Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.264561

ARTIFICIAL INTELLIGENCE AND INFORMATICS

REVIEW

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

Comparative reasoning in radiology artificial intelligence: longitudinal and reference-case approaches

Burak Kocak[1]
Ismail Mese[2]

[1]Basaksehir Cam and Sakura City Hospital, Department of Radiology, İstanbul, Türkiye

[2]Uskudar State Hospital, Department of Radiology, İstanbul, Türkiye

ABSTRACT

Artificial intelligence (AI) in radiology remains predominantly examination-centric, whereas clinical interpretation frequently depends on comparison across time or with relevant prior cases. This narrative review examines comparative reasoning in radiology AI through two related but clinically distinct forms: longitudinal reasoning, which compares examinations from the same patient over time, and reference-case reasoning, which uses clinically analogous cases from other patients to support interpretation. We propose a capability hierarchy that distinguishes simple access to comparative information from higher-level capabilities, including selection of relevant historical context, anatomical and lesion-level correspondence, characterization of interval change, temporal-state interpretation, multi-interval trajectory reasoning, and entity-specific assessment of clinical similarity. This distinction matters because providing prior examinations or retrieving visually similar cases does not itself demonstrate meaningful comparative reasoning. We review developments in prior-examination selection, registration, lesion tracking, temporal vision–language modeling, treatment-response assessment, analogous-case retrieval, and longitudinal report generation. Particular emphasis is placed on evaluation strategies that determine whether model performance depends appropriately on clinically relevant comparative information rather than on additional input volume, global image similarity, acquisition characteristics, or other shortcuts. Key challenges include incorrect prior selection, lesion mismatching, temporal inconsistency, clinically irrelevant reference-case retrieval, data leakage, misleading comparative language, and propagation of historical errors. Future progress will require models that preserve entity identity, reason coherently across multi-visit imaging histories, identify clinically informative reference cases, and demonstrate benefit under realistic clinical conditions. Comparative reasoning is therefore a necessary step in the transition from examination-level toward patient-level radiology AI.

KEYWORDS

Artificial intelligence, longitudinal reasoning, comparative reasoning, reference-case reasoning, diagnostic imaging, medical imaging, radiology

Clinical interpretation in radiology is frequently comparative. Radiologists routinely compare a current examination with prior imaging to assess interval change and better understand the disease state; professional reporting guidance explicitly recommends comparison with prior examinations when available.[1,2] Such comparison is central to common tasks, including assessing lesion growth, treatment response, recurrence, disease progression, and distinguishing progression or recurrence from treatment-related change.[2,3] However, the dominant development paradigm in radiology artificial intelligence (AI) has historically centered on individual examinations, with detection, triage, classification, and related tasks performed predominantly at a single time point.[2]

Assessment of an examination alone captures only part of the longitudinal comparison workflow. Providing a model with a prior examination does not, by itself, constitute a longitudinal comparison process. The process is a set of linked operations: selection of relevant historical examinations, identification of corresponding anatomy or lesions across examinations, and assessment of stability or interval change.[4] Each of these can fail independently.

Corresponding author: Burak Kocak

E-mail: drburakkocak@gmail.com

Received 10 October 2026; revision requested 18 October 2026; accepted 25 October 2026.

Epub: 02.11.2026

DOI: 10.4274/dir.2026.264561

You may cite this article as: Kocak B, Mese I. Comparative reasoning in radiology artificial intelligence: longitudinal and reference-case approaches. *Diagn Interv Radiol.* 02 November 2026 DOI: 10.4274/dir.2026.264561 [Epub Ahead of Print].

The problem compounds when several prior examinations are available, because the model must then identify not only the relevant comparator examination but also the relevant series within it and the corresponding finding location.[4]

Recent primary studies illustrate the breadth of this emerging field. These include tumor-preserving longitudinal registration and automated lesion tracking,[5,6] temporal vision–language models that explicitly incorporate previous images and reports,[7] and benchmarks designed to test reasoning across multiple imaging intervals rather than isolated current–prior pairs.[8] The demonstrated value of additional historical information, however, is not uniform. For instance, in screening mammography, radiologists' comparison with two or more prior examinations was associated with lower recall, higher positive predictive value of recall, and a higher cancer detection rate than comparison with a single prior.[9] In contrast, a recent digital breast tomosynthesis AI study found similar discrimination with and without prior examinations, with priors providing only limited additional benefit, mainly through slightly improved specificity.[10] These contrasting examples indicate that the value of longitudinal information is task- and system-dependent and cannot be assumed from its availability alone.

This narrative review synthesizes work across medical image registration, lesion tracking, vision–language modeling, report generation, image retrieval, and clinical decision support through the lens of comparative reasoning. This includes both same-patient longitudinal comparison and comparison with clinically analogous cases from other patients. It focuses on the capabilities required to move from access to comparative information toward clinically meaningful interpretation, covering historical-context selection, correspondence, temporal interpretation, reference-case comparison, communication, evaluation, safety, and the transition toward patient-level radiology AI.

Relevant literature was identified through targeted and topic-driven searches, supplemented by backward and forward citation tracking. As this is a narrative rather than a systematic review, studies were selected based on their relevance to the conceptual and clinical themes discussed rather than to provide exhaustive coverage.

Main points

- Comparative radiology artificial intelligence (AI) includes same-patient longitudinal reasoning and reference-case reasoning using analogous cases from other patients.
- Providing prior examinations does not by itself demonstrate longitudinal reasoning.
- Meaningful longitudinal reasoning requires appropriate historical-context selection, correspondence across examinations, interpretation of interval change, and integration across multiple time points.
- Clinically useful reference-case comparison should prioritize entity-specific relevance over global image similarity.
- Comparative radiology AI should be evaluated for appropriate use of comparative information as well as clinical utility.

Conceptual foundations of comparative reasoning

For this review, two related but clinically distinct forms of comparative reasoning in radiology AI are distinguished (Figure 1). The first—and the primary focus of this review—is longitudinal comparison, in which current and previous examinations from the same patient are interpreted together to characterize disease evolution; paired longitudinal vision–language approaches provide a direct example of this paradigm.[11] The second is reference-case comparison, in which a current examination is compared with clinically analogous cases from other patients to support diagnosis or differential diagnosis, as demonstrated by a deep learning-based image retrieval system evaluated for radiologist decision support.[12] Longitudinal and reference-case comparison have distinct clinical purposes: the former characterizes change within the same patient over time, whereas the latter supports interpretation by comparison with analogous cases from other patients. They nevertheless share a common principle of identifying clinically relevant entities or regions across images and reasoning about their correspondence, similarity, or difference. Cross-modality imaging may contribute to either form, although it is regarded as a technical dimension of comparison rather than a separate category of reasoning. Comparison with a normative reference population, as in age- and sex-adjusted brain volumetric percentiles or coronary artery calcium percentiles, is also common in clinical radiology; its comparator, however, is a statistical distribution rather than another imaging examination, and it is not considered further here.

Furthermore, a capability hierarchy is proposed that distinguishes access to longitudinal information from increasingly advanced forms of longitudinal reasoning (Table 1 and Figure 2). The individual components of this hierarchy are not necessarily novel, as several have been studied separately. Its contribution is to place these capabilities within a common structure that distinguishes what each level demonstrates and what remains to be evaluated. At the most basic level, an AI system may analyze a single examination without historical context. A system at the next level may receive a current and prior examination but use them primarily as joint inputs for prediction, as demonstrated by models combining current and prior computed tomography (CT) for pulmonary nodule malignancy estimation.[13] Beyond simple access, a system may identify the historical examination, series, report, or reference point most relevant to the current task.[4] More advanced systems establish anatomical or lesion-level correspondence between examinations. Systems at the next level support quantification of interval differences, as illustrated by longitudinal CT pipelines combining detection, segmentation, lesion matching, and change classification.[14] Higher-level systems translate paired-examination information into clinically meaningful temporal states; for example, progression models have classified findings as improved, unchanged, worsened, or new.[15] At the highest level, multi-interval trajectory reasoning requires integration of several examinations across time; temporal models have begun to operationalize this by encoding serial examination trajectories for prediction.[16]

The clinical relevance of this hierarchy can be seen in treatment-response assessment. For example, under Response Evaluation Criteria in Solid Tumors (RECIST) 1.1, a patient may have a sum of target-lesion diameters of 100 mm at baseline, 60 mm at nadir, and 75 mm at a subsequent follow-up. Because progression is assessed relative to the smallest sum on study, the increase from 60 to 75 mm represents a 25% and 15-mm increase and therefore meets the criteria for progressive disease, despite the tumor burden remaining below baseline. This assessment depends on correct identification of the corresponding target lesions across examinations; if lesion correspondence is uncertain or incorrect, the measured change and the resulting temporal interpretation should likewise be regarded as uncertain.

These levels are not interchangeable and should not be interpreted as a mandatory sequence. The ability to accept multiple images does not, by itself, demonstrate that

Table 1. Taxonomy of capabilities relevant to comparative reasoning in radiology artificial intelligence

Category	Capability	Definition	Required information or data	Typical output	What it does not establish
Longitudinal	**Single-examination analysis**	Interpretation of one examination without explicit use of previous imaging or reference cases	Current examination	Detection, segmentation, classification, or diagnosis	Use of historical context or comparative reasoning
	Longitudinal information access	Availability of one or more previous examinations or reports to the model	Current examination + previous examination(s) and/or report(s)	Prediction incorporating multiple inputs	That the model has identified or interpreted temporal relationships
	Historical-context selection	Identification of the previous examination, series, report, reference point, or clinical event most relevant to the current task	Current examination + longitudinal patient history	Selected prior examination, series, report, reference point, or event	Correct correspondence or interpretation of interval change
	Anatomical correspondence	Establishment of spatial or semantic correspondence between anatomical structures across examinations	Current and previous examinations	Registered anatomy or corresponding regions	Identity of individual lesions or clinical meaning of change
	Lesion/entity correspondence	Identification of the same lesion, finding, or clinically relevant entity across examinations	Current and previous examinations, often with spatial and appearance information	Matched lesion or entity across time	Whether the matched entity has clinically progressed or responded
	Change characterization	Detection or quantification of differences in a corresponding entity between examinations	Corresponding anatomy or entity across ≥ 2 examinations	Change in size, volume, burden, morphology, signal, attenuation, enhancement, or another imaging feature	Whether the observed difference represents biological progression, response, or technical variability
	Temporal interpretation	Translation of interval differences into clinically meaningful temporal states	Correspondence + interval information	New, stable, improved, worsened, or resolved	Integration of the complete disease course across several examinations
	Multi-interval trajectory reasoning	Integration of information across several time points while preserving temporal order, entity identity, and consistency	Structured multi-visit imaging history	Coherent interpretation or representation of disease course	Clinical utility unless benefit is independently demonstrated
Reference	**Reference-case comparison**	Comparison of the current examination or finding with clinically analogous cases from other patients	Current examination + reference-case collection	Similar cases, clinically relevant similarity/difference assessment, and diagnostic support	Same-patient temporal reasoning

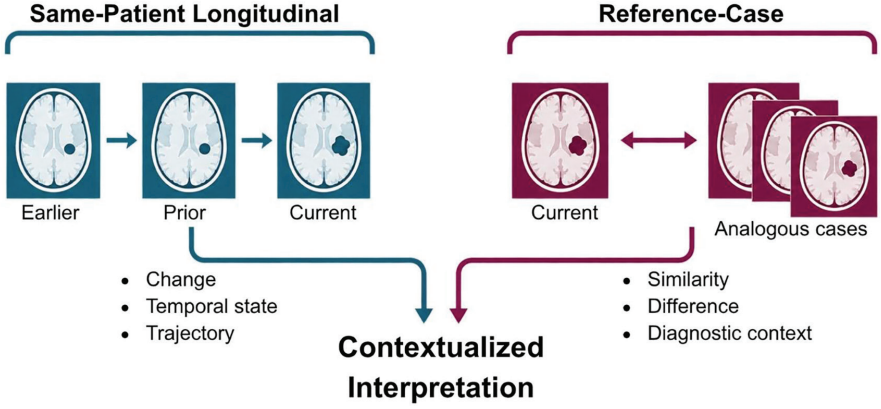


Figure 1. Comparative reasoning in radiology artificial intelligence. A current finding can be interpreted using two complementary forms of comparison: the patient's own prior examinations to understand temporal evolution (i.e., longitudinal comparison), or clinically analogous cases from other patients to provide diagnostic context (i.e., reference-case comparison). Cross-modality imaging may contribute to either form.

a model understands temporal relationships.[8,17] Having access to a patient's imaging history is distinct from selecting appropriately within it, because the relevant comparator is determined by the clinical task rather than by recency. Registration or lesion matching establishes spatial or lesion correspondence across examinations but does not itself determine the clinical meaning of change.[18,19] Likewise, quantifying an interval difference and interpreting it clinically are different capabilities.[20,21] Pairwise comparison between two examinations is not equivalent to multi-interval longitudinal reasoning, which requires integrating temporally ordered evidence into a globally consistent interpretation across multiple visits (Figure 3).[8]

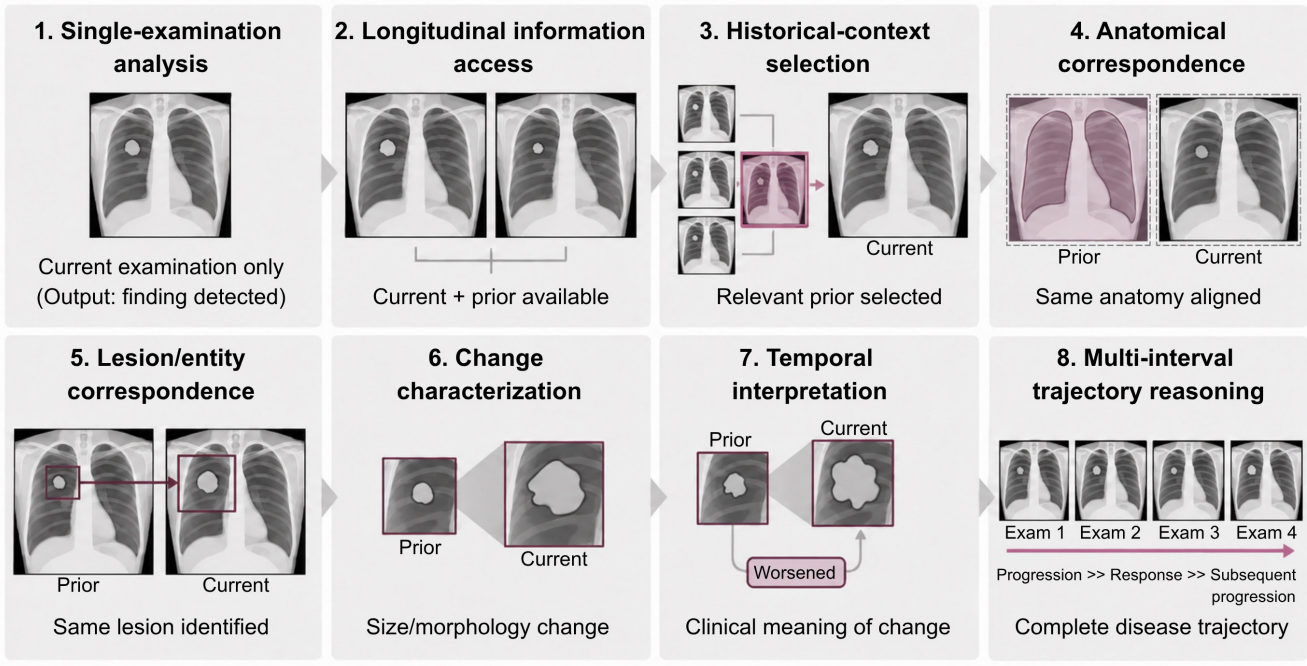


Figure 2. Capabilities in longitudinal radiology artificial intelligence. The schematic distinguishes longitudinal information access from progressively more advanced capabilities, including historical-context selection, anatomical and lesion-level correspondence, change characterization, temporal interpretation, and multi-interval trajectory reasoning. These represent operational capabilities rather than a mandatory sequential architecture, and access to prior examinations alone does not establish longitudinal reasoning.

Recent multi-interval benchmarking supports this last distinction, showing that locally plausible interval-level interpretations may not translate into globally consistent trajectory-level reasoning.[8] This argues for defining longitudinal capability in terms of the temporal relationships and dependencies a model can establish and interpret across examinations, rather than the number of examinations it receives.

Construction of longitudinal context

Longitudinal reasoning begins before direct image comparison. A patient may have numerous previous studies, and the clinically relevant comparator is not necessarily the most recent examination (Figure 4).[4] In some settings, the immediately preceding study is most informative; in others, the relevant reference may be a pretreatment baseline, an earlier examination, or several studies that establish a longer trajectory.[4,9,20] In oncologic response assessment, for example, interpretation may depend on comparison with pretreatment baseline when assessing response, or the smallest previously recorded tumor burden (nadir) when assessing progression, rather than simply the immediately preceding examination.[20] The time interval itself also carries clinical meaning, because expected rates of lesion growth and the evolution of treatment-related changes vary substantially between clinical settings.[22,23]

Consequently, selection of historical information should be considered part of the longitudinal reasoning problem rather than a preprocessing step, because identifying the appropriate comparator depends on the clinical question being asked. Providing a model with every available examination indiscriminately increases input volume without necessarily increasing relevant comparative information and adds computational burden. Human radiologists already perform selective historical retrieval by identifying

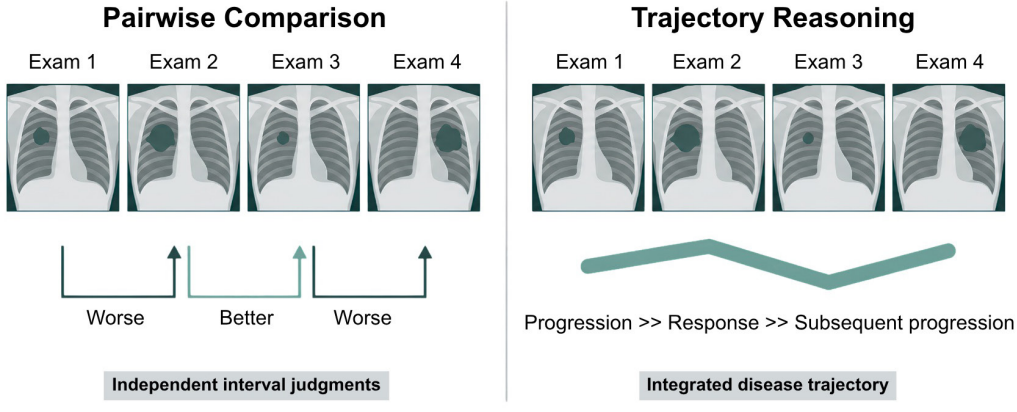


Figure 3. Pairwise comparison versus multi-interval trajectory reasoning. Pairwise approaches generate separate judgments for individual intervals, whereas trajectory reasoning integrates temporally ordered information across the complete sequence to characterize the overall disease course (with subsequent progression illustrated by a new contralateral lesion at examination 4). Correct interval-level judgments, therefore, do not, by themselves, demonstrate coherent multi-interval reasoning.

Table 2. Proposed evaluation framework for comparative radiology artificial intelligence

Evaluation target	Minimum evidence	Stronger evidence	What must be ruled out
Historical-context selection	Accuracy or relevance of the selected previous examination, series, report, or reference point	Reader-defined clinical relevance; downstream effect on interpretation; comparison with simple strategies, such as selecting the most recent prior	Selection of the temporally closest study rather than the clinically appropriate comparator
Anatomical correspondence	Registration or localization accuracy using an appropriate reference standard	Evaluation across scanners, protocols, anatomical deformation, and treatment states	Clinically important local misregistration despite good global alignment
Lesion/entity correspondence	Accuracy of matching the same lesion or entity across examinations	Evaluation in patients with multiple similar lesions, new lesions, resolving lesions, and marked morphological change	Spatial proximity mistaken for true entity identity
Change characterization	Agreement with reference measurements or labels for interval change	Repeatability analysis and robustness to acquisition or reconstruction differences	Technical or measurement variability interpreted as biological change
Temporal interpretation	Accuracy for clinically meaningful states, such as new, stable, improved, worsened, or resolved	Performance stratified by temporal state; comparison with current-examination-only models; correct-prior versus irrelevant-prior experiments	Model infers temporal state from the current examination alone or from dataset-specific shortcuts
Incremental value of longitudinal information	Better performance in the same cases with current + prior input than with the current examination alone	Correct-prior versus irrelevant/mismatched-prior testing; prior-image versus prior-report ablation; perturbation of historical information	Improvement caused by additional input volume, patient identity, acquisition characteristics, or textual patterns rather than meaningful temporal reasoning
Value of longer history	Comparison of one-prior versus multiple-prior input	Incremental performance according to number, timing, and relevance of previous examinations	Additional examinations increase information volume without providing useful trajectory information
Multi-interval trajectory reasoning	Evaluation using ≥ 3 temporally ordered examinations and trajectory-level targets	Comparison of full-history reasoning with independent adjacent-pair predictions; explicit temporal-consistency evaluation	Independent pairwise predictions presented as evidence of trajectory reasoning
Comparative report generation	Correctness of comparative findings and temporal states	Entity-level factuality, temporal consistency, grounding to relevant examinations, and uncertainty assessment	Fluent temporal language without correct comparative evidence
Reference-case comparison	Diagnostic relevance of retrieved analogous cases	Reader study; lesion/entity-specific retrieval analysis; comparison with global-image retrieval	Similar anatomy or acquisition characteristics mistaken for clinically meaningful similarity
Clinical utility	Effect on radiologist performance, workflow, or clinical decision-making in a clinically relevant setting	Multi-reader, workflow, or prospective studies; evaluation of effects on diagnosis, management, or other clinically meaningful outcomes	Benchmark or statistical improvement assumed to translate into clinical benefit

which prior examinations, image series, findings, reports, and clinical events are most relevant to the current question.[4,24]

The optimal longitudinal context may include both images and text. Previous reports provide compact summaries of established findings and earlier interpretations, whereas prior images provide the underlying visual evidence. Longitudinal AI systems have begun to exploit both sources simultaneously; for example, models have combined the previous chest radiograph and its report with the current examination to pre-fill the findings section of the current report.[25] Real-world comparison is further complicated by heterogeneity between current and prior examinations. In a mammography reader study, diagnostic performance differed among cases with prior and current examinations acquired using the same versus different vendors.[26] Surgery, radiation therapy, and systemic treatment between examinations can also produce expected posttreatment changes that affect how interval findings are interpreted.[23,27]

Automated selection of relevant historical information is beginning to be addressed explicitly.[4] Weikert et al.[4] developed TimeLens, which used natural language processing and image matching to identify relevant prior examinations, comparator series, and corresponding finding locations in serial CT and magnetic resonance imaging (MRI); use of the system reduced the average time required for longitudinal finding assessment by approximately 40%. In the literature reviewed here, however, prior-selection methods remain less prominent than downstream approaches for registration, lesion matching, and change analysis. This suggests an important research opportunity: future longitudinal systems should not simply accumulate patient history but should identify which elements of that history are relevant to the current diagnostic task. Once the appropriate comparison has been selected, the next problem is establishing reliable correspondence across examinations.

Correspondence across examinations

Meaningful longitudinal comparison requires establishing that the same anatomy, lesion, or clinical entity is being evaluated across time.[5,18,28] This can be challenging because serial examinations may differ in patient positioning, acquisition parameters, scanner characteristics, and respiratory state;

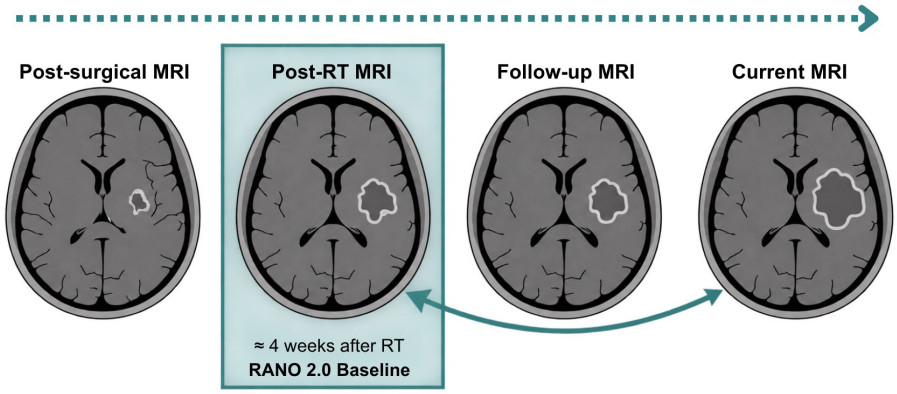


Figure 4. Selection of clinically relevant longitudinal context. The temporally nearest prior examination is not necessarily the clinically defined reference examination for a longitudinal task. The Response Assessment in Neuro-Oncology (RANO) Version 2.0 criteria provide a practical example: in newly diagnosed glioma treated with radiotherapy (RT), the post-RT magnetic resonance imaging (MRI) obtained approximately 4 weeks after treatment serves as the baseline for subsequent response assessment rather than the immediate postsurgical MRI. For subsequent progression assessment, the relevant comparison may shift to the nadir if tumor burden decreases below baseline. The schematic illustrates the broader principle that longitudinal artificial intelligence should identify task-appropriate reference examinations rather than defaulting to the most recent prior, while still integrating information across the longitudinal imaging history.

anatomy may also be altered due to treatment-related changes.[18,28,29]

At the anatomical level, image registration provides a common spatial framework for comparison.[18] Rigid registration primarily addresses translation and rotation, affine registration additionally accommodates transformations such as scaling and shearing, and deformable registration models more complex local anatomical changes.[18] For longitudinal reasoning, however, accurate alignment is not sufficient if the registration process distorts the pathology being measured.[5] This is particularly relevant in treatment-response imaging, where genuine changes in tumor size or morphology should be preserved rather than normalized by the registration process.[5] A recent longitudinal breast MRI study specifically addressed this problem by developing deformable registration that maintained global anatomical alignment while preserving tumor volume.[5]

Anatomical alignment also does not guarantee lesion identity.[6,28,30] Individual lesions may enlarge, shrink, resolve, or substantially change in appearance between examinations, and new lesions may emerge and others disappear.[18,28] Longitudinal lesion tracking therefore requires more than anatomical alignment alone.[6,28,30] Cai et al.[6] demonstrated this principle with Deep Lesion Tracker, which combined lesion appearance with anatomical information to identify corresponding lesions across serial CT examinations. This distinction between anatomical alignment and lesion-specific matching is particularly important in patients with multiple similar lesions, where spatial correspondence alone may be insufficient.[6,28,30]

Correspondence becomes more challenging when exact voxel-wise correspondence is unreliable, such as with substantial treatment-related deformation or when correspondence must be established across modalities.[5,31] Recent work has therefore explored semantic and anatomical embeddings that identify corresponding structures despite substantial differences in visual appearance.[31] For example, the Universal Anatomical Embedding framework was evaluated for both longitudinal CT lesion tracking and cross-modality CT–MRI correspondence for affine and rigid registration, explicitly addressing situations in which structures with similar meaning may have markedly different appearances.[31] In such settings, the relevant longitudinal question may be entity-level correspondence rather than exact voxel overlap: whether a finding or anatomical structure at one time point represents the same clinical entity at another.

This distinction is important for the overall capability hierarchy. Correspondence determines what should be compared; it does not determine what the observed difference means clinically. Once reliable anatomical or lesion identity has been established, the task then shifts from alignment and matching toward temporal interpretation.

Temporal reasoning across serial examinations

Temporal reasoning begins with detection of interval change but extends beyond it. At a basic level, models can identify differences in location, size, morphology, signal intensity, attenuation, enhancement, or disease burden between sequential examinations. These differences can then be translated into clinically meaningful temporal states. For example, an anatomy-specific chest radiograph model has classified interval findings as improved, unchanged, worsened, or new while also localizing newly appearing abnormalities.[15] Such categories move longitudinal analysis closer to clinical radiology reporting and decision-making than raw pixel- or feature-level differences alone.

Quantitative longitudinal analysis provides another level of information. Depending on the clinical task, relevant measurements may include lesion diameter, volume, number of lesions, disease burden, growth rate, enhancement, diffusion, perfusion, or other quantitative imaging biomarkers. Longitudinal models can use these changes for higher-level clinical tasks, such as treatment-response prediction. In a multi-institution study of rectal cancer, Jin et al.[32] integrated pre- and post-treatment multiparametric MRI using a Siamese deep learning framework to predict pathologic complete response; the longitudinal model significantly outperformed models using imaging from a single time point. This illustrates the transition from measuring imaging change toward estimating clinically relevant disease evolution, without implying that model predictions alone are sufficient to determine management.

A major challenge is distinguishing biological change from measurement and acquisition variability. Quantitative imaging studies have shown that apparent longitudinal differences can arise from imaging parameters, segmentation, measurement procedures, and other sources of variability even when no true biological change has occurred.[33,34] Treatment adds a separate layer of complexity because therapy-related imaging changes may mimic progression. In neuro-oncology, for example, the Response Assessment in Neuro-Oncology (RANO) Version 2.0 set of criteria specifically addresses the high frequency of pseudoprogression after radiotherapy and recommends confirmation of progression in appropriate settings.[35] Longitudinal AI therefore needs to interpret observed change in relation to both imaging conditions and treatment chronology.

A major limitation is that many existing evaluations of AI in medical imaging have focused on single images or short-horizon image pairs rather than complete multi-visit trajectories.[8,36] Pairwise comparison is clinically useful, but it is not equivalent to reasoning over a disease course. Across several examinations, serial imaging may reveal progressive enlargement, response followed by recurrence, or different trajectories among individual lesions. These patterns require preservation of temporal order and integration of evidence across multiple intervals rather than independent before-and-after comparisons. Longitudinal modeling studies also suggest that additional time points may provide incremental predictive information.[37] In patients with brain metastases treated with stereotactic radiosurgery, a convolutional gated recurrent unit model integrating four sequential MRI examinations outperformed simple convolutional neural network, radiomics, and diameter-based comparison models, with performance improving as additional follow-up examinations were incorporated.[37]

Multi-interval benchmarks have begun to expose this limitation directly. The Multi-Interval Chest X-Ray (MI-CXR) benchmark evaluates five-visit chest radiograph sequences and separates temporal event localization, interval-wise change reasoning, and global trajectory summarization.[8] Across 14 vision–language models, mean accuracy was 29.3%, only modestly above random guessing. Probing analyses showed that models could generate locally plausible interval interpretations yet fail to combine them into globally consistent decisions across the complete timeline. Thus, good performance on paired examinations should not automatically be interpreted as evidence of longitudinal trajectory reasoning.

Temporal interpretation also needs to account for regional and entity-level heterogeneity. Disease evolution may differ across anatomical regions: one abnormality may improve while another remains unchanged or worsens. Anatomy-specific progression modeling has shown that incorporating regional information can improve progression classification compared with approaches relying only on global image features.[15] Stability itself is also a clinically relevant longitudinal state; dedicated follow-up chest radiograph systems have been developed specifically to identify unchanged examination pairs while still detecting important interval abnormalities.[36] More recently, the CheXTemporal benchmark reported that, in zero-shot evaluation, the vision–language models tested performed better on visually salient worsening than on subtler states, such as stable and resolved, although these findings currently derive from a preprint and require further validation.[38]

The central objective of longitudinal AI should therefore extend beyond detecting that two images differ. The system should determine which entities have changed, in what direction, over what interval, and with what clinical significance. Comparative information becomes valuable when it contributes to a coherent and clinically meaningful interpretation of disease course.

Reference-case and analogous-case reasoning

Reference-case comparison represents the second form of comparative reasoning considered in this review. Instead of comparing a patient with their own prior examinations, the current examination is compared with previously diagnosed cases or other clinically analogous cases from other patients.[39] This approach is particularly relevant when imaging findings are uncommon, visually confusable, or associated with a broad differential diagnosis.[12,40]

Content-based image retrieval (CBIR) provided the historical foundation for this approach by retrieving radiological images according to visual similarity, with semantic and metadata information used to improve clinical relevance.[41] More recently, deep representations and foundation-model embeddings have been applied to retrieval across large and heterogeneous imaging collections.[42] In a 2025 study evaluating 1.6 million radiological images across four modalities and 161 pathologies, selected weakly supervised foundation models, particularly BiomedCLIP, achieved competitive retrieval performance without task-specific fine-tuning; however, retrieval based on pathological features remained less effective than retrieval based on anatomical characteristics.[42] This distinction is important because visual similarity does not necessarily imply clinical

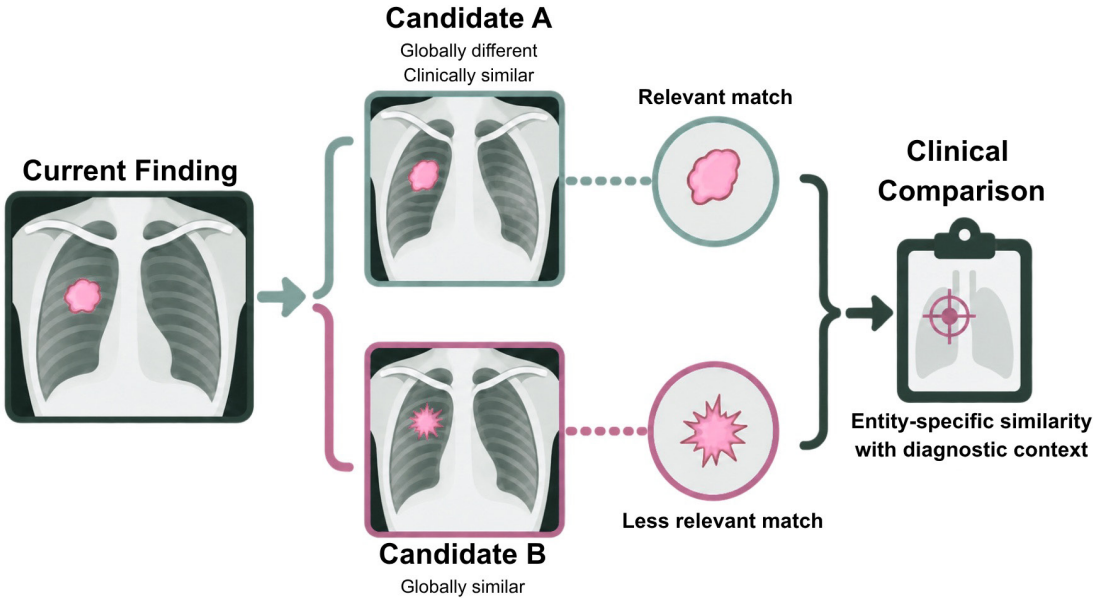


Figure 5. Reference-case reasoning and the distinction between global image similarity and clinical relevance. Candidate A is globally different from the current examination but contains a clinically analogous lesion, whereas Candidate B is globally similar but contains a clinically different finding. Clinically meaningful reference-case reasoning should therefore prioritize similarity of the relevant entity rather than overall image appearance.

similarity (Figure 5).[43] Clinically useful retrieval may therefore require representations that emphasize the specific lesion, anatomical region, or diagnostic entity rather than global image appearance alone.[39,43] Successful retrieval should therefore not be interpreted as evidence of reference-case reasoning unless the retrieved cases are also used appropriately to support diagnostic interpretation.

Reader studies provide direct evidence that analogous-case retrieval can support radiological interpretation.[12,40] In interstitial lung disease CT, use of a deep learning-based CBIR system that retrieved similar cases using regional disease patterns improved both diagnostic accuracy and interreader agreement across readers with different levels of experience.[12] More recently, a multi-reader study of orbital MRI demonstrated improved diagnostic accuracy with retrieved reference cases for uncommon eye and orbital masses, and reading time was lower in reads using the CBIR tool alone than in reads using status quo reference tools alone.[40] Together, these reader studies suggest that reference-case retrieval can improve radiologist diagnostic performance in selected settings, particularly when interpretation involves uncommon or visually overlapping imaging patterns.[12,40]

The value of reference-case reasoning also depends on the quality and composition of the reference database. In CBIR, the size and diversity of the database can influence the reliability of retrieved comparisons.[44] Cases should also have reliable diagnostic labels or reference standards, because weak or noisy labels can compromise the reliability and generalizability of AI systems.[45] Biased or unrepresentative case sampling may limit the applicability of retrieved cases to other patient populations.[45,46]

The conceptual link between longitudinal and reference-case comparison has recently been formalized in MedReCo, which frames radiological comparison as an entity-aware cross-image reasoning problem encompassing both clinically analogous reference-case retrieval and same-patient temporal interpretation.[39] This emerging formulation is conceptually useful; however, for the purposes of this review, analogous-case reasoning should remain distinct from text-centric retrieval-augmented generation or literature search, in which retrieval primarily supplies external knowledge to a generative model rather than a matched imaging case used directly as comparative imaging evidence.[47]

Communication of comparative reasoning

Comparative reasoning becomes clinically useful when its conclusions can be communicated accurately. Longitudinal report generation therefore represents a downstream expression of temporal reasoning rather than the entirety of the reasoning process.

A comparative report should identify which findings have changed, clearly identify corresponding anatomical locations, and use appropriate temporal language, such as stable, improved, worsened, new, or resolved. Such statements should be grounded in imaging evidence from the relevant examinations and should communicate uncertainty when the observed difference is equivocal.[1,48] The European Society of Radiology guideline explicitly recommends clear anatomical localization and description of comparison with prior examinations,[1] and Bruno et al.[48] specifically address clear communication of diagnostic uncertainty.

An important limitation is that models can learn the language of comparison without access to reliable comparative evidence. Report-generation systems trained on routine radiology reports have been shown to generate references to prior examinations that were not actually available to the model, reflecting patterns inherited from reports containing historical comparisons rather than genuine longitudinal interpretation.[49] Even when previous images and reports are provided, correctly translating changes between prior and current observations into text remains difficult. In a longitudinal chest radiograph report-generation study, when the label for the same observation differed between the previous and current reports, 84.42% of generated outputs were incorrect for that observation.[25]

Recent longitudinal report-generation approaches have attempted to incorporate temporal imaging information more explicitly. Dalla Serra et al.[50] aligned and fused current and prior chest radiographs into a joint longitudinal representation and introduced anatomy-based control over generated report content. More recently, Libra incorporated paired current and prior images through a dedicated temporal alignment mechanism within a vision–language model, further illustrating the move toward explicit modeling of interval information during report generation.[51] Other recent approaches have moved toward explicit difference modeling; DDaTR introduced a dynamic difference-aware temporal architecture designed to capture changes between current and prior examinations during longitudinal report generation.[52] These approaches demonstrate technical progress, but the accuracy of comparative language should remain an independently evaluated output of longitudinal reasoning.[51] A fluent statement that a lesion is stable does not, by itself, establish that the underlying temporal interpretation was correct.

Evaluation of comparative radiology artificial intelligence

Evaluation should establish not only whether a comparative radiology AI system performs well but also whether its performance depends appropriately on the comparative information it receives. An improvement after adding a prior examination does not, by itself, demonstrate that the model used that prior correctly.[53] A system may instead rely on entity presence or on acquisition-, patient-, or healthcare-process correlates rather than the intended temporal information.[53-55]

Evaluation should therefore reflect the stages of the comparative-reasoning pipeline (Table 2). Historical-context selection requires assessment of whether the clinically appropriate historical examination and comparison series were identified or retrieved.[4] Anatomical and lesion/entity correspondence require evaluation of spatial alignment and correct identification of the same lesion or finding across examinations.[6] Change characterization requires accurate localization and, where applicable, measurement of interval differences, whereas temporal-state interpretations should distinguish task-appropriate clinically meaningful states, such as new, stable, improved, worsened, and resolved.[6,7,53] Multi-interval trajectory reasoning should additionally demonstrate consistency across the complete timeline rather than only between adjacent examinations.[8] Reference-case reasoning should similarly be evaluated for the clinical relevance of retrieved cases, including whether similarity reflects the relevant lesion or entity rather than global image resemblance. More generally, robust evaluation of medical-imaging AI requires complementary measures selected according to the task, dataset properties, and clinical context.[56]

Controlled comparisons are particularly informative. Current-examination-only performance can be compared in the same cases with performance using the correct prior,

to estimate incremental longitudinal value.[10] Correct priors can be contrasted with irrelevant or mismatched priors to test whether the model responds specifically to meaningful temporal information (Figure 6). Where sufficient data exist, one-prior and multiple-prior settings can determine whether longer histories provide added value. Comparisons between prior images and prior reports may help establish whether observed gains arise from visual or textual historical information.[7]

Perturbation-based evaluation, in which supplied comparative information is altered rather than withheld, may provide further evidence of genuine temporal sensitivity. Ko et al.[53] evaluated longitudinal chest radiograph models by reversing current–prior image order and measuring order sensitivity and temporal consistency, demonstrating how controlled manipulation of comparative information can test whether model behavior reflects the intended temporal relationship.

These strategies should be regarded as proposed evaluation principles rather than universal requirements, since the appropriate design depends on the intended task. Nevertheless, claims of longitudinal reasoning are stronger when evaluation demonstrates that model behavior changes appropriately with the relevance, content, or temporal ordering of the comparison.[53]

Dedicated temporal benchmarks are beginning to provide more task-specific evaluation. Microsoft Chest X-Ray Temporal (MS-CXR-T) introduces radiologist-reviewed temporal labels for paired chest radiographs, including improving, stable, and worsening states, together with a temporal-semantic text task.[7] The MI-CXR benchmark approach extends evaluation beyond image pairs to five-visit timelines and separately assesses event localization, interval-wise change reasoning, and global trajectory summarization.[8] The LUNGUAGE dataset further addresses sequential report interpretation through expert-reviewed longitudinal annotations and LUNGUAGESCORE, which evaluates entities, relations, attributes, and temporal consistency across patient timelines.[57] Metrics of this kind are important because conventional lexical-overlap measures, such as BLEU and ROUGE, do not directly establish whether a comparative statement is factually, clinically, or temporally correct.[57-59] Temporal-state vocabularies also differ across benchmarks and primary studies, which limits direct comparison of reported performance.[7,15] Comparable benchmarks for reference-case retrieval remain less developed, although large-scale retrieval evaluations have begun to assess whether similarity reflects pathological rather than anatomical features.[42]

Clinical utility requires a different level of evidence from benchmark or metric-based performance. Statistical performance alone does not necessarily imply practical clinical relevance, and medical-imaging AI evaluation should ultimately consider the clinical consequences of model performance.[56] A recent digital breast tomosynthesis study provides a useful example: the same AI system achieved an area under the receiver operating characteristic curve of 0.93 both with and without prior examinations.[10] Thus, the availability of clinically realistic historical information should not itself be interpreted as evidence of added clinical value.

Accordingly, evidence of comparative reasoning and evidence of clinical utility should be distinguished. The former can be demonstrated through controlled, task-specific experiments showing appropriate use of comparative information. The latter requires evidence of benefit when the system is integrated into clinical decision-making, and broader claims of generalizability additionally require testing beyond the development setting.[60,61] An established AI reporting guideline, the Checklist for Artificial Intelligence in Medical Imaging (CLAIM), distinguishes internal testing from external testing, and a clinical AI guideline, Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence (DECIDE-AI), emphasizes evaluation in live workflows, including human–AI interaction, safety, implementation, and actual clinical performance.[60,61] Comparative AI should therefore be evaluated separately for task performance, for whether it uses comparative information (e.g., historical information in longitudinal comparison) appropriately, and for whether that capability ultimately improves clinical practice.

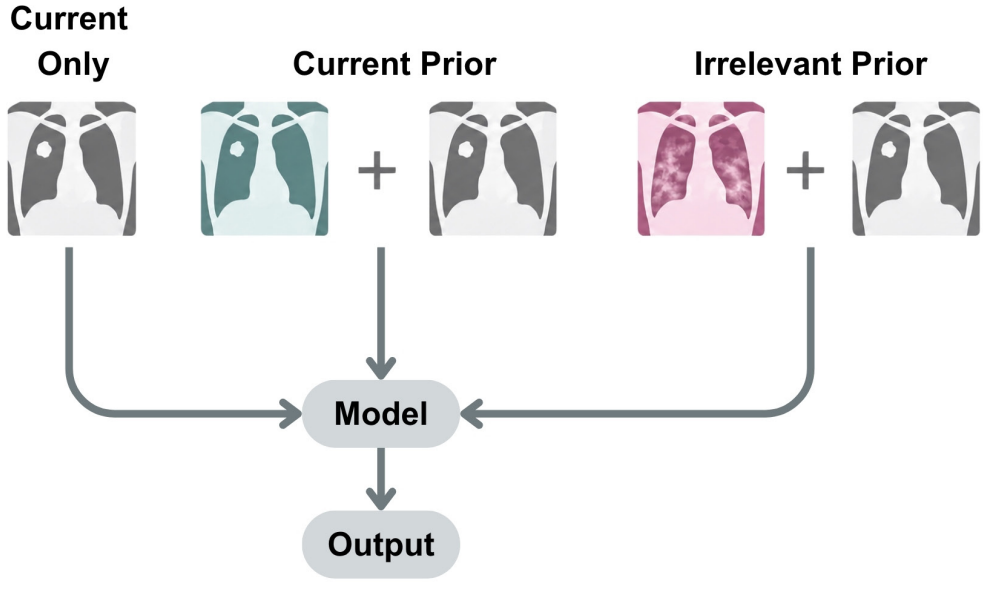


Figure 6. Testing whether longitudinal information is used specifically and appropriately. Providing a prior examination alone does not establish longitudinal reasoning. Stronger evidence is obtained when a clinically relevant prior produces an appropriate task-specific change or benefit compared with current-only input, and this effect is not reproduced by an irrelevant or mismatched prior. Evidence is ambiguous when relevant and irrelevant priors produce similar effects or when model output shows no specific dependence on clinically relevant history, because any apparent benefit may reflect additional input, shortcuts, or nonspecific contextual information rather than genuine temporal reasoning.

Table 3. Failure modes, consequences, and safeguards in comparative radiology artificial intelligence (AI)

Failure mode	Example	Potential consequence	Evaluation or mitigation
Incorrect prior examination or series retrieval	For comparison of an enhancing brain lesion, the system retrieves the precontrast T1-weighted series from the prior MRI instead of the available postcontrast series	Incorrect assessment of response or progression	Explicit evaluation of examination and series retrieval; verification that the characteristics of the selected images, such as sequence and contrast status, are appropriate for the comparison task
Missing historical information	Relevant imaging performed at another institution is unavailable	A pre-existing finding may appear new, or the observed trajectory may be incomplete	Make available history transparent; evaluate performance under incomplete histories
Inappropriate temporal reference point	Current tumor burden compared only with the previous examination rather than the clinically appropriate baseline or nadir	Incorrect temporal classification	Task-specific definition and recording of the appropriate comparison point
Registration failure	Substantial postoperative or treatment-related deformation produces incorrect alignment	False localization or apparent interval change	Registration quality control; pathology-aware methods; automated failure detection where feasible
Lesion/entity mismatch	One metastasis is matched to a different nearby metastasis at follow-up	Incorrect growth or response assessment	Entity-specific matching validation; confidence estimates; human review when correspondence is uncertain
Acquisition-related pseudo-change	Different slice thickness, reconstruction, respiratory state, scanner, or contrast timing alters appearance	Technical variability interpreted as biological change	Robustness testing across acquisition conditions; normalization where appropriate; uncertainty reporting
Treatment-related mimic of progression	Inflammation, pseudoprogression, necrosis, fibrosis, or postoperative change resembles active disease	Incorrect classification of progression or recurrence	Incorporation of treatment chronology and appropriate clinical context; task-specific reference standards
Patient-level data leakage	Different examinations from the same patient occur in both training and test sets	Inflated apparent performance	Patient-level dataset splitting; patient-independent external evaluation
Historical shortcut learning	Model exploits scanner characteristics, patient identity, or report style rather than temporal relationships	Apparent benefit from historical input without genuine comparative reasoning	Current-only controls; mismatched-prior experiments; shortcut and subgroup analyses
Hallucinated comparative language	Generated report refers to a previous examination that was not supplied	False impression that a comparison was actually performed	Explicit grounding of comparative statements to available evidence; factuality and temporal-consistency evaluation
Incorrect temporal language	Report states stable despite measurable progression	Misleading clinical communication	Entity-level assessment of comparative statements rather than lexical similarity alone
Propagation of historical error	Incorrect previous report or AI output is reused as context for a later interpretation	Earlier error may influence subsequent longitudinal conclusions	Preserve provenance; retain access to original imaging; explicitly study error propagation
Automation bias	Reader accepts an incorrect AI judgment of stability or progression	Reduced human diagnostic performance	Calibrated uncertainty; transparent supporting evidence; human-factors evaluation
Loss of entity-level information in patient representations	A global patient representation obscures divergent trajectories of individual lesions	Regional progression may be hidden by overall stability	Entity-level representations and lesion-specific temporal tracking
Failure under real-world heterogeneity	Model trained on standardized paired studies encounters different vendors, protocols, modalities, institutions, or treatment states	Reduced generalizability in routine follow-up	External multi-institution testing; subgroup analysis across protocol, vendor, modality, and institution
Clinically irrelevant reference-case retrieval	A globally similar examination is retrieved although the relevant lesion or entity is not clinically analogous	Misleading diagnostic context or differential diagnosis	Entity- or lesion-specific retrieval evaluation; comparison with global-image retrieval; reader assessment of clinical relevance

Failure modes, safety, and clinical translation

Longitudinal systems introduce multi-stage failure modes because an error at any stage propagates to those that follow (Table 3). A system may retrieve the wrong prior examination, use an inappropriate baseline, or lack access to a clinically relevant study performed elsewhere.[62] The availability and relevance of historical information should therefore be treated as part of the reliability of the longitudinal system rather than assumed to be constant. Reference-case systems introduce a related risk when globally similar examinations are retrieved despite poor similarity of the clinically relevant lesion or entity, potentially providing misleading diagnostic context.

Correspondence errors create another important risk. Registration may be challenged by substantial anatomical or treatment-related change, and lesion tracking becomes difficult when several similar lesions coexist or when lesions appear, disappear, or change markedly in morphology.[30] These limitations are not merely theoretical. In a longitudinal CT lesion-tracking study, an automated registration and segmentation pipeline succeeded for 96% of baseline lesions but only 80% of corresponding follow-up lesions.[30] If lesion identity is incorrect, apparently precise measurements of interval change can still produce an incorrect temporal conclusion.[30]

Longitudinal datasets also require particular attention to data leakage. Repeated examinations from the same patient can allow subject-specific information to appear in both training and evaluation sets when data are split at the examination rather than patient level. Park et al.[63] explicitly identify the placement of different follow-up imaging examinations from the same patient into training and test sets as data leakage, and Chaibub Neto et al. [64] demonstrate the related problem of identity confounding in repeated-measures machine-learning data. In a longitudinal brain MRI study, record-wise splitting produced a cross-validation accuracy of 97.33%, which fell to 38.71% when the trained models were evaluated on previously unseen subjects; subject-wise splitting was substantially less optimistic during cross-validation.[65] Although these numerical results arise from one experimental setting and should not be generalized across longitudinal imaging AI, they demonstrate why patient-level separation is essential when repeated examinations are present.

Communication creates an additional layer of risk. Report-generation systems trained on routine radiology reports can generate references to prior examinations that were not available to the model, creating apparently comparative statements without comparative evidence.[49] Historical reports may themselves contain interpretations that are later found to be discrepant or incorrect. A systematic review and meta-analysis of secondary imaging interpretations found an overall discrepancy rate of 32.2% between primary and secondary interpretations, although the authors also emphasized heterogeneity and limitations in determining which interpretation was correct in cases without a reference standard.[66] Previous information can also influence errors in longitudinal report generation: in one chest-radiograph study, generated observation labels were particularly likely to be incorrect when the corresponding finding changed between the previous and current examinations.[25] Systems that repeatedly consume earlier reports or AI-generated outputs could therefore propagate errors forward. Whether such longitudinal error inheritance occurs systematically has not yet been well characterized and should therefore be regarded as a potential risk rather than an established failure mode.

Human interaction adds another dimension. In a prospective mammography experiment, incorrect Breast Imaging Reporting and Data System suggestions presented as AI output substantially impaired reader performance among inexperienced, moderately experienced, and very experienced radiologists.[67] The system was simulated rather than an actually trained AI model. Although this study did not evaluate longitudinal AI specifically, it demonstrates the broader susceptibility of radiologists to automation bias. More recent eye-tracking evidence using a commercially available mammography AI system similarly showed reduced reader sensitivity in cases with false-negative AI suggestions, indicating that erroneous AI output can alter both diagnostic performance and visual search behavior.[68] Among 10 readers, median sensitivity fell from 71% without AI to 39% with false-negative AI suggestions, accompanied by a reduction in fixation rate. The same mechanism could be consequential when an AI system presents a confident temporal conclusion, such as stable or progressive. Automation bias may be more pronounced for such conclusions than for single-examination findings, because verifying a temporal claim requires retrieving and re-reading prior examinations rather than simply re-inspecting the current one.

Clinical translation should therefore include transparency about which examinations contributed to an output, how longitudinal findings were linked, and how certain the system is about its temporal conclusion. Clinical implementation should also address data privacy and security, particularly when reference cases from other patients or imaging histories across institutions are accessed, with appropriate de-identification and data-governance measures. These requirements align with the international FUTURE-AI consensus framework for trustworthy healthcare AI,[69] whose guiding principles include traceability, usability, robustness, and explainability, and whose recommendations include human oversight, uncertainty quantification, and logging of data access. More historical information is not inherently safer or more informative; its value depends on whether that information is relevant, reliable, correctly associated with the current findings, and interpreted within the appropriate clinical context.

Toward patient-level radiology artificial intelligence

The longer-term development of longitudinal AI requires moving beyond isolated current–prior pairs toward representations of a patient's evolving imaging history. Such systems should not simply store every previous examination: they should maintain clinically relevant information over time and retrieve the appropriate historical context for the current task. Recent patient-level pretraining work combining longitudinal chest CT with clinical variables illustrates the feasibility of learning representations across multiple time points and data types rather than treating each examination independently.[70]

Multi-interval reasoning is a central requirement. A patient-level model should preserve temporal order and entity identity, recognize changes in trajectory, and maintain consistency across several examinations. In pediatric glioma surveillance, Tak et al.[16] developed a temporal deep-learning approach using serial postoperative MRI examinations, including chronological-order learning during pretraining. Recurrence-prediction performance increased as additional historical examinations were incorporated, with gains plateauing after approximately three to six scans depending on the dataset. This is among the few direct estimates of how much imaging history contributes and suggests that the marginal value of additional examinations may decline rather than accumulate indefinitely.[9,16,37] Emerging work has begun to model imaging histories explicitly as spatiotemporal trajectories; for example, a recent breast MRI framework represented relationships across multiple treatment time points and evaluated how much the number of dynamic contrast-enhanced MRI time points and the inclusion of inter-scan time differences contributed to response prediction.[71] Such approaches illustrate the shift from paired-image comparison toward explicit trajectory representation.

Historical selection is another opportunity. Automated systems can already retrieve relevant prior examinations and image series and display corresponding finding locations after a finding is selected on the current examination.[4] Future systems could extend this capability through more adaptive retrieval

or agentic mechanisms that autonomously select and retrieve historical information according to the current clinical task.

Patient-level reasoning will also need to tolerate real-world heterogeneity across scanners, protocols, modalities, institutions, and treatment states. Recent foundation-model work provides evidence that large-scale clinical representation learning can produce features robust across some of these dimensions. NeuroVFM, trained on 5.24 million clinical CT and MRI volumes, learned a shared neuroimaging representation that supported cross-protocol anatomical matching and cross-modality diagnostic transfer, while maintaining MRI performance across manufacturers, field strengths, and medical centers.[72] Importantly, these capabilities should be viewed as infrastructure that may enable longitudinal reasoning, not as evidence that the model itself reasons over patient trajectories.

Foundation models and longitudinal pretraining should therefore be judged by demonstrated temporal behavior rather than by scale or generality alone. Increasingly capable image representations do not remove the need for explicit evaluation of longitudinal reasoning.

Future datasets should move beyond isolated pairs and include structured multi-visit trajectories with lesion or entity correspondence, annotated clinically appropriate comparators, clinically meaningful temporal labels, and realistic acquisition heterogeneity. Evaluation should likewise progress beyond retrospective benchmark accuracy toward external or multisite testing, and toward prospective clinical evaluation assessing effects on radiologist performance and productivity, workflow, clinical decision-making, and other clinically meaningful outcomes.[60,69]

Final remarks

Comparative reasoning is a necessary step beyond examination-centric radiology AI. Clinically meaningful systems must not only access prior examinations or reference cases but also identify the relevant comparison, preserve anatomical and lesion-level correspondence, and interpret temporal change or clinical similarity appropriately. Future progress will depend on rigorous evaluation of whether comparative information is used as intended and whether these capabilities translate into reliable clinical benefit. Ultimately, the goal is patient-level radiology AI that reasons coherently across evolving imaging histories and clinically relevant reference cases.

Building on this broader view of patient-level radiology AI, claims that AI could replace radiologists should be considered in the context of the broader interpretive capabilities required in clinical practice. Radiologists integrate findings across time, interpret them within the clinical and treatment context, and communicate uncertainty. Reliable comparative reasoning is therefore one of several capabilities that AI would need to demonstrate before such claims can be meaningfully assessed.

Acknowledgement

The language of this manuscript was improved using generative AI (ChatGPT 5.6, OpenAI). The authors also used ChatGPT 5.6 and Google Gemini 3.6 Flash to assist in creating selected figure components, which were subsequently edited in Canva. All AI-generated or AI-assisted content was carefully reviewed and supervised by the authors, who take full responsibility for the final content.

Footnotes

Conflict of interest disclosure

Ismail Mese serves as Section Editor and Burak Kocak previously served as Section Editor for *Diagnostic and Interventional Radiology*; neither had any involvement in the peer review of this article.

References

1. European Society of Radiology (ESR). Good practice for radiological reporting. Guidelines from the European Society of Radiology (ESR). *Insights Imaging.* 2011;2(2):93-96. [Crossref]
2. Acosta JN, Falcone GJ, Rajpurkar P. The need for medical artificial intelligence that incorporates prior images. *Radiology.* 2022;304(2):283-288. [Crossref]
3. Noebauer-Huhmann IM, Chaudhary SR, Papakonstantinou O, et al. Soft tissue sarcoma follow-up imaging: strategies to distinguish post-treatment changes from recurrence. *Semin Musculoskelet Radiol.* 2020;24(6):627-644. [Crossref] Erratum in: *Semin Musculoskelet Radiol.* 2020;24(6):e1. [Crossref]
4. Weikert T, Litt HI, Moore WH, et al. Reduction in radiologist interpretation time of serial CT and MR imaging findings with deep learning identification of relevant priors, series and finding locations. *Acad Radiol.* 2023;30(10):2269-2279. [Crossref]
5. Han L, Tan T, Zhang T, et al. Tumor-preserving deformable registration of longitudinal DCE breast MRI during neoadjuvant chemotherapy. *Radiol Artif Intell.* 2026;8(4):e250789. [Crossref]
6. Cai J, Tang Y, Yan K, et al. Deep lesion tracker: monitoring lesions in 4D longitudinal imaging studies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021:15159-15169 Accessed August 22, 2026. [Crossref]
7. Bannur S, Hyland S, Liu Q, et al. Learning To exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023:15016-15027 Accessed August 22, 2026. [Crossref]
8. Cho SS, Han Y, Do J. MI-CXR: a benchmark for longitudinal reasoning over multi-interval chest X-rays. In: Liakata M, Moreira VP, Zhang J, Jurgens D, editors. Findings of the Association for Computational Linguistics: ACL 2026. Stroudsburg (PA): Association for Computational Linguistics; 2026:30241-30273. [Crossref]
9. Hayward JH, Ray KM, Wisner DJ, et al. Improving screening mammography outcomes through comparison with multiple prior mammograms. *AJR Am J Roentgenol.* 2016;207(4):918-924. [Crossref]
10. Moshina N, Larsen M, Holen ÅS, et al. Artificial intelligence for digital breast tomosynthesis screening with and without prior examinations in BreastScreen Norway. *Radiol Artif Intell.* 2026;8(4):e250988. [Crossref]
11. Cho Y, Kim T, Shin H, Cho S, Shin D. Pretraining vision-language model for difference visual question answering in longitudinal chest X-rays. *arXiv [Preprint].* 2024:arXiv:2402.08966. In: Proceedings of the 7th International Conference on Medical Imaging with Deep Learning. PMLR; 2024:263-275. [Crossref]
12. Choe J, Hwang HJ, Seo JB, et al. Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest CT. *Radiology.* 2022;302(1):187-197. [Crossref]
13. Venkadesh KV, Aleef TA, Scholten ET, et al. Prior CT improves deep learning for malignancy risk estimation of screening-detected pulmonary nodules. *Radiology.* 2023;308(2):e223308. [Crossref]
14. Kenneth Portal N, Rochman S, Szeskin A, Lederman R, Sosna J, Joskowicz L. Metastatic lung lesion changes in follow-up chest CT: the advantage of deep learning simultaneous analysis of prior and current scans with SimU-Net. *J Thorac Imaging.* 2025;40(2):e0808. [Crossref]
15. Yu K, Ghosh S, Liu Z, Deible C, Poynton CB, Batmanghelich K. Anatomy-specific progression classification in chest radiographs via weakly supervised learning. *Radiol Artif Intell.* 2024;6(5):e230277. [Crossref]
16. Tak D, Garomsa BA, Zapaishchykova A, et al. Longitudinal risk prediction for pediatric glioma with temporal deep learning. *NEJM AI.* 2025;2(5):10.1056/aioa2400703. [Crossref]
17. Yang X, Miao J, Yuan Y, et al. Medical large vision language models with multi-image

visual ability. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference, Daejeon, South Korea, September 23-27, 2025, Proceedings, Part V. Berlin, Heidelberg: Springer; 2025:402-412. [Crossref]

18. Im JE, Khalifa M, Gregory AV, Erickson BJ, Kline TL. A systematic review on the use of registration-based change tracking methods in longitudinal radiological images. *J Imaging Inform Med.* 2025;38(4):2549-2562. [Crossref]
19. Santoro-Fernandes V, Huff D, Scarpelli ML, et al. Development and validation of a longitudinal soft-tissue metastatic lesion matching algorithm. *Phys Med Biol.* 2021;66(15):10.1088/1361-6560/ac1457. [Crossref]
20. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer.* 2009;45(2):228-247. [Crossref]
21. McKinley R, Wepfer R, Grunder L, et al. Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence. *Neuroimage Clin.* 2020;25:102104. [Crossref]
22. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. *Radiology.* 2017;284(1):228-243. [Crossref]
23. Benveniste MF, Gomez D, Carter BW, et al. Recognizing radiation therapy-related complications in the chest. *Radiographics.* 2019;39(2):344-366. [Crossref]
24. Haygood TM, Mullins B, Sun J, et al. Consultation and citation rates for prior imaging studies and documents in radiology. *J Med Imaging (Bellingham).* 2018;5(3):031409. [Crossref]
25. Zhu Q, Mathai TS, Mukherjee P, Peng Y, Summers RM, Lu Z. Utilizing longitudinal chest X-rays and reports to pre-fill radiology reports. *Med Image Comput Comput Assist Interv.* 2023;14224:189-198. [Crossref]
26. Trieu PDY, Borecky N, Li T, Brennan PC, Barron ML, Lewis SJ. The impact of prior mammograms on the diagnostic performance of radiologists in early breast cancer detection: a focus on breast density, lesion features and vendors using wholly digital screening cases. *Cancers (Basel).* 2023;15(4):1339. [Crossref]
27. Brahmbhatt S, Overfield CJ, Rhyner PA, Bhatt AA. Imaging of the posttreatment head and neck: expected findings and potential complications. *Radiol Imaging Cancer.* 2024;6(1):e230155. [Crossref]
28. Huff DT, Santoro-Fernandes V, Chen S, et al. Performance of an automated registration-based method for longitudinal lesion matching and comparison to inter-reader variability. *Phys Med Biol.* 2023;68(17):175031. [Crossref]
29. Jank E, Liu EY, Delery W, et al. Modern approaches for thoracic image registration and respiratory motion management in oncology. *Radiol Imaging Cancer.* 2025;7(5):e250023. [Crossref]
30. Hering A, Peisen F, Amaral T, et al. Whole-body soft-tissue lesion tracking and segmentation in longitudinal CT imaging studies. In: *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*. PMLR; 2021:312-326. Accessed August 22, 2026. [Crossref]
31. Bai X, Bai F, Huo X, et al. UAE: Universal Anatomical Embedding on multi-modality medical images. *Med Image Anal.* 2025;103:103562. [Crossref]
32. Jin C, Yu H, Ke J, et al. Predicting treatment response from longitudinal images using multi-task deep learning. *Nat Commun.* 2021;12(1):1851. [Crossref]
33. Obuchowski NA, Buckler A, Kinahan P, et al. Statistical issues in testing conformance with the quantitative imaging biomarker alliance (QIBA) profile claims. *Acad Radiol.* 2016;23(4):496-506. [Crossref]
34. Buckler AJ, Danagoulian J, Johnson K, et al. Inter-method performance study of tumor volumetry assessment on computed tomography test-retest data. *Acad Radiol.* 2015;22(11):1393-1408. [Crossref]
35. Wen PY, van den Bent M, Youssef G, et al. RANO 2.0: update to the response assessment in neuro-oncology criteria for high- and low-grade gliomas in adults. *J Clin Oncol.* 2023;41(33):5187-5199. [Crossref]
36. Yun J, Ahn Y, Cho K, et al. Deep learning for automated triaging of stable chest radiographs in a follow-up setting. *Radiology.* 2023;309(1):e230606. [Crossref]
37. Cho SJ, Cho W, Choi D, et al. Prediction of treatment response after stereotactic radiosurgery of brain metastasis using deep learning and radiomics on longitudinal MRI data. *Sci Rep.* 2024;14(1):11085. [Crossref]
38. Prakash E, Gao Y, Wang C, et al. CheXTemporal: a dataset for temporally-grounded reasoning in chest radiography. *arXiv [Preprint].* 2026:arXiv:2605.11304. [Crossref]
39. Zhang T, Zhao Z, Zhang X, et al. A vision-language framework for comparative reasoning in radiology. *arXiv [Preprint].* 2026:arXiv:2606.06407. [Crossref]
40. Rumberger JL, Lim W, Wildfeuer B, et al. Content-based image retrieval assists radiologists in diagnosing eye and orbital mass lesions in MRI. *Sci Rep.* 2025;15(1):11334. [Crossref]
41. Akgül CB, Rubin DL, Napel S, Beaulieu CF, Greenspan H, Acar B. Content-based image retrieval in radiology: current status and future directions. *J Digit Imaging.* 2011;24(2):208-222. [Crossref]
42. Denner S, Zimmerer D, Bounias D, et al. Leveraging foundation models for content-based image retrieval in radiology. *Comput Biol Med.* 2025;196(Pt A):110640. [Crossref]
43. Silva W, Gonçalves T, Härmä K, et al. Computer-aided diagnosis through medical image retrieval in radiology. *Sci Rep.* 2022;12(1):20732. [Crossref] Erratum in: *Sci Rep.* 2023;13(1):1470. [Crossref]
44. Wang XH, Park SC, Zheng B. Assessment of performance and reliability of computer-aided detection scheme using content-based image retrieval approach and limited reference database. *J Digit Imaging.* 2011;24(2):352-359. [Crossref]
45. Hadjiiski L, Cha K, Chan HP, et al. AAPM task group report 273: recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. *Med Phys.* 2023;50(2):e1-e24. [Crossref]
46. 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]
47. Fink A, Rau A, Reisert M, Bamberg F, Russe MF. Retrieval-augmented generation with large language models in radiology: from theory to practice. *Radiol Artif Intell.* 2025;7(4):e240790. [Crossref]
48. Bruno MA, Petscavage-Thomas J, Abujudeh HH. Communicating uncertainty in the radiology report. AJR *Am J Roentgenol.* 2017;209(5):1006-1008. [Crossref]
49. Ramesh V, Chi NA, Rajpurkar P. Improving radiology report generation systems by removing hallucinated references to non-existent priors. *arXiv [Preprint].* 2022:arXiv:2210.06340. [Crossref]
50. Dalla Serra F, Wang C, Deligianni F, Dalton J, O'Neil AQ. Controllable chest X-ray report generation from longitudinal representations. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics; 2023:4891-4904. [Crossref]
51. Zhang X, Meng Z, Lever J, Ho ESL. Libra: leveraging temporal images for biomedical radiology analysis. In: Che W, Nabende J, Shutova E, Pilehvar MT, editors. Findings of the Association for Computational Linguistics: ACL 2025. Stroudsburg (PA): Association for Computational Linguistics; 2025:17275-17303. [Crossref]
52. Song S, Tang H, Yang H, Li X. DDaTR: dynamic difference-aware temporal residual network for longitudinal radiology report generation. *IEEE Trans Med Imaging.* 2025;44(12):5345-5357. [Crossref]
53. Ko H, Jeon K, Choi D, Park CM. Temporal inversion for learning interval change in chest X-rays. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2026:28338-28347. [Crossref]
54. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal.

medRxiv [Preprint]. 2020:2020.09.13.20193565. [Crossref] Update in: This article has been published with [Crossref]

55. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med*. 2019;2:31. [Crossref]
56. Kocak B, Klontzas ME, Stanzione A, et al. Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations. *Eur J Radiol Artif Intell*. 2025;3:100030. [Crossref]
57. Moon JH, Choi G, Rabaey P, et al. LUNGUAGE: a benchmark for structured and sequential chest X-ray interpretation. In: Proceedings of the 7th Conference on Health, Inference, and Learning. PMLR; 2026:148-197. Accessed 2026 Aug 22. [Crossref]
58. Ostmeier S, Xu J, Chen Z, et al. GREEN: generative radiology report evaluation and error notation. In: Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, FL: Association for Computational Linguistics; 2024:374-390. [Crossref]
59. Yu F, Endo M, Krishnan R, et al. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns (N Y)*. 2023;4(9):100802. [Crossref]
60. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;6(4):e240300. [Crossref]
61. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*. 2022;377:e070904. [Crossref]
62. Rosenkrantz AB, Smith SW, Recht MP, Horwitz LI. Perceptions of radiologists and emergency medicine providers regarding the quality, value, and challenges of outside image sharing in the emergency department setting. *AJR Am J Roentgenol*. 2020;214(4):843-852. [Crossref]
63. Park SH, Sul AR, Ko Y, Jang HY, Lee JG. Radiologist's guide to evaluating publications of clinical research on AI: how we do it. *Radiology*. 2023;308(3):e230288. [Crossref]
64. Chaibub Neto E, Pratap A, Perumal TM, et al. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *NPJ Digit Med*. 2019;2:99. [Crossref]
65. Rumala DJ. How you split matters: data leakage and subject characteristics studies in longitudinal brain MRI analysis. In: Wesarg S, Puyol Antón E, Baxter JSH, et al., editors. Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging. Lecture Notes in Computer Science, vol 14242. Cham: Springer Nature Switzerland; 2023:235-245. [Crossref]
66. Rosenkrantz AB, Duszak R Jr, Babb JS, Glover M, Kang SK. Discrepancy rates and clinical impact of imaging secondary interpretations: a systematic review and meta-analysis. *J Am Coll Radiol*. 2018;15(9):1222-1231. [Crossref]
67. Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*. 2023;307(4):e222176. [Crossref]
68. Taib AG, Partridge GJW, Phillips P, et al. Automation bias in action: eye tracking of humans reading screening mammograms with and without AI prompts. *Radiology*. 2026;320(1):e252590. [Crossref]
69. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*. 2025;388:e081554. [Crossref] Erratum in: *BMJ*. 2025;388:r340. [Crossref]
70. Li T, Zuo L, Liu Y, et al. Contrastive patient-level pretraining enables longitudinal and multimodal fusion for lung cancer risk prediction. In: Proceedings of the 8th International Conference on Medical Imaging with Deep Learning. PMLR; 2026:1008-1020. Accessed 2026 Aug 22. [Crossref]
71. Kiechle J, Osuala R, Lang DM, et al. Graph representation learning of longitudinal medical imaging trajectories for treatment response prediction. *arXiv [Preprint]*. 2026:arXiv:2607.04912. [Crossref]
72. Kondepudi A, Rao A, Zhao C, et al. Health system learning enables generalist neuroimaging models. *Nat Med*. 2026;32(8):2831-2837. [Crossref]