics, and proteomics, autoML has enhanced clinical decision-making.[13] In the present study, we aimed to identify effective radiomics features in CT images using autoML and integrate them with clinical features to develop a fusion model for individualized efficacy prediction and progression assessment in patients with advanced NSCLC receiving immunotherapy.

Main points

- Radiomics modeling based on computed tomography images predicted the efficacy of immunotherapy in patients with advanced non-small cell lung cancer effectively.
- Automatic machine learning can integrate multiple algorithms to obtain improved predictive capabilities.
- The diagnostic performance of the radiomics model outperformed that of the clinical model.
- Patients with lower rad scores achieved superior progression-free survival.

Methods

Study design and population

In this retrospective observational single-center study, we reviewed patients with NSCLC who underwent ICI treatment at Huadong Hospital between January 2020 and December 2022. The inclusion criteria were as follows: (1) >18 years; (2) receiving ICI treatment (anti-PD-1/PD-L1) at Huadong Hospital for the first time; (3) a clinically confirmed diagnosis of unresectable locally advanced stage NSCLC [stage III–IV, Union for International Cancer Control/American Joint Committee on Cancer (8th edition)]; and (4) available thin-slice CT images (1–1.25 mm), with lesions delineated and evaluated. The exclusion criteria were as follows: (1) a pathologically confirmed diagnosis of small cell lung cancer; (2) a history of malignancies other than NSCLC; (3) poor CT image quality with artifacts; and (4) failure to extract radiomics features due to other reasons.

Finally, a total of 63 eligible cases were enrolled (Figure 1). The clinical features before receiving ICIs were collected from medical records, including age, gender, smoking history, the time of diagnosis, pathological type, tumor location, the maximum diameter of the primary tumor site, clinical tumor stage, metastatic location, driver gene mutation, the start time and type of ICI treatment, treatment regimen, and disease progression and survival information. The efficacy evaluation

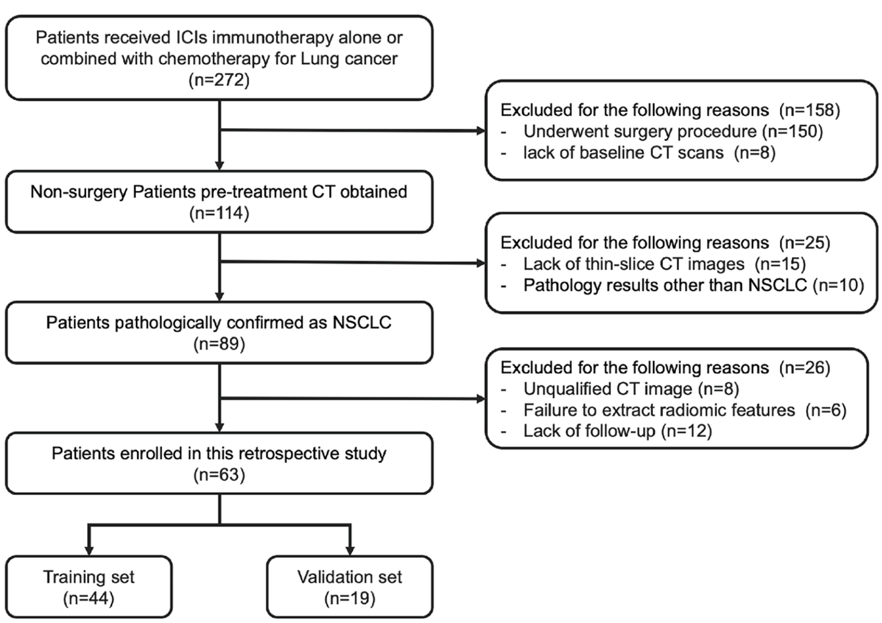


Figure 1. Study flowchart. ICIs, immune checkpoint inhibitors; CT, computed tomography; NSCLC, non-small cell lung cancer.

was based on the immune-related response evaluation criteria in solid tumors,[14] which classifies outcomes as CR, partial response (PR), stable disease (SD), and PD. The disease control rate (DCR) refers to the sum of all patients who were CR, PR, and SD. All the enrolled cases were further separated into a training and a validation cohort randomly after adjusting for potential confounders. The study was approved by the Ethics Committee of Huadong Hospital, and the requirement for informed consort was waived (approval no.: 2022K033, date: 21.02.2022).

Computed tomography image acquisition

The patients in this study were all subjected to non-contrast-enhanced CT performed on two scanners: a Somatom Definition Flash scanner (Siemens Medical Solutions, Erlangen, Germany) and a GE Discovery CT750 HD scanner (GE Healthcare, MO, USA) at 120 kV. The detailed scanning parameters are shown in Supplementary Table 1. The overall scanning range was from the lung apex to the bilateral adrenal gland. During the examination, the patients were instructed to lie in a supine position and inhale deeply with both arms raised.

Target segmentation and radiomics features extraction

According to the target lesions on the axial slices of the initial CT scans, the volumes of interest (VOIs) were manually marked by two experienced radiologists, each with 5 years' expertise in diagnosing chest CT images, to achieve three-dimensional (3D) segmentation using the open-source 3D Slicer software (version 4.13.0; National Institutes of Health).

The extraction of radiomic features from these tumor VOIs was automatically performed using pyRadiomics (version: 3.0.1).[15] To assess the inter-rater reliability between the radiologists, the intraclass correlation coefficient (ICC) was employed, with ICC >0.75 indicating a high level of agreement. The types of radiomic features extracted included grayscale, shape, texture, and wavelet transform features.

Feature selection and model construction

Due to the broad variability in the initial dataset, the data underwent normalization to control the radiomics features within a standardized intensity range. Feature selection was performed within the training cohort. The MLJAR platform, an open-source software based on Python, was employed for predictive feature selection and modeling.[16] This platform is designed to automatically address missing data by implementing strategies such as mean or median imputation to maintain data integrity. It also manages categorical variables by automatically performing encoding transformations, such as one-hot encoding or label encoding, enabling machine learning algorithms to effectively interpret these features. Subsequently, a feature engineering step was undertaken to create "golden features" that possess enhanced predictive power, derived from the original dataset features through operations such as addition, subtraction, multiplication, and division. Throughout the training phase, MLJAR assessed the significance of each feature using techniques such as permutation importance or SHapley Additive exPlanations, providing a quantitative measure of each

Table 1. Basic characteristics of the enrolled patients in the training cohort and validation cohorts

	Training cohort (n = 44)					Validation cohort (n = 19)				
	Total	PR	SD	PD	*P*	Total	PR	SD	PD	*P*
Age										
<60	11	3	3	5	0.484	7	4	0	3	0.731
≥60	33	12	4	17		12	6	1	5	
Gender										
Male	34	14	4	16	0.13	10	6	1	3	0.396
Female	10	1	3	6		9	4	0	5	
Pathological type										
LSCC	13	8	2	3	0.034*	13	7	1	5	0.739
LUAD	31	7	5	19		6	3	0	3	
Tumor location										
Right	30	10	5	15	0.975	13	7	0	6	0.311
Left	14	5	2	7		6	3	1	2	
cT stage										
T1–T2	22	5	3	14	0.179	9	4	1	4	0.509
T3–T4	22	10	4	8		10	6	0	4	
cN stage										
N0	9	2	2	5	0.663	1	0	0	1	0.484
N+	35	13	5	17		18	10	1	7	
cM stage										
M0–M1a	30	11	6	13	0.365	13	8	1	4	0.311
M1b–M1c	14	4	1	9		6	2	0	4	
cTNM stage										
III	11	5	2	4	0.563	4	3	1	0	0.041*
IV	33	10	5	18		15	7	0	8	
Driver gene mutation										
Negative	35	13	6	16	0.533	15	7	1	7	0.577
Positive	9	2	1	6		4	3	0	1	
Smoking status										
Never	22	7	5	10	0.464	12	7	0	5	0.383
Ex- or current	22	8	2	12		7	3	1	3	
Treatment										
Without CHT	9	1	3	5	0.137	5	2	1	2	0.221
With CHT	35	14	4	17		14	8	0	6	
PD-L1 expression										
<50%	33	11	6	16	0.573	11	6	1	4	1.000
≥50%	11	6	1	4		8	5	0	3	

*Means statistical significance existed (Fisher exact probability test, $P < 0.05$). PR, partial response; SD, stable disease; PD, progressive disease; LSCC, lung squamous cell carcinoma; LUAD, lung adenocarcinoma; CHT, chemotherapy; cT stage, clinical tumor stage; cN stage, clinical node stage; cM stage, clinical metastasis stage; cTNM stage, clinical tumor-node-metastasis stage; PD-L1, programmed death-ligand 1.

feature's impact on the model's predictive accuracy and offering insight into the underlying decision-making processes of the model.

Afterward, in the "competition" mode of MLJAR, the software sought the most effective algorithms from a range, including linear regression, light gradient-boosting machine (LightGBM), eXtreme gradient boosting, neural networks (NN), and random forest (RF). Additionally, it considered assembling multiple algorithms to finalize the modeling process. The rad score was obtained by multiplying the coefficients of each feature by its value and then summing the results to get the final value.

The predictive model, which included clinical, radiomics, and fusion models, was developed using the aforementioned autoML algorithms. The efficacy of each model was assessed through receiver operator characteristic (ROC) curves for both the training and validation cohorts. Subsequently, the area under the curve (AUC) was calculated to determine the predictive accuracy of each constructed model.

Statistical analysis

The feature extraction and statistical analysis procedures were conducted using R software (version 3.6.2; http://www.Rproject.org and SPSS 22 (IBM, IL, USA). Categorical variables were analyzed using Fisher's exact test. To evaluate the multi-class ROC curves, both the macro-AUC and micro-AUC were calculated. The macro-AUC averaged the AUC values from each category, whereas the micro-AUC computed the weighted average after evaluating each category independently. Furthermore, model performance was assessed using statistical metrics such as accuracy, precision, recall, and F1-score.

Model performance was evaluated by ROC analysis, and the significance level of curves was compared using the DeLong test. A COX regression analysis was utilized to investigate factors associated with disease progression and survival. Survival rates were analyzed using the Kaplan–Meier method, and survival data comparisons were conducted with the log-rank test. A two-sided *P* value less than 0.05 was considered statistically significant for all tests.

Results

Basic characteristics of patients

The basic characteristics of the patients are listed in Table 1. In total, 63 patients with advanced NSCLC who had received ICIs in our hospital were randomly divided into the training cohort (n = 44, PR: 15, SD: 7, and PD: 22) and the validation cohort (n = 19, PR: 10, SD: 1, and PD: 8) based on the efficacy evaluation (Supplementary Table 2).

In the training cohort, differences were observed in the tumor pathological types of patients with various curative effects [lung squamous cell cancer vs. lung adenocarcinoma (LUAD), 13 vs. 31, *P* = 0.034]. In the validation cohort, a difference in the clinical TNM (cTNM) stage was observed (cTNM III vs. cTNM IV, 4 vs. 15, *P* = 0.041). No differences were observed in age, gender, tumor location, driver gene mutations, smoking history, PD-L1 expression, or combination therapy among the patients (all *P* > 0.05).

Selection of radiomics and clinical golden features

The radiomics feature selection workflow is shown in Figure 2. The VOIs were automatically extracted, yielding a total of 1,219 features. Within the training cohort, the golden features, regarded as the most predictive features, were selected for the subsequent model construction by autoML. Among the radiomics features, based on the superior performance of the LightGBM algorithm, log-sigma-4-0mm_ Glrlm_ Lowgraylevelrunemphasis had the highest mean of feature importance; the top 25 golden features are listed in Supplementary Figure 1. The rad scores for patients undergoing ICI treatment were significantly lower in the DCR group than in the PD group in both the training (0.105 ± 0.284 vs. 0.502 ± 0.318, *P* < 0.001) and the validation cohorts (0.119 ± 0.224 vs. 0.528 ± 0.262, *P* = 0.002) (Supplementary Figure 2).

Among the clinical features, ten golden features were identified and selected for model building using autoML. Among these, the feature representing the combination with chemotherapy (feature 11) was identified as the most critical (Supplementary Figure 3).

Model construction and performance comparison

Based on the input of golden features with the highest importance, different learning algorithms were selected for establishing each model (Supplementary Figure 4). The ensemble algorithm demonstrated the low-

Table 2. Performance evaluation of the clinical, radiomics, and fusion models

	Training cohort			Validation cohort		
	Clinical model	Radiomics model	Fusion model	Clinical model	Radiomics model	Fusion model
Micro-AUC	0.93	0.94	0.93	0.90	0.98	0.97
Macro-AUC	0.92	0.94	0.95	0.92	0.99	0.98
Accuracy	0.80	0.77	0.84	0.74	0.89	0.84
AUC (95% CI)						
PR	0.92 0.646 to 0.953	0.92 0.614 to 0.928	0.89 0.602 to 0.913	0.88 0.474 to 0.895	0.99 0.737 to 1.000	0.96 0.698 to 0.968
SD	0.87 0.549 to 0.925	0.96 0.638 to 1.000	0.98 0.676 to 0.996	1.00 /	1.00 /	1.00 /
PD	0.96 0.664 to 0.965	0.92 0.696 to 0.928	0.91 0.688 to 0.921	0.83 0.605 to 0.934	0.98 0.605 to 1.000	0.94 0.653 to 0.956
***P* value**	0.004*	0.005*	<0.001*	0.015*	0.060	0.010*
Precision						
PR	0.87	0.80	0.80	0.80	0.90	0.90
SD	0.71	0.86	0.86	1.00	1.00	1.00
PD	0.77	0.73	0.86	0.62	0.88	0.75
Recall						
PR	0.76	0.75	0.86	0.80	1.00	0.82
SD	0.56	0.67	0.75	0.50	0.50	1.00
PD	0.94	0.84	0.86	0.71	0.88	0.86
F1-score						
PR	0.81	0.77	0.83	0.80	0.95	0.86
SD	0.85	0.75	0.80	0.67	0.67	1.00
PD	0.63	0.78	0.86	0.67	0.88	0.80

*Means statistical significance existed between the AUC values among the models (DeLong test, *P* < 0.05). AUC, area under the curve; 95% CI, 95% confidence interval; PR, partial response; SD, stable disease; PD, progressive disease.

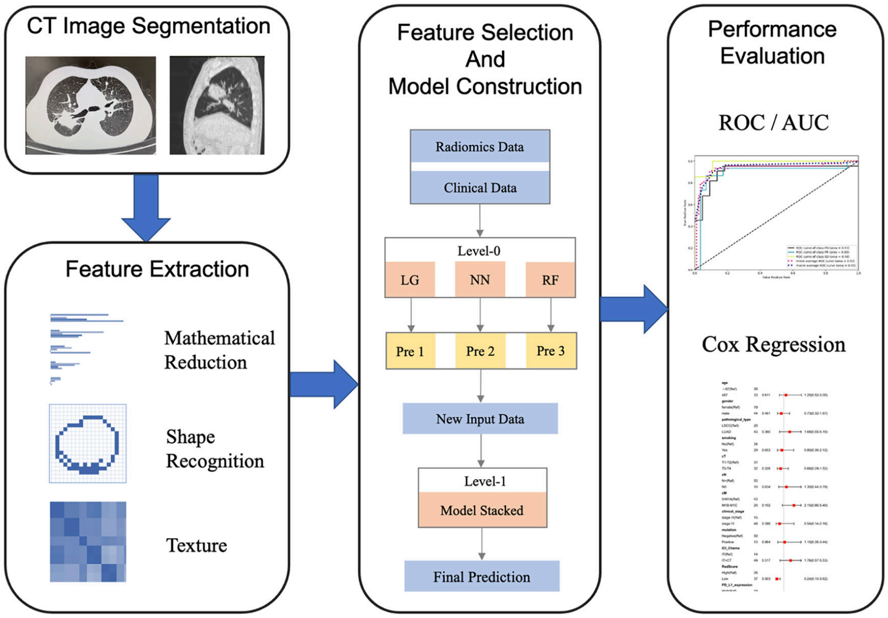


Figure 2. Workflow for the radiomics analysis. ROC, receiver operator characteristic; AUC, area under the curve.

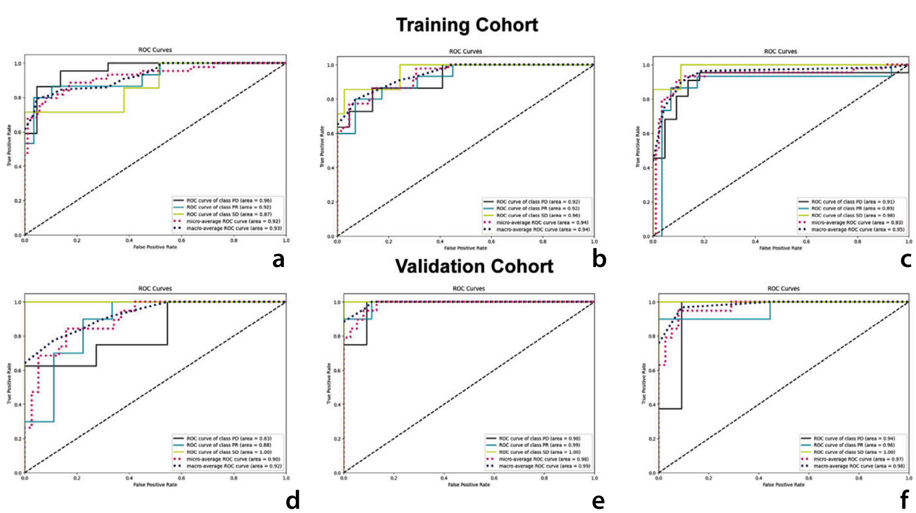


Figure 3. Evaluation of the performance of the different models. **(a)** Clinical model in the training cohort; **(b)** radiomics model in the training cohort; **(c)** fusion model in the training cohort; **(d)** clinical model in the validation cohort; **(e)** radiomics model in the validation cohort; **(f)** fusion model in the validation cohort. ROC, receiver operator characteristic.

est log-loss value in both the clinical and fusion models, indicating greater accuracy and a superior alignment between the predicted results and actual outcomes. In the radiomics model, the performance matched that of LightGBM, also suggesting improved accuracy and consistency.

Our study has shown that in both the radiomics and fusion models, the micro-AUC and macro-AUC were higher than those in the clinical model across the training and validation cohorts. In terms of accuracy, the fusion model scored the highest in the training cohort with 0.84, whereas the radiomics model outperformed the other models in the validation cohort with 0.89. In the training cohort, the radiomics and fusion models both exhibited optimal performance in SD, with an AUC of 0.96 (95% CI, 0.638–1.000) and 0.98 (95% CI, 0.676–0.996), respectively. In the validation cohort, the AUC of the radiomics model in three subgroups (PR, PD, and SD) were all higher than in the clinical and fusion models. Additionally, in the validation cohort, the PR subgroup exhibited better recall values and F1-scores than the SD and PD subgroups in both the clinical and radiomics models, suggesting enhanced predictive performance for this subgroup (Table 2, Figure 3).

Model prediction of progression-free and overall survival

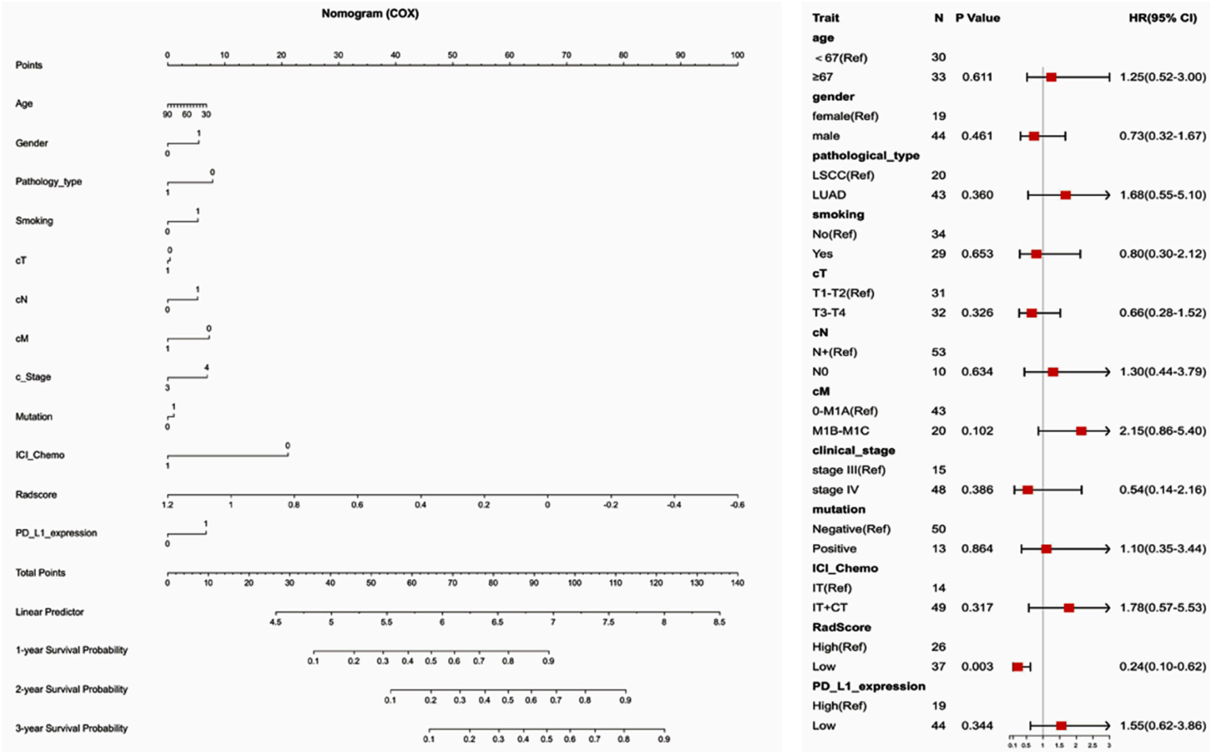
All the enrolled patients were followed up for progression-free survival (PFS) and overall survival (OS), including 30 disease-progressed cases and 8 deaths, with a median follow-up time of 20 months (range: 3–47 months). Based on a nomogram derived from the multivariate COX regression analysis, patients undergoing ICI treatment were divided into high and low rad-score groups, with a threshold of 0.3 (Figure 4a). Regression analysis confirmed that the rad score was a more accurate predictor of progression risk than clinical factors (HR: 0.25, 95% CI: 0.10–0.63, P = 0.004) (Figure 4b). Although there was no significant difference in OS between the high and low rad-score groups (20.2 vs. 21.8 months, P = 0.056), the median PFS was notably longer in the low-score group, at 16.2 months, compared with 13.4 months in the high-score group (P = 0.009) (Supplementary Figure 5). The above data suggest that patients with low rad scores, as determined by the radiomics model, tend to experience less progression following immunotherapy.

Discussion

In the present study, we developed and validated a radiomics-based model using autoML algorithms to non-invasively assess the efficacy of immunotherapy in patients with inoperable advanced NSCLC. The findings revealed that the model, which incorporates features from CT images, displayed robust capabilities for diagnostics as well as for predicting therapeutic efficacy and disease progression.

In addition to PD-L1 expression, recent studies have shown that ICIs are highly effective in patients with high microsatellite instability or deficient mismatch repair (dMMR). Tumor cells with dMMR characteristics tend to have a higher TMB, which leads to the production of a considerable number of neoantigens. These neoantigens facilitate the recruitment of lymphocytes that become tumor-infiltrating lymphocytes, inhibiting tumor growth and enhancing the efficacy of immunotherapy.[17,18] However, these markers are typically identified through pathological immunohistochemistry or next-generation sequencing analysis, which require invasive tissue sampling and are costly. Therefore, there is a need for non-invasive, cost-effective, and accurate predictive methods using radiomics.

Progress in computerized imaging technology has led to the production of high-



a b

Figure 4. Rad score reflecting the risk of progression using COX regression analysis. **(a)** Nomogram of the rad scores and clinical risk factors; **(b)** results of the COX regression analysis.

er-definition images, enhancing radiomics' ability to extract more intricate features than traditional imaging methods. This advancement supports the performance of high-dimensional quantitative analysis, providing additional insights for clinical decision-making.[19] At present, numerous researchers have developed models with refined features that demonstrate high evaluation efficacy in various NSCLC application scenarios. These models have proven effective in predicting lesion benignity and malignancy, lymph node metastasis, driver mutations, and the severity of adverse effects.[20-23] For example, Yoon et al.[24] discovered that CT imaging features could non-invasively predict PD-L1 expression, identifying that validated radiomics models had greater discriminatory power than those generated from clinical features alone in an advanced LUAD cohort. Similarly, Trebeschi et al.[25] identified a non-invasive machine learning biomarker capable of differentiating between responders and non-responders to immunotherapy, and this model achieved an AUC value of 0.83 in lung cancer studies.[24]

In all our models, the predictive performance for the PR subgroup exceeded that for the PD subgroup. These results suggest that our model aided in identifying patients who are likely to benefit from immunotherapy. However, the diagnostic consistency for the SD subgroup in the validation cohort remained uncertain due to the limited sample size. Previous studies typically focused on binary outcomes, such as categorizing responses as effective or ineffective or progressive and non-progressive, which often excluded patients in the SD subgroup. The antitumor effect in the SD subgroup is considered ambiguous, leading to no significant differences in OS compared with the PR or PD subgroups. Although fusion models are generally regarded as having superior predictive capabilities, in this study, they only excelled in the SD subgroup compared with both clinical and radiomics models alone. This occurred because the features extracted from the images, when processed by autoML, might yield diagnoses that contradict clinical features, thereby reducing the predictive accuracy of the fusion model.

In the survival analysis, variations in PFS were observed among patients with differing rad scores ($P = 0.056$), though there was no statistically significant difference in OS. This lack of significance in OS could be due to all patients being in the advanced stages of the disease (cIII–cIV) and exhibiting either lymph node or distant metastasis, both of which are associated with higher risks. In studies with smaller sample sizes and shorter follow-up times, PFS may be a more suitable endpoint than OS, although OS remains the gold standard for measuring clinical benefit.

Furthermore, a positive result in PFS does not always translate to a benefit in OS. This discrepancy can arise because the toxic side effects of a treatment might cause a statistical bias in the PFS assessment, with drugs that have higher side effects potentially showing a "false" PFS advantage during shorter follow-up periods. In this study, the high rad-score group accounted for more than half of the recurrences (median PFS: 13.8 months), whereas the low rad-score group did not reach the median PFS. Median OS was not achieved in either group. The median follow-up time was 20 months, exceeding the median PFS by 6.2 months, which may also indicate robust results.

With the progression in central processing unit and graphics processing unit technology, deep learning and autoML methods have gained popularity.[26] In the present study, various algorithms were sequentially employed to develop clinical, radiomics, and fusion models via autoML. Among these, the ensemble models that integrated multiple classifiers demonstrated superior performance. However, the radiomics model, developed using LightGBM, achieved prediction levels in the training cohort comparable to those of the ensemble model. LightGBM is a framework that implements the gradient-boosting decision tree algorithm. This algorithm is well-regarded in machine learning for its ability to iteratively train weak classifiers to derive an optimal model, notable for its efficient parallel training, improved accuracy, and capability to prevent overfitting.[27,28] In response to the characteristics of the dataset, different machine learning algorithms have demonstrated their respective performance advantages. For instance, Wiesweg et al.[29] applied support vector machine modeling to analyze RNA expression from biopsy samples in patients with advanced NSCLC, identifying seven genes predictive of immunotherapy response. Similarly, using a cytokine-based ICI response index, Wei et al.[30] employed RF modeling to predict responses to ICIs in patients with NSCLC. In the present study, we harnessed autoML to amalgamate multiple algorithms, developing models that exhibited enhanced predictive efficacy. This approach could significantly aid in predicting the effectiveness and survival outcomes of ICI treatment in patients with advanced NSCLC.

The current study has several limitations. First, being a single-center retrospective study with a small sample size in the training cohort, there is a potential impact on the specificity of the models, necessitating the collection of multicenter clinical data to confirm the models' robustness. Second, CT images were obtained from two scanning devices, which might have an adverse effect on radiomics feature extraction caused by uniformity. The MLJAR platform offers capabilities for model interpretation. As the complexity of the autoML models increases, their interpretability decreases, making it difficult for clinicians to understand and trust the model outputs, which could affect the reliability of model outcomes and the quality of decision-making. Moreover, the assessment of PD-L1 expression was limited by the amount of tissue available for fine-needle biopsy, resulting in some patients not being accurately assessed. It is also crucial in practice to select the most suitable combination of autoML algorithms, tailored to the specific characteristics of the data.

Furthermore, although the primary goal of this study was to provide surgical and oncology specialists with a predictive tool for treatment efficacy in patients with advanced NSCLC, challenges have arisen in accurately identifying lesions on CT images. To address this, Jiang et al.[31] developed a multi-scale convolutional NN method that integrates features from different resolutions to segment lung tumors accurately, facilitating the precise and automated tracking of tumor volumes. Integrating similar diagnostic models could enhance the utility of autoML in clinical settings. Moreover, although autoML allows for the training of numerous deep learning models with minimal coding or data input, the performance of these models can vary, and there remains room to improve both efficiency and prediction accuracy. Models that are designed and refined by experts may prove more reliable, and further clarification is needed on their clinical relevance and guidelines for practical diagnosis and treatment.

In conclusion, autoML has the ability to accurately predict the efficacy of immunotherapy and the short-term prognosis of patients with inoperable advanced NSCLC by constructing CT-base radiomics models, aiding the clinical evaluation and screening of a broader population and the development of personalized treatment strategies.

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

The work was supported by the National Natural Science Foundation of China (62203117), the National Key R&D Program of China (2022YFF1203301), Research Project Plan of Shanghai Municipal Health Commission (20214Y0309) and Huadong Hospital Clinical Trial Project (HDLC2022012).

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin*. 2022;72(1):7-33. [CrossRef]
2. Isaka T, Ito H, Nakayama H, Yokose T, Yamada K, Masuda M. Effect of epidermal growth factor receptor mutation on early-stage non-small cell lung cancer according to the 8th TNM classification. *Lung Cancer*. 2020;145:111-118. [CrossRef]
3. Lahiri A, Maji A, Potdar PD, et al. Lung cancer immunotherapy: progress, pitfalls, and promises. *Mol Cancer*. 2023;22(1):40. [CrossRef]
4. Borghaei H, Gettinger S, Vokes EE, et al. five-year outcomes from the randomized, phase III Trials CheckMate 017 and 057: nivolumab versus docetaxel in previously treated non-small-cell lung cancer. *J Clin Oncol*. 2021;39(7):723-733. [CrossRef]
5. Forde PM, Spicer J, Lu S, et al. Neoadjuvant nivolumab plus chemotherapy in resectable lung cancer. *N Engl J Med*. 2022;386(21):1973-1985. [CrossRef]
6. Lu S, Zhang W, Wu L, et al. Perioperative toripalimab plus chemotherapy for patients with resectable non-small cell lung cancer: the neotorch randomized clinical trial. *JAMA*. 2024;331(3):201-211. [CrossRef]
7. Spigel DR, Faivre-Finn C, Gray JE, et al. Five-year survival outcomes from the PACIFIC trial: durvalumab after chemoradiotherapy in stage III non-small-cell lung cancer. *J Clin Oncol*. 2022;40(12):1301-1311. [CrossRef]
8. Antonia SJ, Villegas A, Daniel D, et al. Durvalumab after chemoradiotherapy in stage III non-small-cell lung cancer. *N Engl J Med*. 2017;377(20):1919-1929. [CrossRef]
9. Saad MB, Hong L, Aminu M, et al. Predicting benefit from immune checkpoint inhibitors in patients with non-small-cell lung cancer by CT-based ensemble deep learning: a retrospective study. *Lancet Digit Health*. 2023;5(7):e404-e420. [CrossRef]
10. Thomas A, Pattanayak P, Szabo E, Pinsky P. Characteristics and outcomes of small cell lung cancer detected by CT screening. *Chest*. 2018;154(6):1284-1290. [CrossRef]
11. Ligero M, Garcia-Ruiz A, Viaplana C, et al. A CT-based radiomics signature is associated with response to immune checkpoint inhibitors in advanced solid tumors. *Radiology*. 2021;299(1):109-119. [CrossRef]
12. Chen W, Wang R, Ma Z, et al. A delta-radiomics model for preoperative prediction of invasive lung adenocarcinomas manifesting as radiological part-solid nodules. *Front Oncol*. 2022;12:927974. [CrossRef]
13. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. *Gastrointest Endosc*. 2020;92(4):807-812. [CrossRef]
14. Seymour L, Bogaerts J, Perrone A, et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *Lancet Oncol*. 2017;18(3):e143-e152. [CrossRef]
15. 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]
16. A. Plonska PP. MLJAR. 2021. [CrossRef]
17. Taieb J, Svrcek M, Cohen R, Basile D, Tougeron D, Phelip JM. Deficient mismatch repair/microsatellite unstable colorectal cancer: diagnosis, prognosis and treatment. *Eur J Cancer*. 2022;175:136-157. [CrossRef]
18. Li N, Wan Z, Lu D, Chen R, Ye X. Long-term benefit of immunotherapy in a patient with squamous lung cancer exhibiting mismatch repair deficient/high microsatellite instability/high tumor mutational burden: a case report and literature review. *Front Immunol*. 2023;13:1088683. [CrossRef]
19. Mu W, Jiang L, Zhang J, et al. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. *Nat Commun*. 2020;11(1):5228. [CrossRef]
20. Le NQK, Kha QH, Nguyen VH, Chen YC, Cheng SJ, Chen CY. Machine learning-based radiomics signatures for EGFR and KRAS

mutations prediction in non-small-cell lung cancer. *Int J Mol Sci*. 2021;22(17):9254. [CrossRef]

21. Liu G, Xu Z, Ge Y, et al. 3D radiomics predicts EGFR mutation, exon-19 deletion and exon-21 L858R mutation in lung adenocarcinoma. *Transl Lung Cancer Res*. 2020;9(4):1212-1224. [CrossRef]
22. He L, Huang Y, Yan L, Zheng J, Liang C, Liu Z. Radiomics-based predictive risk score: a scoring system for preoperatively predicting risk of lymph node metastasis in patients with resectable non-small cell lung cancer. *Chin J Cancer Res*. 2019;31(4):641-652. [CrossRef]
23. Shu Y, Xu W, Su R, et al. Clinical applications of radiomics in non-small cell lung cancer patients with immune checkpoint inhibitor-related pneumonitis. *Front Immunol*. 2023;14:1251645. [CrossRef]
24. Yoon J, Suh YJ, Han K, et al. Utility of CT radiomics for prediction of PD-L1 expression in advanced lung adenocarcinomas. *Thorac Cancer*. 2020;11(4):993-1004. [CrossRef]
25. Trebeschi S, Drago SG, Birkbak NJ, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. *Ann Oncol*. 2019;30(6):998-1004. [CrossRef]
26. Faes L, Wagner SK, Fu DJ, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *Lancet Digit Health*. 2019;1(5):e232-e242. [CrossRef]
27. Fu XY, Mao XL, Wu HW, et al. Development and validation of LightGBM algorithm for optimizing of Helicobacter pylori antibody during the minimum living guarantee crowd based gastric cancer screening program in Taizhou, China. *Prev Med*. 2023;174:107605. [CrossRef]
28. Liu X, Zhu B, Dai XW, et al. GBDT_KgluSite: an improved computational prediction model for lysine glutarylation sites based on feature fusion and GBDT classifier. *BMC Genomics*. 2023;24(1):765. [CrossRef]
29. Wiesweg M, Mairinger F, Reis H, et al. Machine learning reveals a PD-L1-independent prediction of response to immunotherapy of non-small cell lung cancer by gene expression context. *Eur J Cancer*. 2020;140:76-85. [CrossRef]
30. Wei F, Azuma K, Nakahara Y, et al. Machine learning for prediction of immunotherapeutic outcome in non-small-cell lung cancer based on circulating cytokine signatures. *J Immunother Cancer*. 2023;11(7):e006788. [CrossRef]
31. Jiang J, Hu YC, Liu CJ, et al. Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images. *IEEE Trans Med Imaging*. 2019;38(1):134-144. [CrossRef]

Supplementary Table 1. Scanning parameters of two scanners

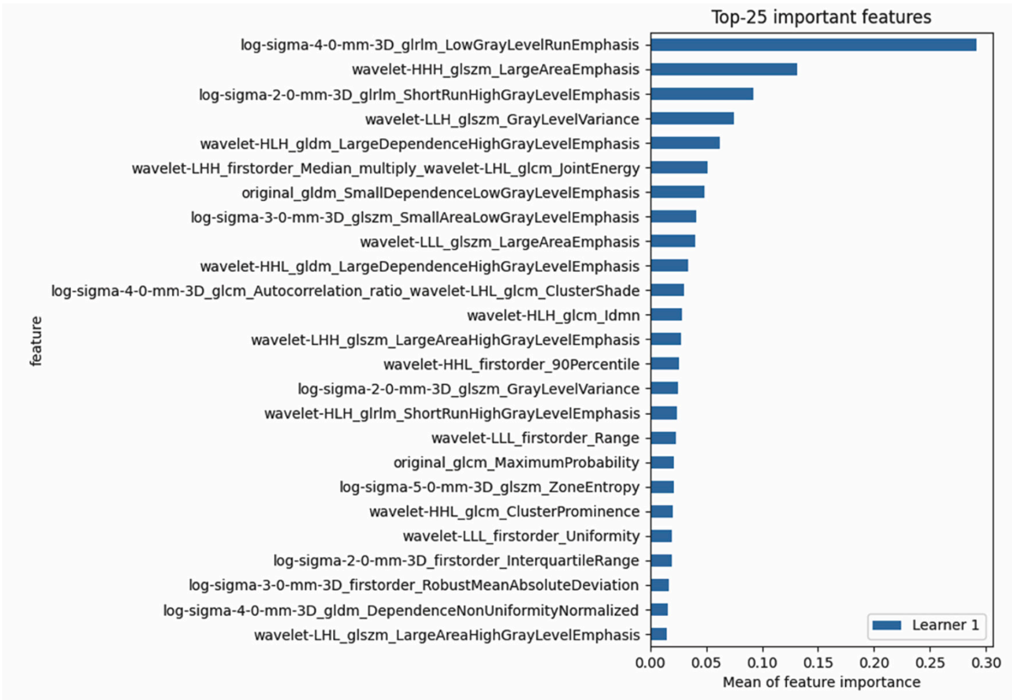
Parameters	GE Discovery CT750 HD	Somatom definition flash
Tube voltage (kVp)	120	120
Tube current (mAs)	200	110
Pitch	0.984:1	1.0
Collimation (mm)	0.625*64	0.6*64
Rotation time (s/rot)	0.5	0.33
SFOV (cm)	50	50
Slice thickness of reconstruction (mm)	1.25	1
Slice interval of reconstruction (mm)	1.25	1
Reconstruction algorithm	STND	Medium sharp

kVp, kilovoltage peak; mAs, milliampere-seconds; SFOV, scan field of view; STND, standard reconstruction algorithm.

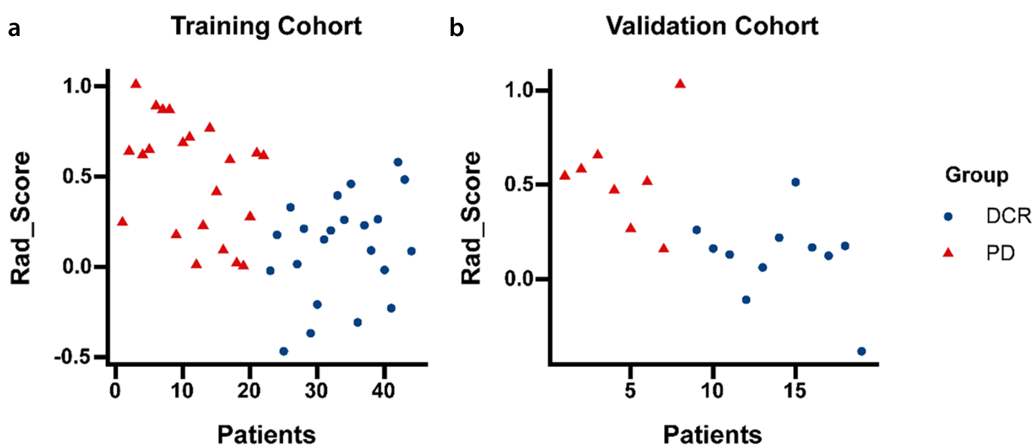
Supplementary Table 2. Evaluation of tumor response to immunotherapy

Tumor response	All patients (n = 63)
CR	0
PR	25
SD	8
PD	30 (47.6%)
DCR (CR + PR + SD)	33 (52.4%)

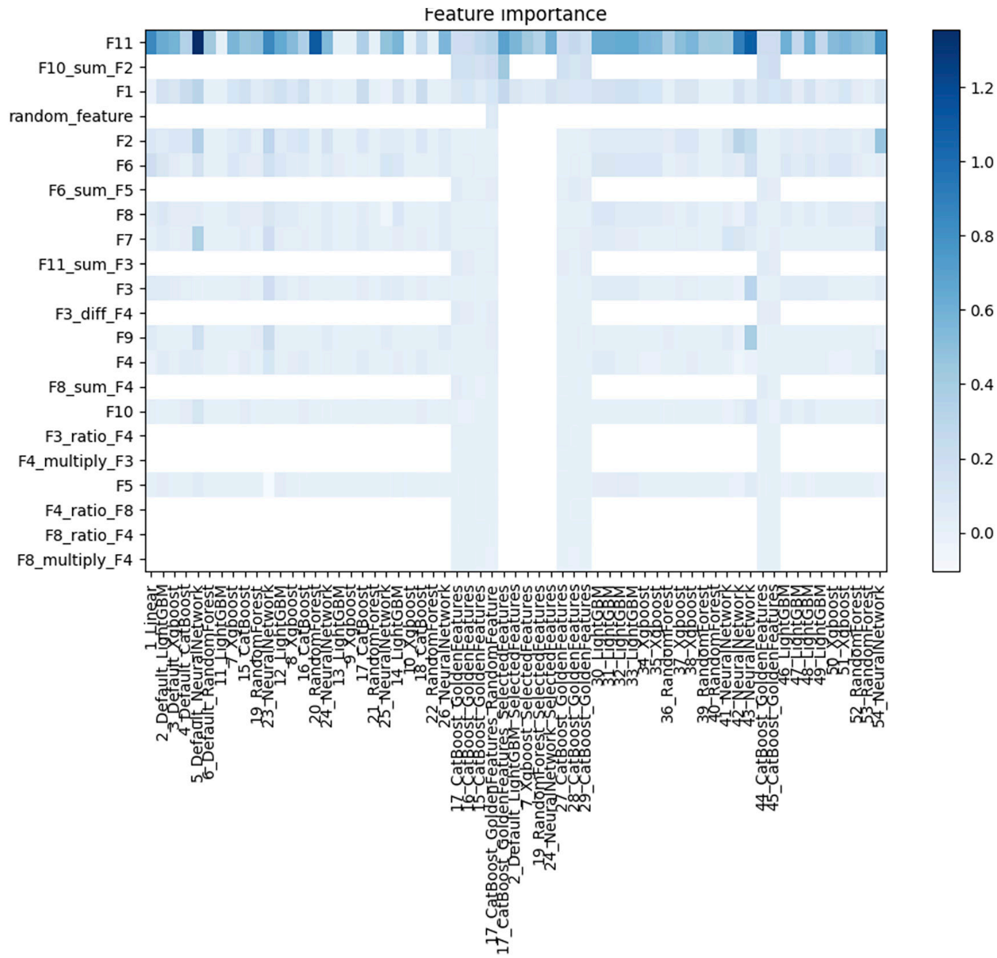
CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease; DCR, disease control rate.



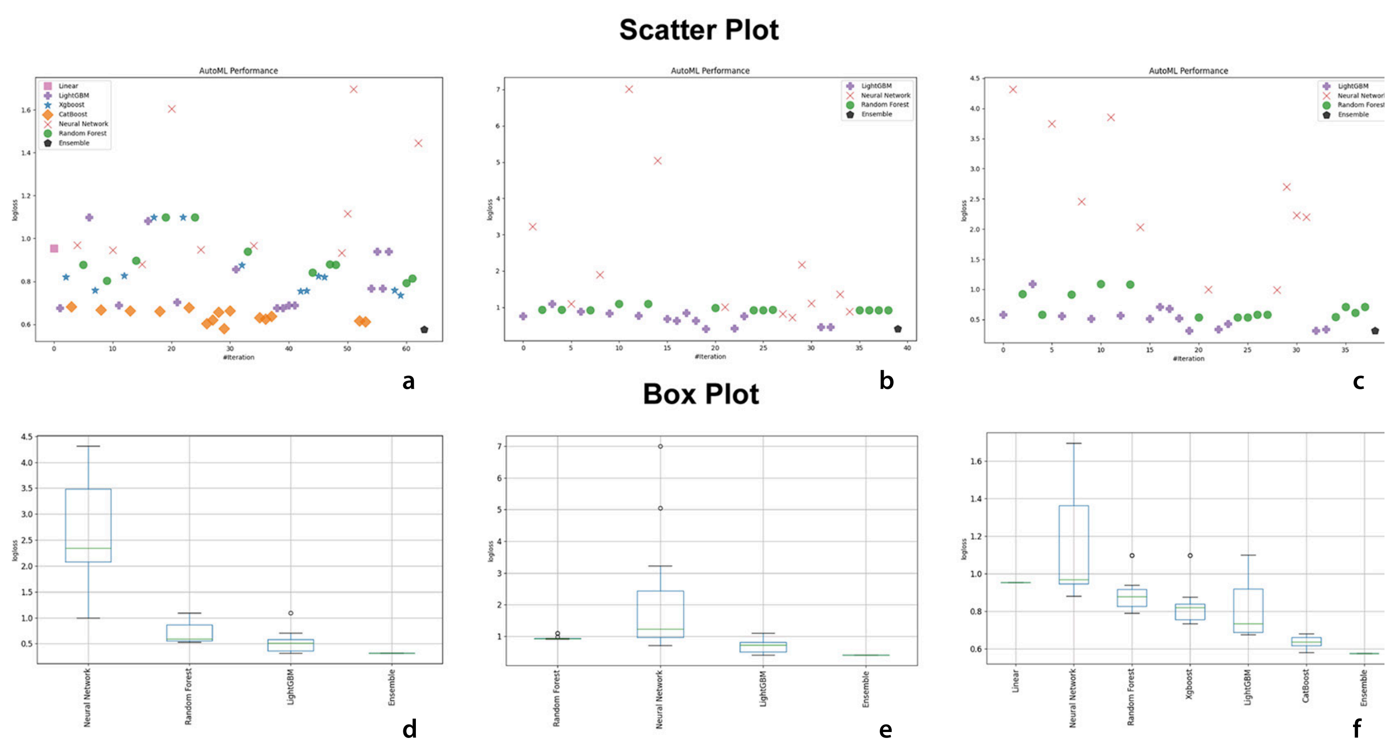
Supplementary Figure 1. Top 25 important radiomics features selected by LightGBM algorithm. LightGBM, light gradient-boosting machine.



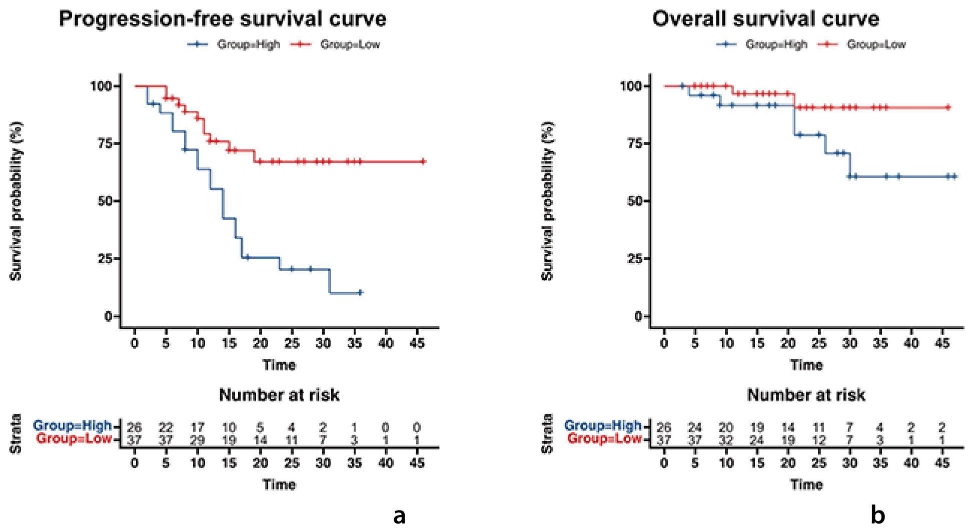
Supplementary Figure 2. The rad scores of patients in DCR and PD subgroups. **(a)** The training cohort; **(b)** the validation cohort. DCR, disease control rate; PD, progressive disease.



Supplementary Figure 3. Predictive clinical features generated by ensemble algorithm.



Supplementary Figure 4. The performance of the detection models in the training cohort. **(a, d)** clinical model; **(b, e)** radiomics model; **(c, f)** fusion model.



Supplementary Figure 5. Survival analyses in different groups of disease progression risk classified by the radiomics model. **(a)** Progression-free survival in different groups of Rad scores ($P < 0.01$); **(b)** Overall survival in different groups of Rad scores ($P = 0.056$).

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.232639

INTERVENTIONAL RADIOLOGY

TECHNICAL NOTE

Transperineal microwave thermoablation for benign prostatic hyperplasia-related lower urinary tract symptoms in an elderly patient

Yaşar Türk[1]
İsmail Devecioğlu[2,3]
Nusret Can Çilesiz[4]
Barış Nuhoğlu[4]

[1]Kolan International Hospital, Clinic of Radiology, İstanbul, Turkey

[2]Tekirdağ Namık Kemal University, Çorlu Faculty of Engineering, Department of Biomedical Engineering, Tekirdağ, Turkey

[3]University of New South Wales Sydney, School of Biomedical Sciences, Sydney, Australia

[4]Biruni University Faculty of Medicine, Department of Urology, İstanbul, Turkey

ABSTRACT

Transperineal prostate microwave thermoablation (TPMT) has been established as a safe means of treating benign prostatic hyperplasia (BPH); however, its effectiveness in addressing BPH-related lower urinary tract symptoms (LUTS) remains unexplored. This case study aims to evaluate the efficacy of TPMT in LUTS attributed to BPH. An 84-year-old man with LUTS due to BPH-induced bladder outlet obstruction, unresponsive to previous medical treatments, and failed prostate artery embolization, underwent TPMT. Three coaxial needles were positioned at the midline, right, and left sides of the hypertrophic transitional zone of the prostate. Microwave energy, with parameters determined using liver data and targeted ablation area, was applied at 2,450 MHz in continuous mode. The tissue temperature was monitored using bilateral thermocouple sensors. The patient exhibited no changes in defecation rhythm, abdominal discomfort, or anorectal pain. Temporary postoperative hematuria was promptly resolved through saline irrigation within 6 hours, and hematological evaluations showed normal results. Significant clinical improvements were observed (e.g., prostate volume, prostate-specific antigen levels) accompanied by an increase in peak flow rate. Thus, TPMT appears to be a promising intervention for bladder outlet stenosis and LUTS induced by BPH.

KEYWORDS

Benign prostatic hyperplasia, microwave ablation, prostate, thermoablation, urinary tract symptoms

Benign prostatic hyperplasia (BPH) has a substantial impact on the social and clinical aspects of geriatric men, and its significance cannot be overstated. It is a prevalent disease, affecting 50% of men in their sixth decade and 90% in their ninth decade, with an annual development of symptoms in 1.5% of men. Those with a prostate size >50 cm^2 face a five-fold increased risk of experiencing clinically mild-to-severe lower urinary tract symptoms (LUTS) and a three-fold elevated risk of significant bladder outlet blockage (peak flow rate 10 mL/sec).[1] These findings indicate a correlation between prostate growth, LUTS, and blockage, particularly in men with larger prostates.

For the past 45 years, transurethral prostate removal has been the gold standard for BPH surgery.[2] However, concerns regarding sexual dysfunction, hospitalization, and cost have prompted the exploration of alternative minimally invasive therapies.[1,2]

Among the alternative methods, microwave ablation has been valuable for treating hyperplasia. When tissue is exposed to microwave radiation (900–2,450 MHz), water molecules begin to oscillate, and the temperature increases due to friction. If the increase in temperature is sufficient, proteins and enzymes start to degrade, resulting in coagulative necrosis.[3] As per the guidelines of the American Urological Association, transurethral microwave thermotherapy (TUMT) is recognized as a potential treatment for LUTS associated with BPH[4] and is widely utilized.[5] Nevertheless, TUMT is contraindicated for patients with urethral strictures, as well as those with penile or urinary sphincter devices or a history of pelvic radiation. Conversely, transperineal prostate microwave thermoablation (TPMT) presents a viable alternative for these individuals.

Corresponding author: Yaşar Türk

E-mail: dryasarturk@gmail.com

Received 20 December 2023; revision requested 15 January 2024; accepted 10 February 2024.

Epub: 29.02.2024

Publication date: 03.03.2025

DOI: 10.4274/dir.2024.232639

You may cite this article as: Türk Y, Devecioğlu İ, Çilesiz NC, Nuhoğlu B. Transperineal microwave thermoablation for benign prostatic hyperplasia-related lower urinary tract symptoms in an elderly patient. *Diagn Interv Radiol.* 2025;31(2):141-144.

TPMT offers the advantage of a shorter duration-typically approximately 5 minutes-compared with the 30 minutes or so required for TUMT. Furthermore, TPMT may lead to a reduction in the risk of sexual dysfunction, persistent irritative voiding symptoms, incontinence, urinary tract infections, and repeated acute urinary retention episodes. Despite these potential benefits, TPMT has been explored in only a limited case series, focusing on the safety and comfort assessment of the device and the impact of microwaves on prostatic tissue *in vivo*, as reported by Bartoletti et al.[1] It is noteworthy that, to the best of our knowledge, there is currently no evidence supporting or refuting the efficacy of TPMT in addressing LUTS. In this case study, TPMT was performed on an 84-year-old patient whose LUTS resulting from bladder outlet obstruction due to BPH were resistant to medical treatment and did not show any improvement after prostate embolization (PE). Surgery was also deemed unsuitable due to the patient's underlying morbidities. The results showed that TPMT was effective in addressing LUTS attributed to BPH.

The patient

An 84-year-old man was presented with LUTS resulting from bladder outlet obstruction attributed to BPH (Figure 1a, Figure 2a). . Prior treatment with α-blockers had proven ineffective. The patient's advanced age and extensive arteriosclerosis of the anterior division of the internal iliac artery resulted in the failure of PE conducted at another medical facility. He experienced severe symptoms, as indicated by an international prostate symptom score of 28. Additionally, the patient exhibited a diminished maximum urinary flow rate of 5.1 mL/sec and incomplete bladder emptying, evidenced by a post-void residual urine volume of 350 mL. Prostate volume (PV) assessment revealed significant enlargement, at 218 mL. Elevated total and free prostate-specific antigen (PSA) serum levels were noted (total PSA: 14.9 ng/mL, free PSA: 7.78 ng/mL), with a free-to-total PSA ratio of >0.2. The PSA density (total PSA/PV) was calculated as 0.068, falling below the 0.15 limit.

Given the patient's advanced age, comorbidities (diabetes mellitus, hypertension, coronary artery disease, atrial fibrillation, cerebrovascular disease), and a poor 10-year prognosis, a decision was made to forego a prostate cancer study.

Technique

The patient received a 7-day course of cefixime (400 mg/day), Ibuprofen (2 × 600 mg/day), and gastroprotective therapy. Notably, the treatment initiation occurred 12 hours before the scheduled procedure.

The patient was positioned in the dorsal lithotomy posture. For anesthesia, lidocaine (10 mL) was administered into the prostate/seminal vesicle angle and the bilateral prostate apex. The coaxial needles and microwave antenna were inserted under guidance from ultrasound (Aplio 500, with a 3.5 MHz Convex ultrasound probe; Toshiba, Japan) and computerized tomography (SOMATOM Scope; Siemens AG, Germany). Three 15-G/13.8 cm coaxial needles (TruGuide; Bard, GA, USA) were strategically placed behind the urethra at the midline, right, and left sides of the hypertrophic transitional zone of the prostate (Figure 1b).

The exposure energy and duration were calculated using liver data, measuring the targeted ablation area around the microwave antenna. The antenna's tip was positioned at a distance >1 cm from the capsule and >0.5 cm from the rectal wall and the urethra, ensuring the preservation of the rectal wall and the urethra during thermoablation. The microwave ablation device, equipped with a 16-G/20-cm microwave ablation antenna (Canyon; Nanjing, China), operated at 2,450 MHz in continuous mode. The midline of the prostate received an exposure power of 20 W for 2 minutes, and the right and left sides had exposure powers of 40 W for 3 minutes and 2 minutes, respectively.

To monitor the temperature of the periprostatic tissue around the treatment site, two interstitial thermocouple sensors were placed bilaterally just outside the prostatic capsule, using two 19-G/20-cm microwave ablation temperature probes (Canyon; Nanjing, China). Post-treatment, the patient underwent immediate magnetic resonance imaging, followed by assessments at 1- and 3-month intervals. An 18-F urethral catheter, placed before TPMT, was removed 2 weeks later.

Main points

- Transurethral prostate resection is currently the recommended benign prostatic hyperplasia (BPH) procedure; however, mortality, sexual dysfunction, hemorrhage hospitalization, and high costs have led to less invasive options.
- BPH treatment using transperineal prostate microwave thermoablation (TPMT) is safe, but its efficacy in treating BPH-related lower urinary tract symptoms (LUTS) is unknown.
- This case study examined TPMT's effectiveness in BPH-related LUTS.
- Transperineal prostate microwave is an effective treatment method in BPH-related LUTS.

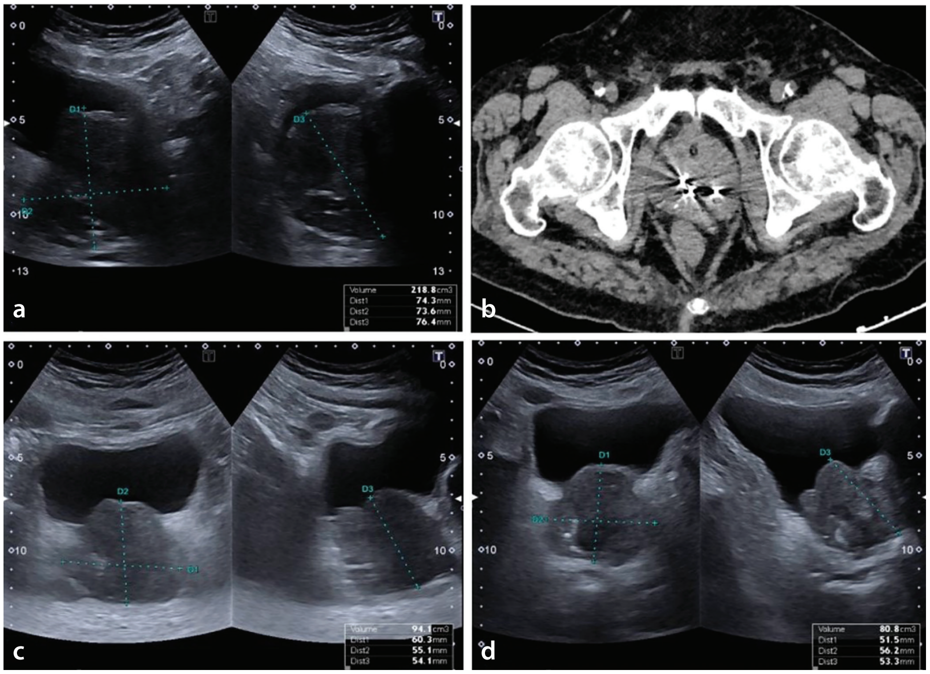


Figure 1. Ultrasound (US) image **(a)** before, computed tomography (CT) image **(b)** during, and US images **(c, d)** after the intervention. **(a)** The prostate gland was enlarged as measured on the US image before transperineal prostate microwave thermoablation (TPMT). **(b)** TPMT was performed under CT guidance. Three coaxial needles are observable in the image. The prostate gland was reduced in size after **(c)** 1 month and **(d)** 3 months post-procedure.

The safety and efficacy of the technique were assessed through a series of tests conducted before (within 30 days), as well as 1 and 3 months after TPMT. The assessments included blood tests, urine culture, complete urinalysis, uroflowmetry, and chest X-rays, as well as comprehensive abdominal, transperineal, and transrectal ultrasounds. These evaluations were repeated at 1 and 3 months post-TPMT to monitor any changes or developments. This study has obtained informed consent.

Results

The procedure was conducted as an outpatient procedure in a day hospital. The patient did not exhibit any abdominal discomfort, anorectal pain, or any changes in defecation rhythm following TPMT. The only transient issue observed was post-operative hematuria, which was promptly addressed through saline irrigation within 6 hours and did not require a blood transfusion. The initial PV of 218.8 mL decreased to 94.1 mL 1 month after TPMT and further to 80.8 mL 3 months after TMPT (Table 1). Subsequent hematological assessments conducted after TPMT consistently indicated normal results, while functional measures showed improvement (Table 1). The ejaculatory function was not recorded due to the limited sexual activity reported by the elderly patient. Meanwhile, a significant reduction in the prostate gland size was noted (Figure 1c, d, Figure 2b-d). The patient experienced approximately 3 weeks of dysuria following TPMT, which necessitated catheterization for 2 weeks post-procedure.

Discussion

We did not observe any of the clinical indications reported in the literature,[6] such as orchitis, prostatic abscess, urethral burn, urinary tract infections, or severe urinary retention. The episode of urinary incontinence lasting approximately 2 weeks may be attributed to the external urethral sphincter undergoing degeneration due to prolonged and severe BPH. We opted to retain the bladder catheter for 2 weeks to facilitate the reduction of the adenoma after consultation with the relevant urologists. Consequently, we recommend choosing a course of action for a bladder catheter on an individual basis. The occurrence of dysuria lasting approximately 3 weeks following TPMT is consistent with the patterns observed in previous minimally invasive thermal interventions for BPH.[6]

Bartoletti et al.[1] conducted open prostatectomy procedures in three groups at different intervals following TPMT. Their findings demonstrated that treating BPH with microwave thermotherapy is a safe, tolerable, and repeatable procedure, especially when employing a dedicated probe (AMICA-PROBE). The presented case further supports the versatility of using a common microwave probe. Notably, these microwave probes offer the advantage of customization for various anatomical sites. The studies targeting BPH using microwave thermoablation mostly occupy the transurethral method, which may not be suitable for some patients due to urethral

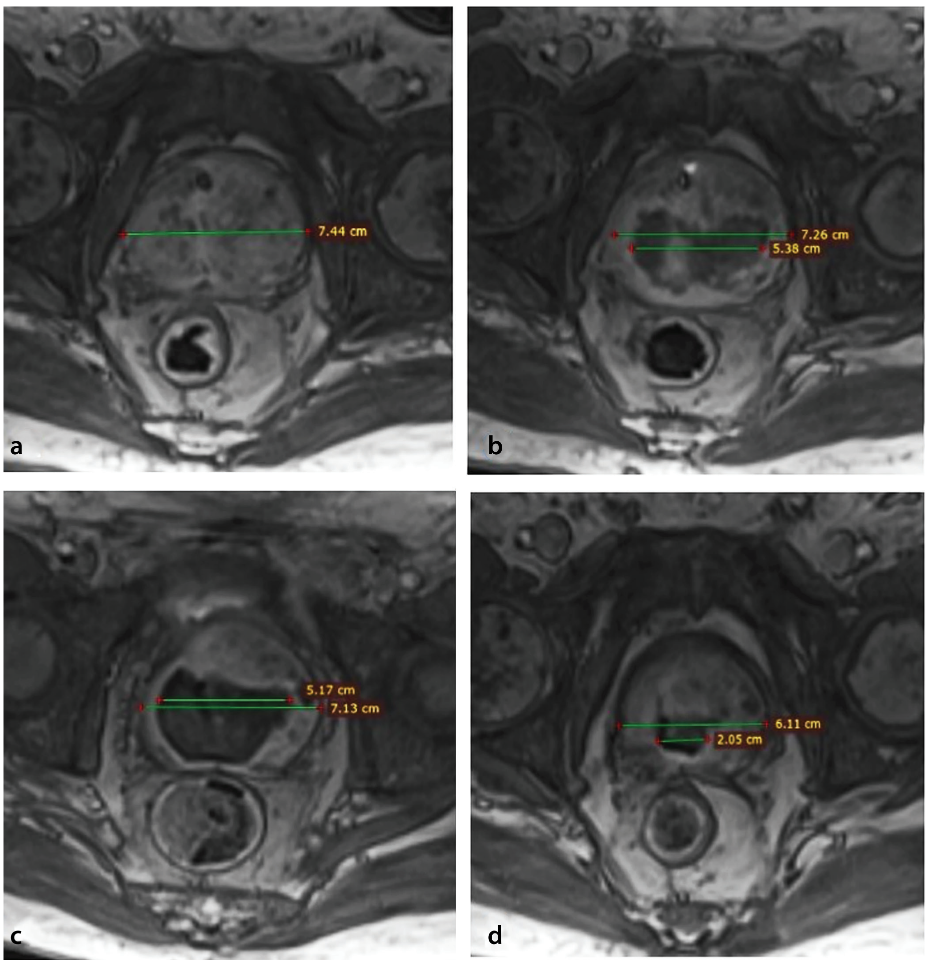


Figure 2. Contrast-enhanced T1-weighted magnetic resonance images before and after the intervention. The shrinkage of the prostate gland is observable immediately after transperineal prostate microwave thermoablation (TPMT) (transverse diameter of the prostate and the ablated zone: 7.26 and 5.38 cm) **(b)**, as well as in the follow-up controls 1 month (7.13 and 5.17 cm) **(c)** and 3 months after TPMT (6.11 and 2.05 cm) **(d)** in comparison with before TPMT condition (7.44 and 0.00 cm) **(a)**.

Table 1. Clinical measures related to the severity of lower urinary tract symptoms before and after the transperineal prostate microwave thermoablation procedure

	BII	IPSS	IPSS-QoL	Q_{max} (mL/sec)	PVR (mL)	Prostate size		T-PSA (ng/mL)	F-PSA (ng/mL)
						X × Y × Z (mm)	PV (mL)		
Before	11	28	6	5.1	350	74.3 × 73.6 × 76.4	218.8	14.9	7.78
After 1 month	5	15	3	7.3	150	60.3 × 55.1 × 54.1	94.1	1.32	0.56
After 3 months	2	7	1	11.5	90	51.5 × 56.2 × 53.3	80.8	1.19	0.44

BII, benign prostatic hyperplasia impact index; IPSS, international prostate symptom score; IPSS-QoL, IPSS quality of life index; Q_{max}, peak flow rate; PVR, postvoid residual volume; PV, prostate volume; PSA, prostate-specific antigen; T-PSA, total PSA; F-PSA, free PSA.

strictures or a history of radiation therapy in the pelvic region. To reliably compare the outcomes of these studies with the TPMT procedure using a common microwave probe, a further large-scale clinical study is needed. Although similar microwave ablation settings (e.g., frequency and operation mode) can be used for the treatment of both BPH and prostate cancer, further studies are still needed to optimize the duration and the power of microwave radiation in the treatment of LUTS due to BPH. Although the aged participant already reported limited sexual activity, we do not expect any significant effects on ejaculatory functions because the technique maintains the prostate capsule, urethra, and ejaculatory pathways similar to cryoablation. Nevertheless, further studies are required to confirm this.

In conclusion, the utilization of interstitial microwave ablation antennas in TPMT appears to be a promising treatment for LUTS caused by BPH. However, the need for randomized clinical studies is imperative to comprehensively assess the effectiveness and practicality of implementing TPMT in clinical settings.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Bartoletti R, Cai T, Tinacci G, et al. Transperineal microwave thermoablation in patients with obstructive benign prostatic hyperplasia: a phase I clinical study with a new mini-choked microwave applicator. *J Endourol*. 2008;22(7):1509-1518. [CrossRef]
2. Couteau N, Duquesne I, Frédéric P, et al. Ejaculations and benign prostatic hyperplasia: an impossible compromise? A comprehensive review. *J Clin Med*. 2021;10(24):5788. [CrossRef]
3. Carrafiello G, Laganà D, Mangini M, et al. Microwave tumors ablation: principles, clinical applications and review of preliminary experiences. *Int J Surg*. 2008;6(Suppl 1):65-69. [CrossRef]
4. Hoffman RM, MacDonald R, Monga M, Wilt TJ. Transurethral microwave thermotherapy vs transurethral resection for treating benign prostatic hyperplasia: a systematic review. *BJU Int*. 2004;94(7):1031-1036. [CrossRef]
5. Lerner LB, McVary KT, Barry MJ, et al. Management of lower urinary tract symptoms attributed to benign prostatic hyperplasia: AUA GUIDELINE PART II-Surgical Evaluation and Treatment. *J Urol*. Oct 2021;206(4):818-826. [CrossRef]
6. Sessa F, Polverino P, Siena G, et al. Transperineal laser ablation of the prostate (TPLA) for lower urinary tract symptoms due to benign prostatic obstruction. *J Clin Med*. 2023;12(3):793. [CrossRef]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242673

INTERVENTIONAL RADIOLOGY

ORIGINAL ARTICLE

Hepatic arterial infusion chemotherapy combined with toripalimab and surufatinib for the treatment of advanced intrahepatic cholangiocarcinoma

Songlin Song[1,2*]
Yiming Liu[1,2*]
Yanqiao Ren[1,2*]
Chuansheng Zheng[1,2]
Bin Liang[1,2]

[1]Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[2]Hubei Provinve Key Laboratory of Molecular Imaging, Wuhan, China

PURPOSE

The aim of the present study is to report the clinical results of patients with advanced intrahepatic cholangiocarcinoma (ICC) who received combination therapy of hepatic arterial infusion chemotherapy (HAIC), toripalimab and surufatinib.

METHODS

The study cohort consisted of 28 patients with advanced ICC who were treated with HAIC (mFOLFOX6 regimen, Q3W) in combination with intravenous toripalimab (240 mg, Q3W) and oral surufatinib (150 mg, once daily). The cohort had 14 male and 14 female patients. The baseline characteristics of the study cohort were obtained. The tumor response and drug-associated toxicity were assessed and reported.

RESULTS

During the follow-up period (median follow-up time: 11.3 months; range: 4–19 months), four patients died of tumor progression. The objective response rate and disease control rate were 58% and 79%, respectively. The mPFS was 9.5 months, and the overall survival rate was 83.3%. The most frequent adverse events were nausea and vomiting (100%) and abdominal pain (85.7%). Serious complications related to death were not observed.

CONCLUSION

The combination treatment schedule for advanced ICC demonstrated positive efficacy and safety profiles.

CLINICAL SIGNIFICANCE

This study provides promising clinical guidance for the treatment of advanced cholangiocarcinoma and is expected to modify the treatment strategy for this disease.

KEYWORDS

Intrahepatic cholangiocarcinoma, hepatic arterial infusion chemotherapy, toripalimab, surufatinib

***Joint first authors**

Corresponding author: Chuansheng Zheng, Bin Liang

E-mail: hqzcsxh@sina.com, bliang@hust.edu.cn

Received 18 January 2024; revision requested 04 March 2024; accepted 21 April 2024.

Epub: 03.06.2024

Publication date: 03.03.2025

DOI: 10.4274/dir.2024.242673

Carcinoma of the biliary tract can be classified according to the tumor location as either intrahepatic cholangiocarcinoma (ICC) or extrahepatic cholangiocarcinoma. ICC is the second most common primary liver cancer after hepatocellular carcinoma (HCC), and it accounts for 5%–10% of primary malignancies of the liver.[1] The first-line treatment for ICC is surgical resection. However, approximately 75% of ICCs are diagnosed at an advanced stage, and surgery is not possible for these patients.[2] As a result, the overall prognosis for ICC is very poor, with a median survival of less than 4 months for patients not treated through surgery.[3,4] The current preferred first-line chemotherapy for locally advanced ICC is gemcitabine plus cisplatin (GEMCIS). However, the reported median survival for patients with advanced ICC treated with GEMCIS is only 11.7 months.[5] Therefore, new therapeutic strategies for advanced ICC are needed.

You may cite this article as: Song S, Liu Y, Ren Y, Zheng C, Liang B. Hepatic arterial infusion chemotherapy combined with toripalimab and surufatinib for the treatment of advanced intrahepatic cholangiocarcinoma. *Diagn Interv Radiol.* 2025;31(2):145-151.

Hepatic arterial infusion chemotherapy (HAIC) is a well-established transcatheter therapy for hepatic malignancies. With the use of an intraarterially inserted catheter, HAIC may effectively deliver highly concentrated doses of chemotherapy to the tumor bed while sparing the surrounding liver parenchyma.[6] HAIC has been demonstrated to be safe and effective for the treatment of advanced liver-confirmed and unresectable ICC.[7] Given the promising efficacy of targeted therapy and immunotherapy in various malignant tumors, the combination of HAIC with tyrosine kinase inhibitors (TKIs) and programmed cell death protein 1 (PD-1) inhibitors for treating advanced ICC has recently been investigated.[8,9]

Toripalimab is a humanized anti-PD-1 immunoglobulin G4 (IgG4) monoclonal antibody. This drug has demonstrated promising efficacy and safety profiles for treating urologic cancer, melanoma, and gastric cancer. Surufatinib is a small molecule inhibitor of vascular endothelial growth factor (VEGF) receptors 1, 2, and 3, fibroblast growth factor receptor (FGFR) 1, and colony-stimulating factor 1 receptor. Similar to toripalimab, surufatinib has demonstrated promising clinical efficacy and positive tolerability and safety profiles in patients with advanced solid tumors, such as neuroendocrine neoplasms and thyroid tumors. Recently, the use of toripalimab and surufatinib for treating unresectable ICC has been reported.[10,11]

However, the use of HAIC plus toripalimab and surufatinib for the treatment of unresectable ICC has not yet been reported. Therefore, we conducted this study to explore the efficacy and safety of this triple combination treatment for the treatment of unresectable ICC.

Methods

Our study was a retrospective cohort study. The clinical outcomes of patients with advanced ICC who received HAIC + toripalimab + surufatinib maintenance combination therapies between July 2021 and Oct 2023 at Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, were analyzed (Figure 1). The inclusion criteria before treatment were (1) age >18 years, (2) histologically confirmed diagnosis of ICC through ultrasonic-guided biopsy, (3) previous systemic and/or locoregional therapy, (4) Eastern Cooperative Oncology Group (ECOG) performance status 0–2, (5) tumor size evaluable using the Response Evaluation Criteria in Solid Tumors (RECIST; version 1.1) guidelines,[12] (6) liver, renal, and hematological functions compatible with chemotherapy, and (7) life expectancy ≥3 months. Patients were excluded if they had (1) severe infection or heart, liver, or lung failure, (2) other malignant tumors, (3) uncontrolled ascites, or (4) incomplete medical information or were lost to follow-up.

All patients were informed of the purpose of this study, and written consent was obtained. All study protocols were approved by the Ethics Committee of Union Hospital, Tongji Medical College, Huazhong University of Science and Technology (UHCT-IEC-SOP-014-01-02, 2023/07/20) in accordance with the 1975 Declaration of Helsinki.

The baseline characteristics included the patients' demographics, presence of extrahepatic metastases, previous therapies, tumor stage, tumor dimension determined through enhanced computed tomography (CT) and/or magnetic resonance imaging (MRI), and tumor marker levels [carbohydrate antigen 19-9: (CA19-9)].

Treatment procedures and regimens

To perform HAIC, an intraarterial catheter was inserted through the femoral artery using the method described by Irie.[13] A 5F catheter was inserted through the right femoral artery using the Seldinger method. After localization of the ICC, a 5F heparin-coated polyurethane catheter (Braun Medical, Chasseneuil du Poitou, France) was placed at the depth of the gastroduodenal artery (3–5 cm from the origin) to avoid dislocation of the catheter tip, and a side hole (2–3 mm in a longitudinal direction) was made at the level of the common hepatic artery with scissors. The other end of the catheter was connected to the injection port, which was implanted in a subcutaneous pocket created in the right thigh. The gastroduodenal artery and right gastric artery were occluded with steel coils to prevent gastroduodenal injury by the chemotherapeutic agents (Figure 2).

When the blood supply to the HCC stemmed partly from the extrahepatic artery (e.g., a replaced/accessory right hepatic artery from the superior mesenteric artery, replaced/accessory left hepatic artery from the left gastric artery, or other extrahepatic collateral vessels), the artery was first embolized with coils to redistribute the flow of the whole hepatic artery perfusion from multiple arteries to a single artery. This step ensured effective hepatic intraarterial infusion through a single infusion catheter. In this study, arterial port implantation was not suitable for 20 patients as a result of vascular anatomical variation (e.g., the right gastric artery could not be embolized). Therefore, an alternative approach was used to temporarily insert the catheter, and the catheter was removed after chemotherapy.

In the present study, we used an mFOLFOX6 regimen for HAIC (oxaliplatin: 85 mg/m^2 for 2 h on day 1; calcium folinate: 200 mg/m^2 for 2 h on day 1; 5-Fu: 400 mg/m^2 for bolus

Main points

- The current first-line intravenous chemotherapy regimen for advanced intrahepatic cholangiocarcinoma (ICC) is deemed unsatisfactory due to its short survival period.
- Tyrosine kinase inhibitors, whether used alone or in combination with immunotherapy, have shown limited efficacy in treating advanced ICC.
- Combining hepatic arterial infusion chemotherapy with toripalimab and surufatinib has demonstrated a significant improvement in the survival period of patients with advanced ICC.

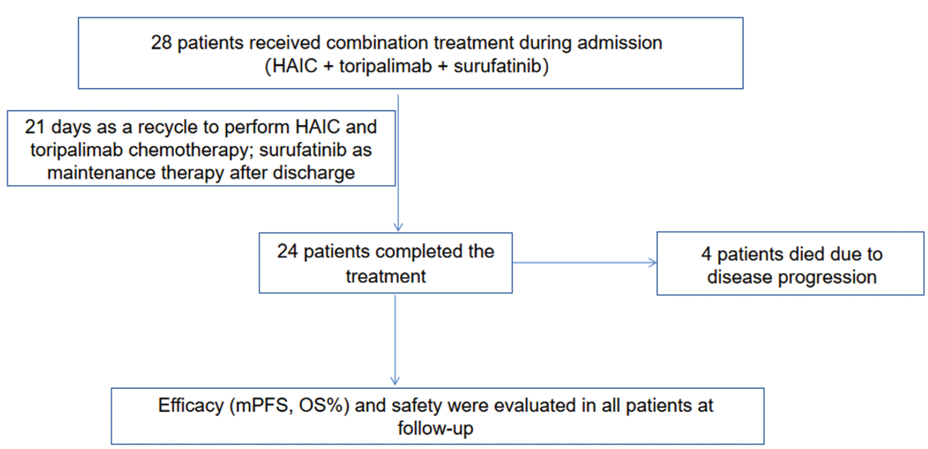


Figure 1. Study treatment flowchart. HAIC, hepatic arterial infusion chemotherapy; mPFS, median progression-free survival; OS, overall survival.

on day 1, followed by 2400 mg/m^2 for 46 h; Q3W). Following HAIC therapy, the patients also received intravenous toripalimab (240 mg on day 3, Q3W) and oral surufatinib (150 mg, once daily). The treatment was performed until unacceptable toxicity or disease progression occurred (Figure 3). Toxicity was recorded and evaluated in accordance with the National Cancer Institute Common Toxicity Criteria for Adverse Events (NCI-CTCAE; version 5.0) guidelines.

Tumor response

The disease responses after therapy were classified as complete response (CR), partial response (PR), stable disease (SD), or progressive disease (PD). Every two cycles, the response to therapy was assessed using RECIST through CT or MRI. A CR was defined as the complete disappearance of all target lesions, a PR was defined as a ≥30% decrease in the maximum diameter of the target lesion compared with the baseline maximum diameter, PD was defined as a ≥20% increase in the maximum diameter of the target lesion, and SD was defined as disease meeting neither the PR nor PD criteria. For responses other than PD, the combination treatment was repeated. Clinical visits, laboratory testing for blood counts, liver functionality, and tumor marker levels (CA19-9) were performed monthly.

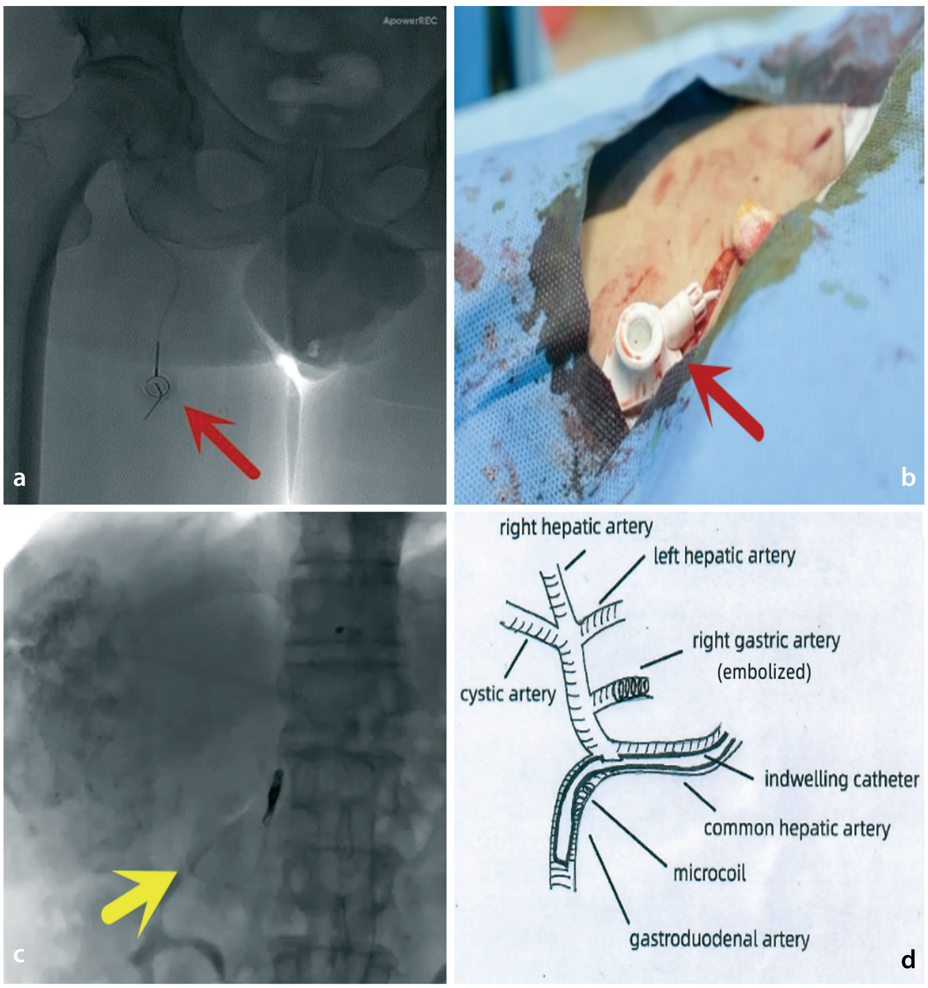


Figure 2. Intraoperative diagram of hepatic arterial infusion chemotherapy. The arterial port (Bard Access Systems, USA) is usually implanted subcutaneously 2 cm below the right groin of the patient, and the arterial catheter (**a** and **b**, red arrow) is then connected. The distal end of the arterial catheter is located in the gastroduodenal artery (**c**, yellow arrow). The distal hole of the arterial catheter is opened to allow the microcatheter into the gastroduodenum through the lateral hole, and the spring ring is implanted through the microcatheter to fix the arterial catheter into the gastroduodenal artery. The ductus arteriosus is compressed, and the chemotherapy drug flows to the internal hepatic artery through the lateral pore (if the right gastric artery is found, embolization is performed at the same time. D, schematic diagram).

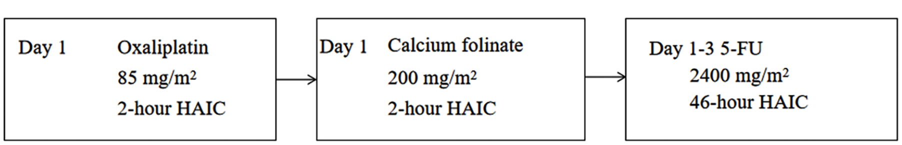


Figure 3. Schedule of chemotherapy administration. HAIC, hepatic arterial infusion chemotherapy.

Tolerability

Toxicity was evaluated according to the NCI-CTCAE guidelines. In the case of toxicity of grade 3 or above, treatment was temporarily suspended. Toxicity was evaluated every 2 or 3 days for each patient. After confirming that the toxicity had resolved to grade 1 or below, treatment was resumed with the same regimen. If toxicity of grade 3 or above was observed again after retreatment, treatment was temporarily suspended, and the patient resumed treatment at a reduced dose after the resolution was verified.

Statistical analysis

The statistical analyses were performed using SPSS 26.0 (IBM, Armonk, NY, USA). Continuous variables are expressed as the means and standard deviations or medians with ranges where appropriate. Qualitative variables are described as percentages or frequencies. The median progression-free survival (mPFS) and overall survival rate (OS %), objective response rate (ORR), disease control rate (DCR), and CR, PR, SD, and PD rates were also calculated.

Results

Baseline characteristics

The included patients' baseline characteristics are summarized in Table 1. Among the 28 patients, 10 had abdominal lymph node metastasis and 2 had portal vein tumor thrombus. All the patients were confirmed to have ICC through pathological examination. Of the 16 patients with abnormal CA 19-9 values, 12 (75%) exhibited a reduction from baseline. Additionally, 18 patients (64%) with an initial ECOG performance status >0 improved during therapy. Pain improved in 20 of the 28 initially symptomatic patients (71%).

Clinical outcomes of the combination therapy

The response and survival outcomes are summarized in Table 2. The mean number of treatment cycles for all participants was 6.4 cycles (range: 4–10 cycles). With regard to an

Table 1. The baseline characteristics of included patients		
Characteristics		All patients (n = 28)
Age, year, median (range)		51.5 (28–62)
Gender (f/m)		14/14
PLT, × 10^9/L, median (range)		203 (89–449)
ECOG performance status		
0		0
1		24
2		4
WBC, × 10^9/L, median (range)		4.75 (2.33–13.1)
HB, g/L, median (range)		118 (74–155)
ALT, U/L, median (range)		55 (11–110)
AST, U/L, median (range)		44 (23–90)
TBIL, μmol/L, median (range)		17.5 (5.5–45.8)
ALB, g/L, median (range)		37.2 (26.5–41.9)
Child–Pugh class		
A		14
B		14
Portal vein tumor thrombus		2
Abdominal lymph node metastasis		10
Pretreatment		
TACE		5
PTCD		4
Systemic chemotherapy		19
CA 19-9 (U/mL)		
Baseline	≥1200	16
3 months later	<1200	12
Number of treatment cycles		6.4 (4–10)
Follow-up time (months)		11.3 (4–19)

PLT, platelet; ECOG: Eastern Cooperative Oncology Group; WBC, white blood cell; AST, aspartate aminotransferase; ALT, alanine aminotransferase; TBIL, total bilirubin; ALB, albumin; TACE, transcatheter arterial chemoembolization; PTCD: percutaneous transhepaticcholangial drainage.

Table 2. Tumor response	
Responses	All patients (n = 28)
CR, n (%)	2 (7%)
PR, n (%)	14 (50%)
SD, n (%)	6 (21%)
PD, n (%)	6 (21%)
ORR, n (%)	16 (57%)
DCR, n (%)	22 (79%)

CR, complete response; PR, partial response; SD, stable disease; PD, progression disease; ORR, objective response rate; DCR, disease control rate.

early response, only two CRs were observed (Figure 4), and 14 and 6 patients achieved a PR and SD, respectively. However, six patients had PD. The ORR (CR + PR/all patients) and DCR (CR + PR + SD/all patients) were 57% and 79%, respectively.

The mPFS of the patients was 9.5 months (median follow-up time: 11.3 months; range: 4–19 months). The cumulative survival rate at 1 year was 83.3% (Figure 5). Four patients died of tumor progression.

Adverse effects

All patients were evaluated for adverse effects and complications related to the implantable port system. Port systems were successfully implanted in eight patients, with other patients receiving alternative methods. No complications were considered to be catheter-related toxicity. The adverse effects at initial treatment are summarized in Table 3. No treatment-related deaths occurred.

All 28 patients (100%) developed nausea and vomiting, but no severe cases of nausea or vomiting were observed. Twenty-four of the 28 (85.7%) patients developed abdominal pain, and two patients experienced severe abdominal pain after oxaliplatin injection. In these two patients, no significant improvement in symptoms was observed after lidocaine injection. Subsequently, when an intravenous infusion of butorphanol tartrate was used, the pain was significantly relieved. Mild diarrhea was observed in four patients (14.3%), and mild neurotoxicity was observed in two patients (7.1%).

Regarding blood toxicity, eight patients developed leukopenia, including two with grade 3 leukopenia. Moreover, 12 patients developed thrombocytopenia, including 2 with grade 3 thrombocytopenia. Severe blood toxicity complications in both of these patients were subsequently corrected through splenic artery embolization.

All patients were treated for at least four cycles. During the follow-up period, six patients discontinued treatment because of disease progression.

Discussion

Recent phase III clinical trials have demonstrated that treatment with GEMCIS in combination with durvalumab or pembrolizumab significantly improved OS compared with conventional chemotherapy alone with similar safety profiles in patients with unresectable or metastatic biliary tract cancers.[14,15] These studies have prompted further exploration of the triple combination treatment of HAIC, TKIs, and immune checkpoint inhibitors. In the present study, we demonstrated that the combination of HAIC, surufatinib, and toripalimab achieved promising patient survival (mPFS, 9.5 months; 1-year survival rate, 83.3%) as well as sufficient tumor response (ORR, 57%; DCR, 79%). In addition, the combination treatment-related adverse events were manageable.

Systemic chemotherapy can increase OS and improve the quality of life of patients with advanced ICC.[14] Several combination chemotherapy regimens have been investi-

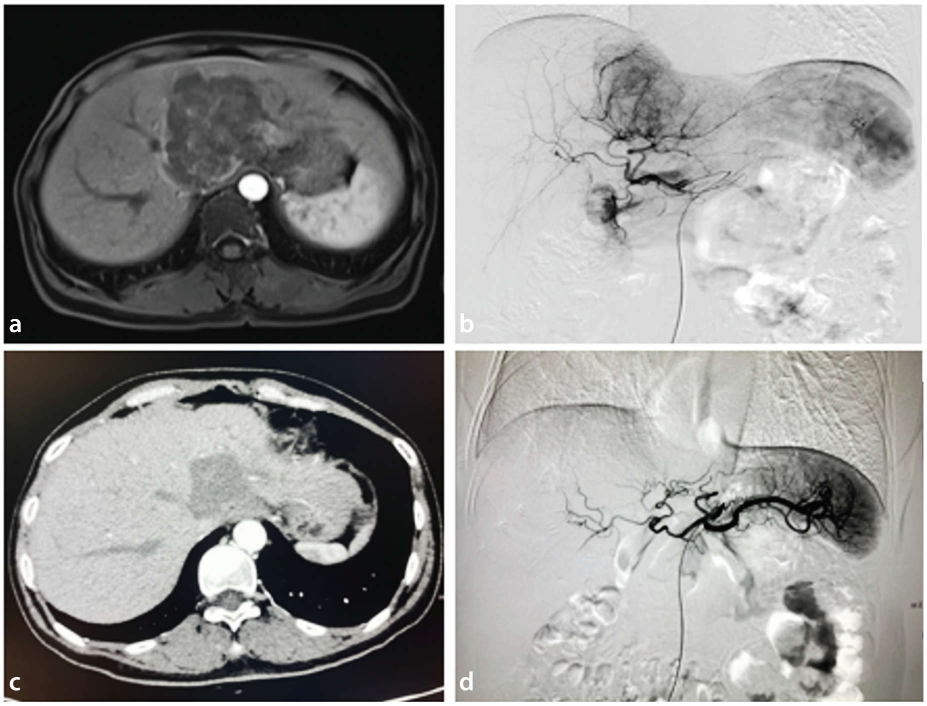


Figure 4. A 51-year-old female patient with ICC (confirmed through puncture biopsy). Magnetic resonance imaging revealed a large mixed density shadow in the liver (62 × 55 × 53 mm) (**a**). Digital subtraction angiography revealed an increased tortuous hepatic artery and obvious tumor staining (**b**). After three cycles of combination treatment, the computed tomography arterial phase demonstrated that the enhancement degree of the lesions was significantly reduced (**c**). Digital subtraction angiography indicated that the tumor staining had disappeared (**d**). ICC, intrahepatic cholangiocarcinoma.

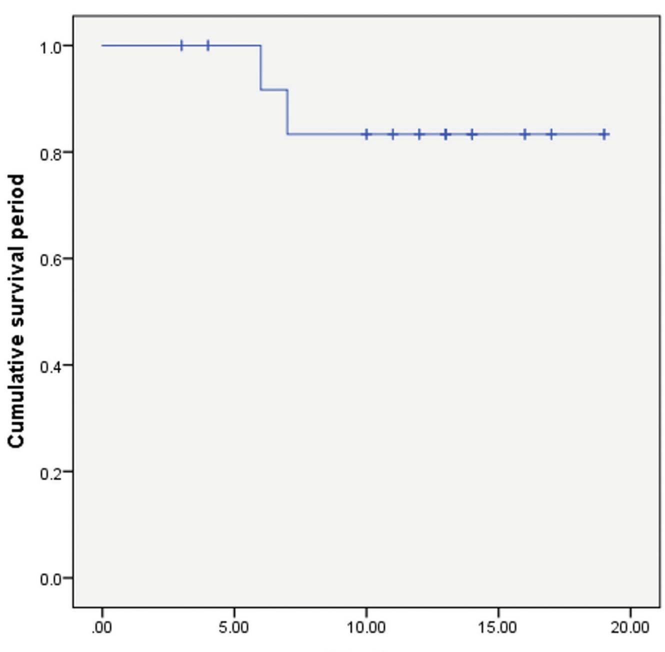


Figure 5. Kaplan–Meier curves illustrate the patient survival rates during the follow-up period.

gated, including gemcitabine/capecitabine, with an ORR of 25%, and gemcitabine/oxaliplatin, with an ORR of 50%.[14] A multicenter, open-label, phase 1 trial revealed that nivolumab (a PD-1 inhibitor) monotherapy had antitumor activity in Japanese patients with advanced cholangiocarcinoma, yielding an ORR of 3.3%, a median OS of 5.2 months, and an mPFS of 1.4 months. However, the combination therapy with nivolumab and chemotherapy achieved improved survival benefits in terms of a higher ORR (33.3%), longer median OS (15.4 months), and longer mPFS (4.2 months).[16] The multicenter, global, phase-3 TOPAZ-1 trial reported that GEMCIS chemotherapy plus durvalumab could significantly increase the median OS by 1.3 months (median OS: 12.8 vs. 11.5 months) when used as the first-line treatment for unresectable and metastatic cholangiocarcinoma compared with GEMCIS chemotherapy alone. In another study, the ORR was 26.7% in a GEMCIS chemotherapy plus durvalumab group, which surpassed that in the GEMCIS chemotherapy group.[17] Systemic chemotherapy has only limited benefits. The median OS after GEMCIS therapy is still <1 year.[18,19] Some studies have evaluated the use of cisplatin in combination with a bolus of 5-FU and epirubicin, with tumor ORRs ranging from 10% to 35% and a median OS of 11 months.[10,20] Another study by Valle et al.[3] reported a median OS of 11.7 months from the ABC-02 trial, and this was also reported in a study by Fu et al.[21] These rates were higher in our study than in previous studies. However, Shi et al.[10] reported on the efficacy of toripalimab combined with lenvatinib and GEMOX as first-line therapy for advanced ICC. The median OS and PFS were 22.5 and 10.2 months, respectively. Our data demonstrated a similar clinical application prospect (mPFS of 9.5 months, and the cumulative survival rate from the time of diagnosis was 83.3% at 1 year. The rationale for the use of HAIC can be summarized as follows. First, ICCs are usually confined to the liver, and patients mainly die of liver failure. Second, some drugs result in high hepatic extraction after the first pass. Moreover, the blood supplied to the upper biliary tree and gallbladder is derived from the hepatic artery.[17,22] The administration of oxaliplatin through the hepatic artery provides a high drug concentration in the perfused blood, and systemic complications are much lower.[23]

Few studies have focused on systemic ICC treatments, and most of these studies did not yield clear results. Therefore, it is difficult to draw a conclusion about which is preferable for systemic or locoregional therapies. Moreover, limited data related to maintenance therapy for ICC are available. For this reason, the present study adopted toripalimab and surufatinib maintenance therapy, and innovative data for the treatment method were reported.

Toripalimab is a humanized anti-PD-1 IgG4 monoclonal antibody approved for clinical trials by the US Food and Drug Administration (FDA) and China's National Medical Products Administration. This drug has demonstrated promising efficacy and safety profiles for use in treating urologic cancer, melanoma, and gastric cancer.[21,24-26]

Table 3. Adverse events

Events	All patients (n = 28)	
	Any grade	Grade 3 to 4
Nausea and vomiting	28	0
Abdominal pain	24	2
Diarrhea	4	0
Neurotoxicity	2	0
Leukopenia	8	2
Thrombopenia	12	2
Mucositis	0	0
Infection	0	0

Surufatinib is a multikinase inhibitor that targets VEGF receptors 1 to 3, FGFR 1, and colony-stimulating factor 1 receptors. A high expression level of VEGF was detected in 53.8% of ICCs and was considered to be involved in hematogenous metastasis. The FGFR signaling pathway is also abnormally activated in ICC and is associated with an unfavorable prognosis.[27,28] Finally, considering that surufatinib and chemotherapy regimens can significantly upregulate PD-L1 expression, using these therapies with anti-PD-1 treatment may significantly enhance their effects. Notably, combined therapy with an anti-PD-1 antibody and surufatinib has been reported to be useful for the treatment of several cancer types, and the FDA has approved the combination of surufatinib and toripalimab for treating advanced endometrial cancer and advanced renal cell carcinoma.

Maintenance therapy cannot be performed in unfit patients who are not clinically indicated for chemotherapy. For this reason, maintenance therapy is usually performed only in those who respond to HAIC, primarily to prolong the benefits of HAIC on survival. Maintenance therapy has demonstrated promising results in terms of tumor response, survival, and progression delay in many types of cancers. These results may suggest a possible advantage of maintenance therapy for ICC. Therefore, combining anti-PD-1 therapy with the combination of surufatinib and HAIC for the treatment of ICC seems reasonable. Our findings suggest that HAIC combined with toripalimab and surufatinib may be a new and promising treatment approach for advanced ICC.

The present study has certain limitations that must be considered. The main limitation of this study is that it is retrospective, and the number of participants was relatively limited. A technical limitation is that polymerase chain reaction and DNA sequencing have not yet been performed to detect antimicrobial resistance genes, whose characterization is also essential for surveillance, infection control, and therapeutic purposes. In additional studies, we will perform genome-wide sequencing to identify a pathogen by comparing its sequence to a database of known pathogens to determine its closest relatives.

In conclusion, future randomized controlled studies are needed to enhance the reliability of the findings because of the short follow-up duration. Finally, as previously mentioned, the sample size should be increased to obtain more conclusive results.

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

This work was supported by grant from National Nature Science Foundation of China (no. 81873919).

References

1. Li QJ, He MK, Chen HW, et al. Hepatic arterial infusion of oxaliplatin, fluorouracil, and leucovorin versus transarterial chemoembolization for large hepatocellular carcinoma: a randomized phase III trial. *J Clin Oncol*. 2022;40(2):150-160. [CrossRef]
2. Yao KJ, Jabbour S, Parekh N, Lin Y, Moss RA. Increasing mortality in the United States from cholangiocarcinoma: an analysis of the National Center for Health Statistics Database. *BMC Gastroenterol*. 2016;16(1):117. [CrossRef]
3. Valle J, Wasan H, Palmer DH, et al. Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer. *N Engl J Med*. 2010;362(14):1273-1281. [CrossRef]
4. Park J, Kim MH, Kim KP, et al. Natural history and prognostic factors of advanced cholangiocarcinoma without surgery, chemotherapy, or radiotherapy: a large-scale observational study. *Gut Liver*. 2009;3(4):298-305. [CrossRef]
5. Shaib Y, El-Serag HB. The epidemiology of cholangiocarcinoma. *Semin Liver Dis*. 2004;24(2):115-125. [CrossRef]
6. Brown DB, Cardella JF, Sacks D, et al. Quality improvement guidelines for transhepatic arterial chemoembolization, embolization, and chemotherapeutic infusion for hepatic malignancy. *J Vasc Interv Radiol*. 2006;17(2 Pt 1):225-232. [CrossRef]
7. Yang Z, Fu Y, Wu W, et al. Comparison of hepatic arterial infusion chemotherapy with mFOLFOX vs. first-line systemic chemotherapy in patients with unresectable intrahepatic cholangiocarcinoma. *Front Pharmacol*. 2023;14:1234342. [CrossRef]
8. Wang Y, Wei Z, Zhang Z, et al. Hepatic arterial infusion chemotherapy with or without lenvatinib for unresectable cholangiocarcinoma: a single-center retrospective study. *Hepat Oncol*. 2023;10(2):HEP49. [CrossRef]
9. Huang Y, Du Z, Kan A, et al. Clinical and biomarker analyses of hepatic arterial infusion chemotherapy plus lenvatinib and PD-1 inhibitor for patients with advanced intrahepatic cholangiocarcinoma. *Front Immunol*. 2024;15:1260191. [CrossRef]
10. Shi GM, Huang XY, Wu D, et al. Toripalimab combined with lenvatinib and GEMOX is a promising regimen as first-line treatment for advanced intrahepatic cholangiocarcinoma: a single-center, single-arm, phase 2 study. *Signal Transduct Target Ther*. 2023;8(1):106. [CrossRef]
11. Xu J, Bai Y, Sun H, et al. A single-arm, multicenter, open-label phase 2 trial of surufatinib in patients with unresectable or metastatic biliary tract cancer. *Cancer*. 2021;127(21):3975-3984. [CrossRef]
12. Gerber DE, Schiller JH. Maintenance chemotherapy for advanced non-small-cell lung cancer: new life for an old idea. *J Clin Oncol*. 2013;31(8):1009-1020. [CrossRef]
13. Irie T. Intraarterial chemotherapy of liver metastases: implantation of a microcatheter-port system with use of modified fixed catheter tip technique. *J Vasc Interv Radiol*. 2001;12(109):1215-1218. [CrossRef]
14. Oh DY, Ruth He A, Qin S, et al. Durvalumab plus gemcitabine and cisplatin in advanced biliary tract cancer. *NEJM Evid*. 2022;1(8):EVIDoa2200015. [CrossRef]
15. Kelley RK, Ueno M, Yoo C, et al. Pembrolizumab in combination with gemcitabine and cisplatin compared with gemcitabine and cisplatin alone for patients with advanced biliary tract cancer (KEYNOTE-966): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet*. 2023;401(10391):1853-1865. [CrossRef]

16. Ueno M, Ikeda M, Morizane C, et al. Nivolumab alone or in combination with cisplatin plus gemcitabine in Japanese patients with unresectable or recurrent biliary tract cancer: a non-randomised, multicentre, open-label, phase 1 study. *Lancet Gastroenterol Hepatol*. 2019;4(8):611-621. [CrossRef]

17. Kemp A. Imfinzi plus chemotherapy significantly improved overall survival in 1stline Advanced Biliary Tract Cancer in Topaz-1 Phase Iii Trial at Interim Analysis. [CrossRef]

18. Valle JW, Furuse J, Jitlal M, et al. Cisplatin and gemcitabine for advanced biliary tract cancer: a meta-analysis of two randomised trials. *Ann Oncol*. 2014;25(2):391-398. [CrossRef]

19. Raderer M, Hejna MH, Valencak JB, et al. Two consecutive phase II studies of 5-fluorouracil/leucovorin/mitomycin C and of gemcitabine in patients with advanced biliary cancer. *Oncology*. 1999;56(3):177-180. [CrossRef]

20. Rogers JE, Law L, Nguyen VD, et al. Second-line systemic treatment for advanced cholangiocarcinoma. *J Gastrointest Oncol*. 2014;5(6):408-413. [CrossRef]

21. Fu J, Wang F, Dong LH, et al. Preclinical evaluation of the efficacy, pharmacokinetics and immunogenicity of JS-001, a programmed cell death protein-1 (PD-1) monoclonal antibody. *Acta Pharm Sin.* 2017;38(5):710-718. [CrossRef]

22. Raderer M, Hejna MH, Valencak JB, et al. Two consecutive phase II studies of 5-fluorouracil/leucovorin/mitomycin C and of gemcitabine in patients with advanced biliary cancer. *Oncology*. 1999;56(3):177-180. [CrossRef]

23. Vexler AM, Mou X, Gabizon AA, Gorodetsky R. Reduction of the systemic toxicity of cisplatin by intra-arterial hepatic route administration for liver malignancies. *Int J Cancer*. 1995;60(5):611-615. [CrossRef]

24. Tang B, Yan X, Sheng X, et al. Safety and clinical activity with an anti-PD-1 antibody JS001 in advanced melanoma or urologic cancer patients. *J Hematol Oncol*. 2019;12(1):7. [CrossRef]

25. Keam SJ. Toripalimab: first global approval. *Drugs.* 2019;79(5):573-578. [CrossRef]

26. Wang F, Wei XL, Wang FH, et al. Safety, efficacy and tumor mutational burden as a biomarker of overall survival benefit in chemo-refractory gastric cancer treated with toripalimab, a PD-1 antibody in phase Ib/II clinical trial NCT02915432. *Ann Oncol*. 2019;30(9):1479-1486. [CrossRef]

27. Yoshikawa D, Ojima H, Iwasaki M, et al. Clinicopathological and prognostic significance of EGFR, VEGF, and HER2 expression in cholangiocarcinoma. *Br J Cancer*. 2008;98(2):418-425. [CrossRef]

28. Javle M, Lowery M, Shroff RT, et al. Phase II study of BGJ398 in patients with FGFR-altered advanced cholangiocarcinoma. *J Clin Oncol*. 2018;36(3):276-282. [CrossRef]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242732

INTERVENTIONAL RADIOLOGY

ORIGINAL ARTICLE

Single-center 10-year retrospective analysis of Amplatzer Vascular Plug 4 embolization for pulmonary arteriovenous malformations with feeding arteries of <6 mm

Jung Guen Cha
Jongmin Park
Byunggeon Park
Seo Young Park
So Mi Lee
Jihoon Hong

Kyungpook National University, School of Medicine, Department of Radiology, Daegu, Republic of Korea

PURPOSE

To evaluate the efficacy and safety of Amplatzer Vascular Plug 4 (AVP4) embolization in pulmonary arteriovenous malformations (PAVMs) with small- to medium-sized feeding arteries (<6 mm) and to identify factors affecting persistence and the main persistence patterns after embolization.

METHODS

Between June 2013 and February 2023, we retrospectively reviewed 100 patients with 217 treated PAVMs. We included PAVMs with feeding arteries <6 mm, treated with AVP4 embolization, and followed adequately with computed tomography (CT). Technical success was defined as flow cessation observed on angiography. Persistence was defined as less than a 70% reduction of the venous sac on CT. We evaluated adverse events for each embolization session. Patterns of persistence were assessed using follow-up angiography. Univariate and multivariate analyses were performed to evaluate factors affecting persistence based on the 70% CT criteria.

RESULTS

Fifty-one patients (48 women, 3 men; mean age: 50.8 years; age range: 16–71 years) with 103 PAVMs met the inclusion criteria. The technical success rate was 100%. The persistence rate was 9.7% (10/103), and the overall adverse event rate was 2.9% (3/103) during a mean follow-up of 556 days (range: 181–3,542 days). In two cases, the persistence pattern confirmed by follow-up angiography involved reperfusion via adjacent pulmonary artery collaterals. The location of embolization relative to the last normal branch of the pulmonary artery was the only factor substantially affecting persistence.

CONCLUSION

Embolization with AVP4 appears to be safe and effective for small- to medium-sized PAVMs. The location of the embolization relative to the last normal branch of the pulmonary artery was found to be the main determinant of persistence.

CLINICAL SIGNIFICANCE

Given the increasing demand for the treatment of small PAVMs, AVP4 embolization could be considered a viable and effective option for managing PAVMs with feeding arteries <6 mm.

KEYWORDS

Arteriovenous malformation, computed tomography, embolization, pulmonary, vascular plug

Corresponding author: Jihoon Hong

E-mail: blushain@gmail.com

Received 29 February 2024; revision requested 02 April 2024; accepted 22 April 2024.

Epub: 03.06.2024

Publication date: 03.03.2025

DOI: 10.4274/dir.2024.242732

Pulmonary arteriovenous malformation (PAVM) describes a direct connection between the pulmonary artery and vein, which can lead to paradoxical embolism and result in serious complications, such as stroke and brain abscess.[1] Endovascular embolization has emerged as the preferred treatment for PAVM.[2] The once conventional "3 mm rule," which rec-

You may cite this article as: Cha JG, Park J, Park B, Park SY, Lee SM, Hong J. Single-center 10-year retrospective analysis of Amplatzer Vascular Plug 4 embolization for pulmonary arteriovenous malformations with feeding arteries of <6 mm. *Diagn Interv Radiol.* 2025;31(2):152-160.

ommended treating feeding arteries larger than 3 mm, no longer holds universal acceptance.[3,4] The current consensus now supports embolization for feeding arteries that are 2–3 mm or larger or when catheterization is feasible.[5] Nonetheless, the choice of embolic materials for small PAVMs remains limited, and the results from coil embolization in these cases are generally less favorable.[6]

The issue of persistence following PAVM embolization is substantial, often necessitating further interventions.[7] To address this, research has been conducted on the effectiveness of various embolic materials, including coils,[6,8] Amplatzer Vascular Plugs (AVPs),[9,10] and microvascular plugs (MVPs).[11,12] Despite the longstanding use of coils, their associated persistence rates are notably high.[6,8,13] While venous sac embolization yields favorable outcomes, employing multiple detachable coils is costly and extends procedural times.[14] More recently introduced MVPs have demonstrated promising results, although they are more expensive, and their long-term efficacy remains uncertain.[15]

AVPs are composed of a braided nitinol mesh and are noted for their low risk of migration in high-flow vessels or short landing zones, which permits device repositioning and provides the potential for single-device occlusion.[16,17] The latest generation, Amplatzer Vascular Plug 4 (AVP4), features a small-profile catheter with a 0.038-inch luminal diameter, suitable for navigating small- to medium-sized vessels and handling vascular tortuosity. Since its introduction for PAVM embolization in 2014, several studies have reported on the use of AVP4, with persistence rates ranging from 0% to 16%.[10,15,18] However, many of these studies have been limited by small sample sizes or the inclusion of different generations of AVP.

Main points

- Amplatzer Vascular Plug 4 embolization was performed on 103 pulmonary arteriovenous malformations (PAVMs) with small- to medium-sized feeding arteries (<6 mm). This resulted in a persistence rate of 9.7% (10/103) based on the 70% reduction criteria on computed tomography and an overall adverse event rate of 2.9% (3/103).
- Follow-up angiography conducted on 28 PAVMs identified persistence in 2 PAVMs, both of which showed reperfusion via adjacent pulmonary artery collaterals.
- The only substantial factor affecting persistence was the location of the embolization relative to the last normal branch of the pulmonary artery.

Consequently, this retrospective single-center study aims to evaluate the efficacy and safety of AVP4 embolization in PAVMs with small- to medium-sized feeding arteries (<6 mm). Additionally, this study seeks to identify factors affecting persistence and to delineate the main persistence patterns following AVP4 embolization.

Methods

This retrospective study received approval from the Institutional Review Board of Kyungpook National University Hospital (KNUH 2023-12-027). All participants provided informed consent prior to the procedure.

Patient selection

The study cohort included patients who underwent endovascular embolization for PAVM from June 2013 to February 2023. The eligibility criteria for inclusion were as follows: (1) treatment-naïve PAVM with a feeding artery diameter of <6 mm; (2) embolization performed using the AVP4; and (3) availability of both initial and follow-up computed tomography (CT) scans before and after embolization. The exclusion criteria were as follows: (1) underwent additional embolization sessions for the same lesion without an intervening follow-up CT; or (2) had a follow-up period of <6 months. Data on clinical history, physical examination, and PAVM characteristics were extracted from electronic medical records. Adverse events associated with the procedure during hospitalization and outpatient follow-up were also meticulously analyzed.

Embolization procedure

Vascular access was obtained via the right femoral vein, followed by intravenous administration of a heparin sodium bolus (3,000–5,000 IU; JW Pharmaceutical, Seoul, Korea). Subsequent pulmonary angiography facilitated the selective catheterization of the juxta-sac feeding artery using a coaxial system composed of a 6-Fr guiding catheter (Flexor Shuttle Guiding Sheath; Cook Medical, Bloomington, IN, USA) and a 5-Fr diagnostic catheter (Torcon NB Advantage, MPA; Cook Medical, or Glidecath, Angled Taper; Terumo, Tokyo, Japan). In cases involving challenging navigation due to small tortuous feeders, a triaxial system was employed, which included a 1.98-Fr microcatheter (Masters Parkway Soft; Asahi Intecc, Tokyo, Japan). The procedure began with the microcatheter, followed by the advancement of a 5-Fr hydrophilic-coated catheter over it. The size of the AVP4 (Abbott, Plymouth, MN, USA) ranged from 30% to 300% oversizing, based on preprocedural CT and selective angiography findings. The AVP4 was advanced into position within the 5-Fr catheter by pushing the guidewire to the catheter tip, and then the catheter was withdrawn to deploy the device. Proper placement of the plug was verified by injecting a contrast medium through the guiding catheter; if necessary, the plug was recaptured, repositioned, and redeployed. Depending on the operator's preference, additional coil embolization was performed occasionally to expedite flow cessation and provide reinforcement. Complete cessation of PAVM flow was confirmed in all patients through the completion of the digital subtraction angiography (DSA).

Acquisition and protocol for computed tomography and follow-up digital subtraction angiography

Initial and follow-up CT scans were primarily conducted using contrast-enhanced CT with multidetector-row scanners (Revolution EVO, Optima CT660, LightSpeed16; GE Healthcare, Chicago, IL, USA; SOMATOM Force, SOMATOM Definition Edge; Siemens Healthineers, Erlangen, Germany). For these examinations, a contrast agent (80–100 mL) was intravenously injected at a rate of 1.5–2 mL/s. CT images targeting the area of interest were reconstructed with a slice thickness of 2.5 mm in both transverse and coronal orientations. Follow-up CT scans were scheduled at 6 and 12 months post-embolization and subsequently every 2–3 years to monitor the persistence or resolution of PAVMs.[19]

DSA was conducted on previously treated PAVMs, particularly in cases in which multiple PAVMs were treated across separate sessions. The procedure typically began with either right or left pulmonary angiography, utilizing an injector with injection rates of 10–15 mL/s and volumes of 20–30 mL per injection. For more detailed assessments, selective angiography was performed at the segmental pulmonary artery levels, using injection rates of 3–5 mL/s and volumes of 9–15 mL. In certain instances, more precise visualization was achieved through meticulous manual injections at the distal levels of the pulmonary arteries.

Imaging analysis

All imaging obtained before, during, and following AVP4 embolization was reviewed by two experienced cardiovascular radiologists who were blinded to the outcomes of

PAVM embolization. Discrepancies between radiologists were resolved by consensus.

The analysis included reviewing the location, multiplicity, complexity (categorized as simple vs. complex), and original vessel diameters, along with their changes (feeding artery, venous sac, and draining vein) between the initial and final post-procedural CT scans. Changes in vessel diameter were quantified as reduction rates and recorded separately. Additionally, the origin of the last normal branch of the pulmonary artery was documented on the initial CT and during procedural DSA,[20] and the embolization location relative to this branch (either proximal or distal) was confirmed on post-procedural CT. The distance from the plug to the sac was also evaluated using pre- and post-procedural CT scans.

During the procedural imaging of AVP4 embolization, the size and number of plugs, the plug oversizing ratio, the type and number of additional coils, and the procedure time were all documented. Technical success was defined as the complete cessation of flow in the PAVM upon completion of the DSA. Treatment outcomes were assessed using the widely accepted CT criteria, where occlusion was defined as a 70% reduction in the venous sac in pre- and post-procedure comparisons (referred to as the 70% CT criteria).[21] Persistence was noted when the reduction rate of the venous sac was less than 70%. Procedure time was recorded from the femoral vein puncture to the completion of angiography, exclusively for sessions treating a single PAVM to ensure accurate assessment.

Adverse events were classified according to the Society of Interventional Radiology standards.[22] Both peri-procedural and post-procedural adverse events were documented for each embolization session.

To investigate factors affecting persistence, variables such as sex, age, smoking history, use of antithrombotic agents, lobar location, complexity, multiplicity, feeding artery diameter, venous sac diameter, plug oversizing ratio, sac-to-plug distance, embolization location relative to the last normal branch, and additional coil embolization were evaluated.

Angiographically confirmed cases by follow-up DSA were analyzed to determine patterns of persistence. Persistence was classified as resulting from (a) recanalization of a previously treated feeding artery, (b) reperfusion via adjacent pulmonary artery collaterals, or (c) the presence of a previously unrecognized feeder (incomplete treatment).[7] Reperfusion from systemic arteries was not assessed.

Statistical analysis

Continuous variables were expressed as the mean and range, whereas categorical variables were reported as the frequency (percentage). Multivariate logistic regression analysis was performed to identify factors affecting persistence using odds ratios (OR) and confidence intervals (CI). This analysis utilized the R software package (version 4.0.3, The R Foundation for Statistical Computing, Vienna, Austria). Variables that achieved a *P* value of <0.20 in the univariate analysis were selected as input variables for the multivariate analysis, which was conducted using a backward stepwise method. A *P* value of <0.05 was considered statistically significant.

Results

Fifty-one patients [48 women and 3 men; mean age: 50.8 years (range: 16–71)] with 103 PAVMs met the inclusion criteria and were included in the analysis (Figure 1). Among these patients, 9 (17.6%) exhibited symptoms of hereditary hemorrhagic telangiectasia, and 22 (43.1%) presented with symptoms attributable to PAVM. Sixteen patients (31.3%) had multiple PAVMs, averaging 2.26 lesions per patient (range: 1–10). Of the 103 PAVMs analyzed, 97 (94.2%) were classified as simple, with the remaining identified as complex. The mean diameter of the feeding arteries was 3.00 mm (range: 1.50–5.70 mm). The mean follow-up period was 556 days (range: 181–3,542 days). The characteristics of the patients and the PAVMs are summarized in Table 1.

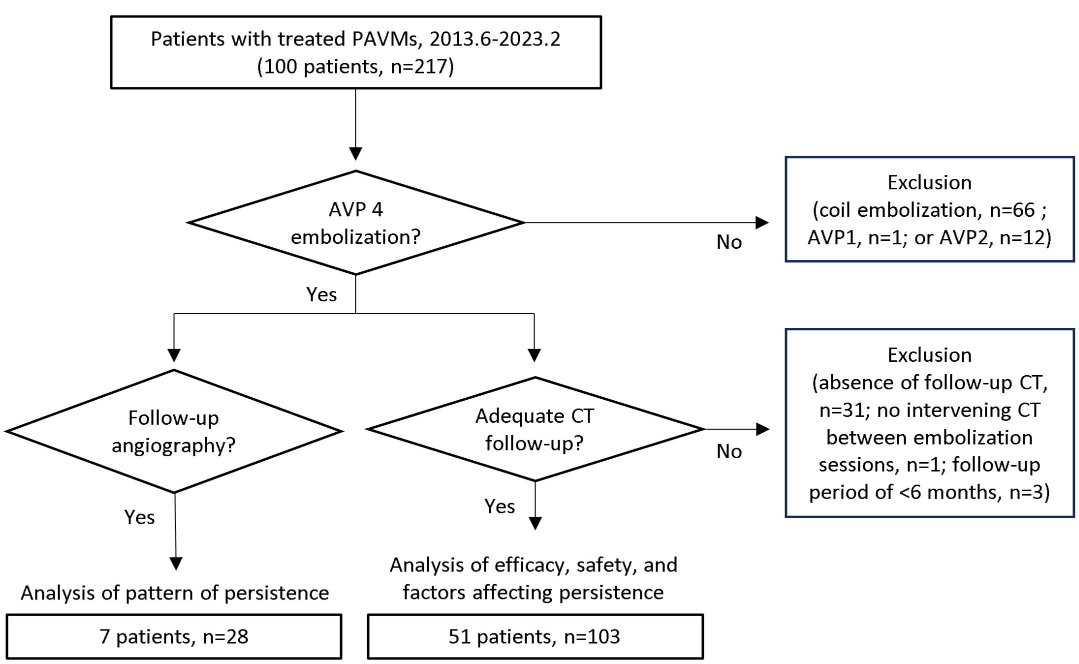


Figure 1. Flow chart summarizing patient enrollment according to study eligibility criteria. PAVM, pulmonary arteriovenous malformation; AVP, amplatzer vascular plug; CT, computed tomography.

All 103 PAVMs were successfully treated with AVP4 embolization across 59 sessions (Figure 2). On average, 1.75 PAVMs were treated per session (range: 1–8). The mean size and number of AVP4 devices used per PAVM were 6.34 mm and 1.09, respectively. Additional coils were used in 9 PAVMs (8.7%), with an average of 2.89 coils per PAVM (range: 1–5). The mean procedure time for sessions treating a single PAVM was approximately 39.62 minutes (range: 18–96 minutes). Details of the AVP4 embolization procedures are summarized in Table 2.

Table 1. Patient demographics and characteristics of pulmonary arteriovenous malformation

Parameters	Value
Patient factor (n = 51)	
Sex (men/women)	3 (5.8)/48 (94.2)
Mean age (range) in years	50.8 (16–71)
Presence of HHT symptoms	9 (17.6)
Symptomatic patients	22 (43.1)
Respiratory	11 (21.6)
Stroke	8 (15.7)
Brain abscess	4 (7.8)
Smoking history	8 (15.7)
Use of antithrombotic agents	9 (17.6)
Multiple PAVMs	16 (31.3)
Mean number of PAVMs per patient (range)	2.26 (1–10)
PAVM factor (n = 103)	
Simple/complex	97 (94.2)/6 (5.8)
Lobar location	
RUL/RML/RLL	16 (15.5)/27 (26.2)/20 (19.4)
LUL/LLL	16 (15.5)/24 (23.3)
Mean feeding artery diameter (range) (mm)	3.00 (1.50–5.70)
<2 mm	12 (11.7)
<3 mm, ≥2 mm	54 (52.4)
<6 mm, ≥3 mm	37 (35.9)
Mean venous sac diameter (range) (mm)	6.91 (2.40–22.25)
Origin of last normal branch	
Sac/junction/proximal feeding artery	27 (26.2)/41 (39.8)/35 (34)
Mean follow-up periods (range) (day)	556 (181–3542)

Data represent the number of patients or PAVMs, with percentages in parentheses unless specified otherwise. HHT, hereditary hemorrhagic telangiectasia; PAVMs, pulmonary arteriovenous malformations; RUL, right upper lobe; RML, right middle lobe; RLL, right lower lobe; LUL, left upper lobe; LLL, left lower lobe.

Table 2. Details of AVP4 embolization (59 sessions for 103 pulmonary arteriovenous malformations)

Embolization factor (n = 103)	Value
Mean number of AVP4 per PAVM (range)	1.09 (1–2)
Mean size of AVP4 (range) (mm)	6.34 (4–8)
Mean plug oversizing ratio (range) (%)	122.4 (35–300)
Mean plug-to-sac distance (range) (mm)	3.90 (0–26.0)
>10 mm	12 (11.7%)
≤10 mm	91 (88.3%)
Embolization location relative to the last normal branch	
Proximal	47 (45.6%)
Distal	56 (54.4%)
Additional coil embolization	9 (8.7%)
Mean number of additional coils (range)	2.88 (1–5)

Data represent the number of PAVMs with percentages in parentheses unless specified otherwise. PAVM, pulmonary arteriovenous malformation; AVP4, Amplatzer Vascular Plug 4.

The technical success rate for AVP4 embolization was 100%. The persistence rate of the treated PAVMs, using the 70% CT criteria, was 9.7% (10/103). Stratified by embolization type, the persistence rates were 9.6% (9/94) for AVP4 alone and 11.1% (1/9) for AVP4 combined with coil embolization. During the 59 sessions for 103 PAVMs, three mild adverse events were reported (5.1% per session): two instances of self-limiting pleuritic chest pain and one case of transient bradycardia. There were no severe adverse events, with an overall adverse event rate of 2.9% per PAVM lesion.

Follow-up DSA was conducted for 28 (27.2%) of the 103 PAVMs at a mean interval of 436 days. Among these, occlusion was observed in 26 PAVMs, whereas the remaining 2 (7.1%) exhibited persistence due to reperfusion via adjacent pulmonary artery collaterals (Figure 3). When comparing outcomes between DSA and the 70% CT criteria, 25 out of 26 angiographically occluded PAVMs showed venous sac reductions exceeding 70% on CT, resulting in concordant findings. However, one PAVM demonstrated a reduction rate of 57.3%, leading to discordance between the two modalities. The two angiographically reperfused PAVMs showed venous sac reductions of 34.7% and 49.2%, respectively, aligning the findings across both modalities.

In both univariate and multivariate analyses, the location of embolization relative to the last normal branch of the pulmonary artery was identified as the only significant factor affecting persistence (OR: 0.18; 95% CI: 0.03–0.81; $P < 0.05$) (Table 3).

Discussion

The findings of this study affirm the efficacy and safety of AVP4 embolization for small- to medium-sized PAVMs with diameters of <6 mm, showing a persistence rate of 9.7% (10/103) based on the 70% CT criteria and an overall adverse event rate of 2.9% during an average follow-up period of 556 days. Follow-up DSA, conducted in 27.1% of this cohort, revealed persistence in 2 PAVMs, predominantly due to reperfusion via adjacent pulmonary artery collaterals. The location of embolization relative to the last normal branch of the pulmonary artery was identified as the only substantial factor affecting persistence according to the CT criteria.

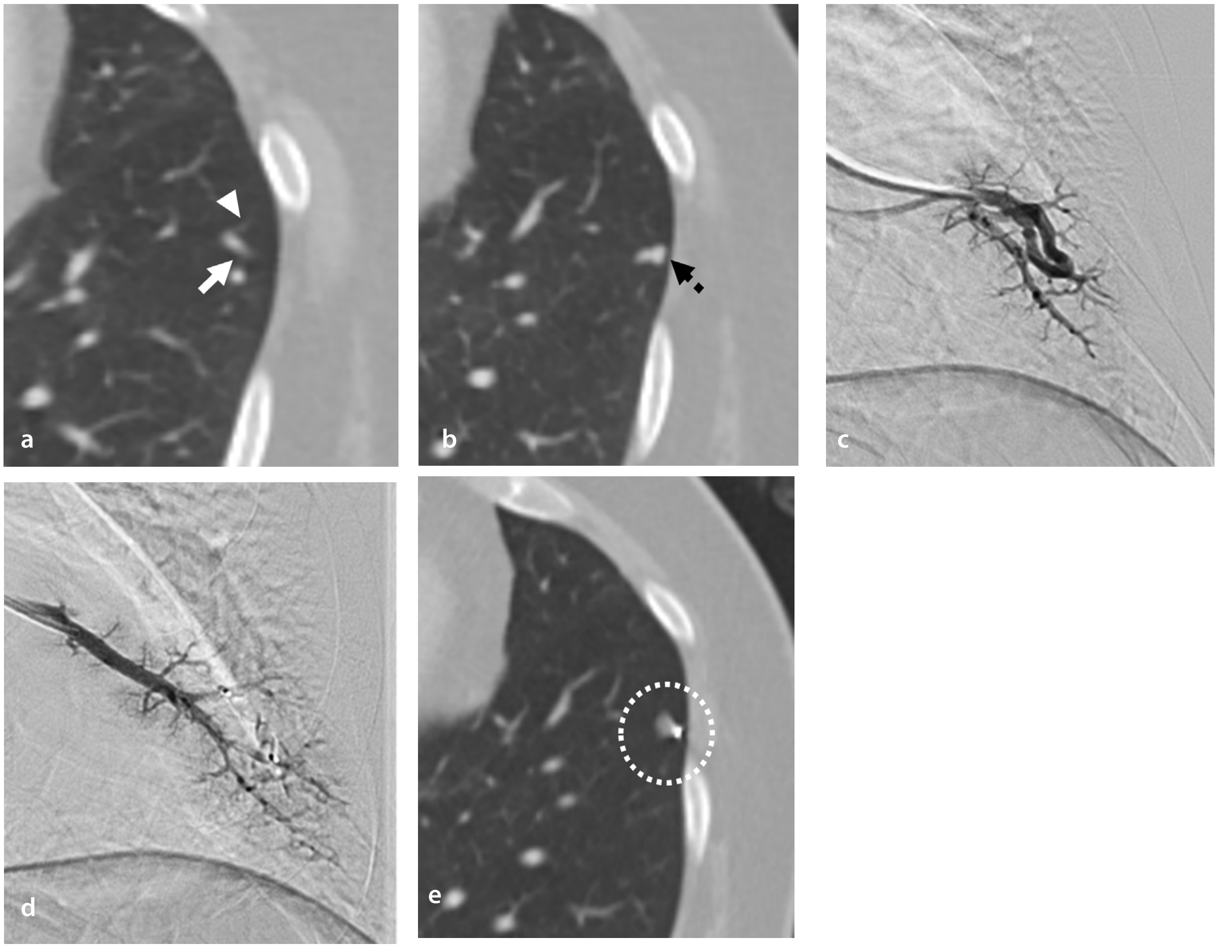


Figure 2. A 51-year-old woman with an incidentally detected simple pulmonary arteriovenous malformation (PAVM). **(a, b)** Pre-embolization computed tomography (CT) images show the distal feeding artery and venous sac of a simple PAVM located in the left lower lobe (LLL). The vessel diameters are as follows: feeding artery (arrow in **a),** 1.53 mm, and venous sac (dotted arrow in **b**), 3.63 mm. The last normal branch of the pulmonary artery (arrowhead in **a**) is identified within the junction between the feeding artery and the sac. **(c)** Angiography conducted after selecting the distal feeding artery shows a simple PAVM in the LLL. **(d)** Completion angiography following the deployment of a 6 mm Amplatzer Vascular Plug 4 (AVP4) in the juxta sac-feeding artery shows complete occlusion of the PAVM with no residual shunt flow. Notably, the embolization location is distal to the last normal branch of the pulmonary artery. **(e)** A CT scan performed at a 2-year follow-up shows the disappearance of the venous sac (dotted circle), with a venous sac reduction rate of 100%. Only AVP4 is visible.

Various generations of AVPs have been employed for PAVM embolization, with reported persistence rates ranging from 0% to 16%.[9,10,15,18,23-25] Some studies have suggested superior outcomes with AVP compared with coils.[26,27] Nonetheless, there remains a scarcity of studies specifically focusing on AVP4. Rabellino et al.[18] defined a successful outcome as a venous sac reduction of ≥30% in their early experience with 7 patients, achieving success across all cases over an average follow-up of 20.1 months. A more recent study in 2019[10] involving 19 PAVMs reported a persistence rate of 16% using 70% CT criteria over an average follow-up of 14 months. Ratnani et al.[15] specifically analyzed AVP4 and reported a persistence rate of 12.5% (1/8) over an average follow-up of 1,239 days, defining persistence based on sustained sac perfusion observed in CT angiography (CTA) or pulmonary angiography. While the outcomes of these small case series generally align with those of the current study, varying assessment criteria make precise comparisons challenging.

Pulmonary angiography is considered the gold standard, but it poses difficulties for routine use due to its invasiveness.[19] The use of sac perfusion on CTA to assess persistence raises concerns about retrograde venous filling from adjacent normal branches.[13] Presently, the 70% CT criteria are the most widely adopted, yet recent discussions highlight concerns regarding their specificity.[13,28,29] Additional research and consensus are necessary to refine and agree on criteria that address these concerns effectively.

The recently introduced MVP has demonstrated superior results compared with the AVP, boasting a low persistence rate of 0%–6%.[11,12,15,30] AVP, composed of a finely braided nitinol mesh, exhibits several structural challenges in comparison to MVP, which features a polytetrafluoroethylene (PTFE)-covered nitinol cage. Particularly, introducing a 5-Fr catheter up to the juxta-sac feeding artery in cases of very small or tortuous feeding arteries can be technically challenging compared with using a 2.4- or 2.8-Fr microcatheter, as utilized with MVP. In our practice, we often overcome this challenge by using a hydrophilic-coated 5-Fr catheter (Glidecath, Angled Taper; Terumo) with appropriate angulation and advancing it over the microcatheter.

A concern exists that AVP4 may become lodged within this soft and flexible 5-Fr catheter during delivery. To address this issue, we primarily employ smaller-sized AVP4s (4–6 mm) -adequate for most small-sized feeding arteries- and advance a 6-Fr guiding

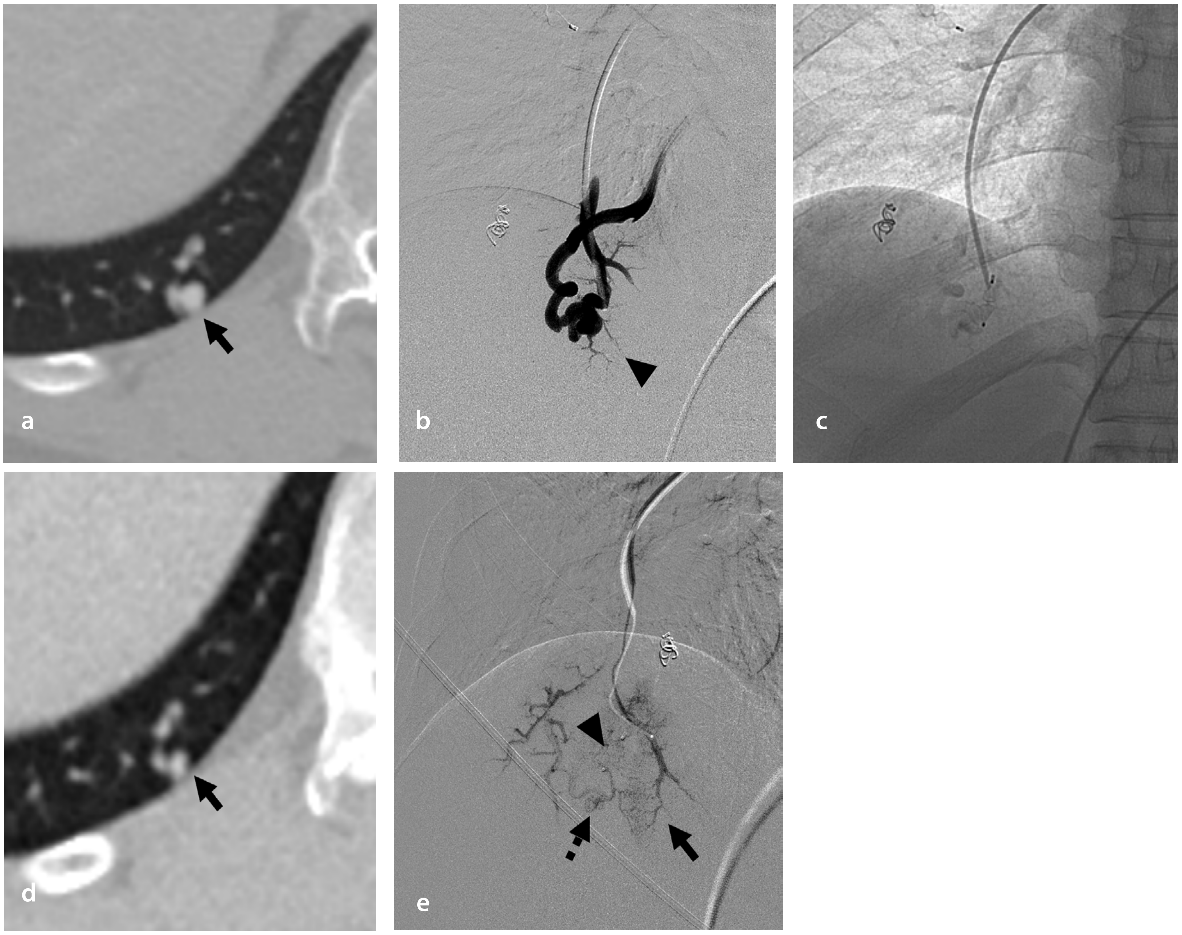


Figure 3. A 57-year-old woman with definite hereditary hemorrhagic telangiectasia and multiple pulmonary arteriovenous malformations (PAVM) (at least 6) in both lungs. **(a)** Pre-embolization computed tomography (CT) image shows a simple PAVM in the right lower lobe. The vessel diameters are as follows: feeding artery, 3.12 mm, and venous sac (arrow), 4.75 mm. **(b)** Angiography conducted after superselecting the distal feeding artery shows a simple PAVM. The last normal branch of the pulmonary artery is identified within the venous sac (arrowhead). **(c)** A 5 mm Amplatzer Vascular Plug 4 is deployed in the juxta sac-feeding artery of the PAVM. Notably, the embolization location is proximal to the last normal branch of the pulmonary artery. **(d)** CT performed at the 3-year follow-up shows a reduction in the diameter of the venous sac (arrow) to 3.1 mm, representing a reduction rate of 34.7%. **(e)** Subsequent angiography shows successful occlusion of the previously treated feeding artery (arrowhead). However, contrast opacification of the venous sac (dotted arrow) is observed due to reperfusion via adjacent pulmonary artery collateral (arrow).

catheter as distally as possible to provide support while routinely performing continuous saline flushing in the catheter to minimize friction between the plug and catheter wall. Furthermore, unlike MVP, which induces immediate flow cessation due to its PTFE cover, AVP4 relies on inducing thrombosis through its nitinol mesh, requiring patience and repeated monitoring for occlusion. The patient's coagulation status may influence this process and raise concerns about the potential migration of *in-situ* thrombus on the device surface, leading to paradoxical embolism.[16,17] To mitigate these risks, we employ a strategy of reinforcement with several additional coils if flow cessation is not achieved within 5–10 minutes or by confirming flow cessation collectively after completing treatment for all PAVMs in cases of multiple PAVMs to save time. Consequently, we achieved a relatively short procedure time (mean: 39.62 minutes), and no procedure-related paradoxical embolisms were reported.

On the financial side, AVP4 offers a more cost-effective alternative than MVP. The mean number of AVP4 devices used in this study, 1.1 per PAVM, is comparable to the 1.1–1.3 used in previous MVP studies[11,12] despite the substantially higher cost of the MVP device.[15] Additionally, the routine use of a microcatheter for MVP delivery adds to overall expenses. While MVP has not yet received approval for use in many countries, including ours, AVP4 remains a favorable option in centers where it is available, offering both clinical efficacy and cost-effectiveness.

In this cohort, the majority of PAVMs featured small-sized feeding arteries, with 64.1% measuring less than 3 mm and 11.7% measuring less than 2 mm. Stein et al.[6] reported on coil embolization for 141 PAVMs with feeding arteries smaller than 3 mm; the persistence rate noted was 21%, which is higher than the 10% reported in other studies targeting PAVMs with feeding arteries of 3 mm or larger. However, in our study, there was no substantial difference in persistence rates between PAVMs with feeding arteries of 3 mm or smaller (9.4%) and those larger than 3 mm (10.3%). This outcome may highlight the advantage of AVP4 over coils, as AVP4 allows for sufficient oversizing and smooth delivery if the catheter reaches the target vessel, regardless of vessel size. In the case of the MVP, there are reports of successful treatments for feeding arteries as small as 1.3 mm;[12] however, there is a lack of studies focusing on small PAVMs or evaluating long-term outcomes. Under these circumstances, AVP4 emerges as a favorable treatment option for small PAVMs.

Table 3. Univariate and multivariate analyses of factors affecting persistence based on 70% CT criteria

Factors	Univariate analysis			Multivariate analysis		
	OR	95% CI	*P* value	OR	95% CI	*P* value
Sex			0.025			0.120
Men	1.00	Reference		1.00	Reference	
Women	0.15	0.03, 0.86		0.25	0.04, 1.56	
Age	0.96	0.91, 1.00	0.069	Stepwise eliminated		
Smoking history			0.773			
Yes	1.00	Reference				
No	1.23	0.32, 5.99				
Antithrombotic agent			0.632			
Yes	1.00	Reference				
No	1.69	0.28, 32.61				
Multiplicity			0.373			
Single	1.00	Reference				
Multiple	2.08	0.49, 14.31				
Complexity			0.460			
Simple	1.00	Reference				
Complex	2.43	0.11, 21.23				
Lobar location			0.163	Stepwise eliminated		
Upper or middle lobe	1.00	Reference				
Lower lobe	2.54	0.71, 10.40				
Feeding artery diameter	1.30	0.58, 2.75	0.501			
Venous sac diameter	0.99	0.80, 1.16	0.889			
Sac to plug distance	1.10	0.99, 1.22	0.067	Stepwise eliminated		
Plug oversizing ratio	1.00	0.98, 1.01	0.751			
Additional embolization			0.903			
Yes	1.00	Reference				
No	1.11	0.25, 7.75				
Location of embolization			0.018			0.041
Proximal to LNB	1.00	Reference		1.00	Reference	
Distal to LNB	0.14	0.02, 0.61		0.18	0.03, 0.81	

OR, odds ratio; CI, confidence interval; LNB, last normal branch of pulmonary artery.

After coil embolization, recanalization through a previously treated feeder is the predominant persistence pattern, reported to exceed 90%.[7] Factors such as coil packing density, the use of oversized coils, and the distance between the coil and the venous sac have been identified as substantial factors affecting persistence rates.[6,31,32] In a recent study by Shimohira et al.[13], the location of embolization relative to the last normal branch of the pulmonary artery was determined to be a substantial factor in persistence, as assessed by CT, time-resolved MR angiography, and DSA. However, similar detailed studies focusing on AVP are lacking.

In this study, reperfusion via adjacent pulmonary artery collaterals was observed in both cases where angiographically confirmed persistence occurred, specifically when proximal embolization was performed because the last normal branch was within the sac. This location was the only substantial factor affecting persistence. In this reperfusion mechanism, the shunt or feeder size is usually very small, rendering additional treatment technically challenging and generally less successful than the recanalization pattern.[7,8] Although this study highlighted the excellent cross-sectional occlusion capabilities of AVP4, achieving complete prevention of persistence in PAVMs where the last normal branch is located within the sac may ultimately require sac embolization.[13,20,33] There are documented cases in which successful outcomes were achieved through venous sac coiling combined with feeding artery plug embolization in such scenarios.[34] Nonetheless, further studies involving a larger cohort are necessary to validate these findings and refine treatment protocols.

Some limitations of this study should be acknowledged. First, it was a retrospective study with a relatively small sample size. Second, owing to the widespread availability of chest CT scans and health screenings, most patients in the study presented with incidentally detected simple PAVMs. Given that treatment outcomes are less favorable for complex PAVMs, the persistence rate of these malformations may have been underestimated. Moreover, follow-up DSA was performed only in patients with multiple PAVMs, which introduced potential bias. Additionally, reperfusion via the systemic artery was not evaluated. Lastly, an important variable related to the use of AVP -occlusion time- was not measured.

In conclusion, AVP4 embolization proved to be safe and effective for treating small- to medium-sized PAVMs (<6 mm), demonstrating a low persistence rate based on the 70% CT criteria. The primary pattern observed in angiographically confirmed persistence was reperfusion via adjacent pulmonary artery collaterals. Concerning treatment outcomes based on CT criteria, the only factor affecting persistence was the location of the embolization relative to the last normal branch of the pulmonary artery.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Majumdar S, McWilliams JP. Approach to pulmonary arteriovenous malformations: a comprehensive update. *J Clin Med*. 2020;9(6):1927. [CrossRef]
2. White RI Jr, Pollak JS, Wirth JA. Pulmonary arteriovenous malformations: diagnosis and transcatheter embolotherapy. *J Vasc Interv Radiol*. 1996;7(6):787-804. [CrossRef]
3. Trerotola SO, Pyeritz RE, Bernhardt BA. Outpatient single-session pulmonary arteriovenous malformation embolization. *J Vasc Interv Radiol*. 2009;20(10):1287-1291. [CrossRef]
4. Todo K, Moriwaki H, Higashi M, Kimura K, Naritomi H. A small pulmonary arteriovenous malformation as a cause of recurrent brain embolism. *AJNR Am J Neuroradiol*. 2004;25(3):428-430. [CrossRef]
5. Müller-Hülsbeck S, Marques L, Maleux G, et al. CIRSE standards of practice on diagnosis and treatment of pulmonary arteriovenous malformations. *Cardiovasc Intervent Radiol*. 2020;43(3):353-361. [CrossRef]
6. Stein EJ, Chittams JL, Miller M, Trerotola SO. Persistence in coil-embolized pulmonary arteriovenous malformations with feeding artery diameters of 3 mm or less: a retrospective single-center observational study. *J Vasc Interv Radiol*. 2017;28(3):442-449. [CrossRef]
7. Woodward CS, Pyeritz RE, Chittams JL, Trerotola SO. Treated pulmonary arteriovenous malformations: patterns of persistence and associated retreatment success. *Radiology*. 2013;269(3):919-926. [CrossRef]
8. Shimohira M, Kawai T, Hashizume T, et al. Reperfusion rates of pulmonary arteriovenous malformations after coil embolization: evaluation with time-resolved MR angiography or pulmonary angiography. *J Vasc Interv Radiol*. 2015;26(6):856-864. [CrossRef]
9. Letourneau-Guillon L, Faughnan ME, Soulez G, et al. Embolization of pulmonary arteriovenous malformations with amplatzer vascular plugs: safety and midterm effectiveness. *J Vasc Interv Radiol*. 2010;21(5):649-656. [CrossRef]
10. Lee SY, Lee J, Kim YH, et al. Efficacy and Safety of Amplatzer Vascular Plug Type IV for embolization of pulmonary arteriovenous malformations. *J Vasc Interv Radiol*. 2019;30(7):1082-1088. [CrossRef]
11. Bailey CR, Arun A, Towsley M, et al. MVP™ micro vascular plug systems for the treatment of pulmonary arteriovenous malformations. *Cardiovasc Intervent Radiol*. 2019;42(3):389-395. [CrossRef]
12. Mahdjoub E, Tavolaro S, Parrot A, Cornelis F, Khalil A, Carette MF. Pulmonary arteriovenous malformations: safety and efficacy of microvascular plugs. *AJR Am J Roentgenol*. 2018;211(5):1135-1143. [CrossRef]
13. Shimohira M, Kiyosue H, Osuga K, et al. Location of embolization affects patency after coil embolization for pulmonary arteriovenous malformations: importance of time-resolved magnetic resonance angiography for diagnosis of patency. *Eur Radiol*. 2021;31(7):5409-5420. [CrossRef]
14. Hayashi S, Baba Y, Senokuchi T, Nakajo M. Efficacy of venous sac embolization for pulmonary arteriovenous malformations: comparison with feeding artery embolization. *J Vasc Interv Radiol*. 2012;23(12):1581. [CrossRef]
15. Ratnani R, Sutphin PD, Koshti V, et al. Retrospective comparison of pulmonary arteriovenous malformation embolization with the polytetrafluoroethylene-covered nitinol microvascular plug, Amplatzer plug, and coils in patients with hereditary hemorrhagic telangiectasia. *J Vasc Interv Radiol*. 2019;30(7):1089-1097. [CrossRef]
16. Lopera JE. The Amplatzer vascular plug: review of evolution and current applications. *Semin Intervent Radiol*. 2015;32(4):356-369. [CrossRef]
17. Wang W, Li H, Tam MD, Zhou D, Wang DX, Spain J. The amplatzer vascular plug: a review of the device and its clinical applications. *Cardiovasc Intervent Radiol*. 2012;35(4):725-740. [CrossRef]
18. Rabellino M, Serra M, Peralta O, et al. Early experience with the Amplatzer vascular plug IV for the occlusion of pulmonary arteriovenous malformations. *J Vasc Interv Radiol*. 2014;25(9):1333-1337. [CrossRef]
19. Faughnan ME, Mager JJ, Hetts SW, Palda VA, Ratjen F. Second International Guidelines for the Diagnosis and Management of Hereditary Hemorrhagic Telangiectasia. *Ann Intern Med*. 2020;173(12):989-1001. [CrossRef]
20. Maruno M, Kiyosue H, Hongo N, Matsumoto S, Mori H. Where is the origin of the last normal branch from feeding artery of pulmonary arteriovenous malformations? *Cardiovasc Intervent Radiol*. 2018;41(12):1849-1856. Erratum in: *Cardiovasc Intervent Radiol*. 2019;42(10):1510. [CrossRef]
21. Lee DW, White RI Jr, Egglin TK, et al. Embolotherapy of large pulmonary arteriovenous malformations: long-term results. *Ann Thorac Surg*. 1997;64(4):930-940. [CrossRef]
22. Khalilzadeh O, Baerlocher MO, Shyn PB, et al. Proposal of a new adverse event classification by the Society of Interventional Radiology Standards of Practice Committee. *J Vasc Interv Radiol*. 2017;28(10):1432-1437. Erratum in: *J Vasc Interv Radiol*. 2018;29(1):146. [CrossRef]
23. Hart JL, Aldin Z, Braude P, Shovlin CL, Jackson J. Embolization of pulmonary arteriovenous malformations using the Amplatzer vascular plug: successful treatment of 69 consecutive patients. *Eur Radiol*. 2010;20(11):2663-2670. [CrossRef]
24. Kucukay F, Özdemir M, Şenol E, Okten S, Ereren M, Karan A. Large pulmonary arteriovenous malformations: long-term results of embolization with Amplatzer vascular plugs. *J Vasc Interv Radiol*. 2014;25(9):1327-1332. [CrossRef]
25. Abdel Aal AK, Ibrahim RM, Moustafa AS, Hamed MF, Saddekni S. Persistence of pulmonary arteriovenous malformations after successful embolotherapy with Amplatzer vascular plug: long-term results. *Diagn Interv Radiol*. 2016;22(4):358-364. [CrossRef]
26. Andersen PE, Duvnjak S, Gerke O, Kjeldsen AD. Long-term single-center retrospective follow-up after embolization of pulmonary arteriovenous malformations treated over a 20-year period: frequency of re-canalization with various embolization materials and clinical outcome. *Cardiovasc Intervent Radiol*. 2019;42(8):1102-1109. [CrossRef]
27. Tau N, Atar E, Mei-Zahav M, et al. Amplatzer Vascular plugs versus coils for embolization of pulmonary arteriovenous malformations in patients with hereditary hemorrhagic telangiectasia. *Cardiovasc Intervent Radiol*. 2016;39(8):1110-1114. [CrossRef]
28. Hong J, Lee SY, Cha JG, et al. Pulmonary arteriovenous malformation (PAVM) embolization: prediction of angiographically-confirmed recanalization according to PAVM Diameter changes on CT. *CVIR Endovasc*. 2021;4(1):16. [CrossRef]
29. Belanger C, Chartrand-Lefebvre C, Soulez G, et al. Pulmonary arteriovenous malformation (PAVM) reperfusion after percutaneous

embolization: sensitivity and specificity of non-enhanced CT. *Eur J Radiol*. 2016;85(1):150-157. [CrossRef]

30. Latif MA, Bailey CR, Motaghi M, et al. Postembolization Persistence of pulmonary arteriovenous malformations: a retrospective comparison of coils and amplatzer and micro vascular plugs using propensity score weighting. *AJR Am J Roentgenol*. 2023;220(1):95-103. Erratum in: *AJR Am J Roentgenol*. 2023;220(1):155. [CrossRef]
31. Milic A, Chan RP, Cohen JH, Faughnan ME. Reperfusion of pulmonary arteriovenous malformations after embolotherapy. *J Vasc Interv Radiol*. 2005;16(12):1675-1683. [CrossRef]
32. Pollak JS, Saluja S, Thabet A, Henderson KJ, Denbow N, White RI Jr. Clinical and anatomic outcomes after embolotherapy of pulmonary arteriovenous malformations. *J Vasc Interv Radiol*. 2006;17(1):35-45. [CrossRef]
33. Nagai K, Osuga K, Kashiwagi E, et al. Venous sac and feeding artery embolization versus feeding artery embolization alone for treating pulmonary arteriovenous malformations: draining vein size outcomes. *J Vasc Interv Radiol*. 2021;32(7):1002-1008. [CrossRef]
34. Hundt W, Kalinowski M, Kiessling A, et al. Novel approach to complex pulmonary arteriovenous malformation embolization using detachable coils and Amplatzer vascular plugs. *Eur J Radiol*. 2012;81(5):732-738. [CrossRef]