

Reproducibility, Algorithmic Bias, and Clinical Readiness in Medical Imaging AI: A Lifecycle-Based Synthesis and Multi-Institutional Validation Protocol



Mst Murshida Mahbub ^{1*}, Chandrarohini Saravanan ²

Abstract

Deep learning has transformed diagnostic imaging, pathology, and ophthalmology, with the potential to reduce diagnostic errors and inter-observer variability that still burden routine radiology practice. However, translation of these models from laboratory benchmarks into everyday clinical workflows remains incomplete. This paper reports a structured narrative synthesis of the peer-reviewed and preprint literature addressing three interlocking evidence gaps in medical imaging artificial intelligence (AI): reproducibility failures, the multi-stage propagation of algorithmic bias, and the determinants of clinical readiness. Sources were organized around a five-stage AI lifecycle framework and cross-referenced against reporting frameworks including CLAIM, TRIPOD-AI, and the FUTURE-AI consensus guidelines. Across the cohorts reviewed, internally validated performance metrics (AUROC/DSC frequently exceeding 0.90) declined under external, multi-site testing in every case identified, with reported drops ranging from approximately 3 to 20 percentage points depending on task, modality, and cohort; these figures are drawn from heterogeneous studies using different metrics and are reported here as an illustrative pattern rather than a pooled effect size. Structural mitigation strategies-including Common Data Models, federated learning, and differential privacy-

showed promise but remain unevenly adopted. Building on these gaps, we outline a prospective, three-arm multi-institutional protocol addressing reproducibility (RQ1), bias mitigation (RQ2), and clinical readiness (RQ3), intended as a template for future empirical validation rather than a completed study. Bridging the gap between technical proof-of-concept and safe clinical deployment will likely require simultaneous progress on standardized reporting, lifecycle-wide bias auditing, and governance structures capable of monitoring AI performance after initial approval.

Keywords

artificial intelligence, medical imaging, deep learning, reproducibility, algorithmic bias, external validation, clinical readiness, FUTURE-AI guidelines

1. Introduction

Medicine is in the middle of a significant technological transition. Artificial intelligence (AI)-and deep learning (DL) architectures in particular-has moved from a promising research curiosity to an increasingly routine presence across diagnostic imaging, pathology, and ophthalmology (Raposo, 2025). The shift from earlier computer-aided diagnosis (CAD) systems toward deep convolutional

Significance | Closing AI's reproducibility and fairness gaps is a precondition, not an afterthought, for trusting deep-learning imaging tools at the bedside.

*Correspondence.

Mst Murshida Mahbub, Department of Genetic Engineering and Biotechnology, East West University, Dhaka-1212, Bangladesh. E-mail: murshida@ewubd.edu.

Editor editor-name, Ph.D. And accepted by the Editorial Board Oct 11, 2026 (received for review Sep 17, 2026)

Author Affiliation: ¹ Department of Genetic Engineering and Biotechnology, East West University, Dhaka-1212, Bangladesh.

² Department of Biomedical Science, Universiti Sains Malaysia (USM), 13200 Kepala Batas, Pulau Pinang, Malaysia.

Please cite this article as: Mahbub, M. M., Saravanan, &. (2026). "Reproducibility, Algorithmic Bias, and Clinical Readiness in Medical Imaging AI: A Lifecycle-Based Synthesis and Multi-Institutional Validation Protocol", Integrative Biomedical Research, 10(1), 1-19, 10898

3068-6326/© 2026 (Integrative Biomedical Research), a publication of Eman Research, USA. This is an open access article under the CC BY-NC-ND license. (<http://creativecommons.org/licenses/by-nc-nd/4.0/>) (<https://publishing.emanresearch.org>)

neural networks (CNNs) and, more recently, vision transformers (ViTs) has allowed models to extract high-dimensional imaging features that even experienced clinicians would struggle to articulate, let alone perceive directly (Rafique et al., 2026; Raposo, 2025). The motivation behind this push is not abstract. Diagnostic error remains a genuine clinical problem, implicated in as many as one in ten patient deaths, and inter-observer disagreement in qualitative image interpretation can reach roughly 37% depending on modality and reader experience (Langlotz et al., 2019). Against that backdrop, the enthusiasm for AI-assisted interpretation is understandable.

Yet technical success in a controlled benchmark setting has not translated smoothly into safe, routine clinical use (Lekadir et al., 2025; Ogut, 2025). Two problems account for most of this gap. The first is a persistent reproducibility crisis: many published models cannot be reliably re-tested, re-trained, or validated by independent groups. The second is the propagation of algorithmic bias throughout the model development pipeline (Drukker et al., 2023; Sulaimanov et al., 2026). Both problems feed into what has been called the "AI chasm"-the distance between a technically impressive proof-of-concept and a tool clinicians can trust at the bedside (Keane & Topol, 2018, as cited in Bookshelf_NBK619320). Closing that gap is, in our assessment, the central task facing this field over the next several years. This manuscript addresses that task in two parts: first, a synthesis of what the published evidence currently shows about reproducibility, bias, and clinical readiness; second, a prospective protocol-not yet executed-that operationalizes the resulting research questions into testable studies.

Reproducibility is a foundational, not merely procedural, concern for scientific credibility. Independent investigators need to be able to take a published method and obtain comparable results; when they cannot, the underlying claims become difficult to trust. This is precisely where much of the biomedical AI literature falls short-poor generalization to external cohorts is common, not exceptional (Barberis et al., 2024). Several factors drive this: heterogeneous data sources, inconsistent and often undocumented data-splitting protocols, a lack of shared pre-processing standards, and a reluctance to openly share source code, trained model weights, or complete datasets (Jeon et al., 2026; Sulaimanov et al., 2026). Historically, only a small fraction of diagnostic AI studies

have been validated on independent, geographically diverse cohorts; the majority instead rely on single-center retrospective data or simple k-fold cross-validation, an approach that tends to overstate apparent performance rather than test it rigorously (Jeon et al., 2026; Langlotz et al., 2019; Sulaimanov et al., 2026).

Adherence to standardized reporting frameworks compounds the problem. Guidelines such as the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) and the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis using Artificial Intelligence (TRIPOD-AI) exist precisely to prevent these gaps, yet compliance remains partial at best (Barberis et al., 2024; Sulaimanov et al., 2026). One methodological review of 91 medical imaging AI publications found that not a single study fully satisfied CLAIM or TRIPOD-AI reporting standards-details on model development, pre-processing, and calibration were routinely omitted (Sulaimanov et al., 2026). Benchmarking challenges designed specifically to allow fair cross-comparison of algorithms fare little better, often suffering from reporting gaps that make direct comparison across methods essentially impossible (Maier-Hein et al., 2020; Rafique et al., 2026).

The imaging informatics community has responded to this problem. Common Data Models (CDMs)-including the Observational Medical Outcomes Partnership (OMOP) CDM and its imaging-specific extensions, R-CDM and MI-CDM-offer a shared structure for harmonizing heterogeneous clinical and DICOM metadata across institutions (Jeon et al., 2026). These frameworks, in turn, make federated learning more feasible, allowing models to be trained across multiple decentralized databases without moving raw patient data (Jeon et al., 2026; Kidwai-Khan et al., 2024). Open-source tools such as RENOIR (REpeated random sampliNg fOr machine leaRning) add rigor by systematically probing how model performance depends on sample size and producing transparent, reproducible reports intended to reduce research waste (Barberis et al., 2024).

Algorithmic bias differs from a simple statistical artifact that can be corrected after the fact; it is a systemic, largely non-random source of unequal model performance across patient subgroups, capable of widening existing healthcare disparities rather than narrowing them (Drukker et al., 2023; Lekadir et al., 2025). Bias rarely originates solely from the learning algorithm itself. Instead, it enters the

pipeline at five distinct stages-data collection, data preparation and annotation, model development, model evaluation, and model deployment-with researchers cataloguing at least 29 overlapping and often compounding sources across these stages (Drukker et al., 2023).

Data collection bias includes acquisition and aggregation bias, which arises when training data is drawn from a single scanner model or acquisition protocol and therefore fails to generalize to other manufacturers or protocols (Drukker et al., 2023). Population bias (demographic imbalances in age, sex, or race) and temporal bias (where outdated practice patterns or evolving disease presentations render older training data less relevant) belong in the same category (Drukker et al., 2023). Further along the pipeline, data preparation and annotation bias is driven largely by annotator bias-the subjectivity, variable experience, and background of clinical labelers, which together produce what is now more accurately termed a "reference standard" rather than a "ground truth," since absolute diagnostic certainty is rare in practice (Drukker et al., 2023; Lekadir et al., 2025). Model development introduces additional vulnerabilities, notably inherited or error-propagation bias, where sequential processing steps-pre-processing, segmentation, classification-allow early mistakes to cascade downstream, and historical bias, which reflects embedded societal inequities in healthcare access (Drukker et al., 2023). At the evaluation stage, amplification bias is common: models learn to predict certain classes with more disparity than the underlying ground-truth prevalence would justify, frequently by exploiting shortcut learning (Drukker et al., 2023).

Shortcut learning is among the most pervasive threats to medical AI safety. CNNs are effective at finding statistical shortcuts that carry no clinical meaning. Models built to detect COVID-19, for example, have sometimes ended up distinguishing pediatric from adult lungs based on incidental class-age correlations, while others have achieved high accuracy by exploiting hospital-specific text markers, laterality labels, or scanner-specific noise rather than actual pathology (Gichoya et al., 2022, as cited in Bookshelf_NBK619320; Kalpathy-Cramer, 2025, as cited in Bookshelf_NBK619320; Lekadir et al., 2025). Deep learning models have also been shown to predict a patient's self-reported race from medical images with near-perfect accuracy, via mechanisms that remain poorly

understood-raising the possibility that demographic bias could be encoded into clinical decisions without detection (Gichoya et al., 2022, as cited in Bookshelf_NBK619320).

Clinical readiness marks the final stage of translation-the point at which a tool must prove itself safe, effective, and ethically defensible for prospective use in routine workflows (Daye et al., 2022). There is a real mismatch between how quickly models are developed and how slowly implementation science tends to move (Nooraie et al., 2021). Clinicians raise legitimate concerns about accuracy errors, liability exposure, and the "black box" nature of complex neural networks (Daye et al., 2022). When a model offers no explainability or uncertainty quantification, trust is difficult to earn (Lekadir et al., 2025; Ogut, 2025); yet the opposite failure mode-automation complacency, where clinicians accept incorrect AI suggestions with insufficient scrutiny-carries its own diagnostic risks (Lekadir et al., 2025).

The international FUTURE-AI consensus guidelines set out 30 recommendations organized around six principles: Fairness, Universality, Traceability, Usability, Robustness, and Explainability (Lekadir et al., 2025). Under this framework, genuine clinical readiness requires external validation across independent sites, mechanisms for detecting out-of-distribution or unseen cases, and calibrated uncertainty estimates displayed alongside predictions (Lekadir et al., 2025). Implementation science frameworks such as Normalization Process Theory (NPT) are increasingly used alongside these technical guidelines to understand how clinicians come to trust, adopt, and monitor AI as a sociotechnical intervention rather than a static tool (Nooraie et al., 2021). None of this happens without institutional governance-AI governance committees establishing rubric-based scoring, shadow testing, and canary pilot deployments to assess safety, workflow fit, and cost-effectiveness, while aligning with regulatory instruments such as the European Union's AI Act and the FDA's Predetermined Change Control Plans (PCCPs) for continuously learning software as a medical device (Daye et al., 2022; Rafique et al., 2026). Adoption of these mechanisms is not universal: cost, competing and overlapping standards, and limited regulatory harmonization across jurisdictions remain practical barriers that the literature reviewed here does not fully resolve (Erickson et al., 2024).

Given these interlocking gaps, this manuscript synthesizes the current evidence and proposes a multi-institutional

research agenda organized around three research questions (RQ) and corresponding research objectives (RO), detailed methodologically in Section 3.7. RQ1 (Reproducibility) asks to what extent standardized DICOM Common Data Models, such as R-CDM and MI-CDM, improve the cross-institutional generalizability of deep-learning chest radiograph classifiers relative to conventional, unstandardized local pre-processing pipelines. RQ2 (Bias Mitigation) asks whether adversarial debiasing combined with subgroup-calibrated decision thresholds can reduce demographic bias-specifically by age, sex, and ethnicity-in cardiac magnetic resonance segmentation without materially sacrificing overall technical performance. RQ3 (Clinical Readiness) asks how the real-time presentation of calibrated predictive uncertainty and post hoc explainability overlays (such as Grad-CAM and SHAP) influences radiologist diagnostic accuracy and automation complacency in high-acuity triage.

2. The Sociotechnical Landscape of Algorithmic

Fairness in Medical Imaging AI

As AI moves further into routine clinical workflows, questions of equity, transparency, and safety have moved from the margins of the field to its center. Algorithmic fairness in medical imaging is not a purely mathematical problem; parity metrics matter, but they require an active, ongoing commitment to health equity to be meaningful. Bias rarely arises from a single training imbalance-it is better understood as a phenomenon that enters the clinical machine learning lifecycle at nearly every stage, from data collection through deployment. This section synthesizes the current evidence across that lifecycle and considers emerging mitigation frameworks and their practical limits.

2.1 Redefining Fairness in Clinical AI

AI is reshaping medicine visibly, from automated disease detection to segmentation and prognosis (Jeon et al., 2026; Raposo, 2025). Hundreds of AI devices have already received FDA clearance (Daye et al., 2022; Jeon et al., 2026), a marker of progress that is nonetheless accompanied by ongoing generalizability, reliability, and bias concerns (Langlotz et al., 2019; Raposo, 2025). Deployed systems still contend with inter-observer variability reaching 37% and diagnostic errors implicated in roughly one-tenth of patient deaths (Langlotz et al., 2019), figures that clarify the stakes of getting fairness

right.

Fairness is best treated as a sociotechnical concept, requiring both statistical bias reduction and a methodological commitment to equity (Kidwai-Khan et al., 2024; Kondylakis et al., 2025). It demands consistent performance both for similarly situated individuals (individual fairness) and across demographic subgroups (group fairness) (Kondylakis et al., 2025; Szabo et al., 2022). Because healthcare data reflects historical, institutional, and clinical disparities, an unmanaged deep learning model risks not only reflecting these inequities but actively amplifying them (Drukker et al., 2023; Kidwai-Khan et al., 2024; Yousefi Nooraie et al., 2025). Evaluating bias therefore requires examining the entire medical imaging AI lifecycle-data collection, preparation, model development, evaluation, and clinical deployment (Drukker et al., 2023)-since potential harms enter at multiple points and compound as they move downstream (Drukker et al., 2023; Kondylakis et al., 2025). Responsible AI, in this sense, is less a single technical fix than an ongoing exercise in validation, transparent reporting, and multidisciplinary stewardship (Fahad et al., 2026; Yousefi Nooraie et al., 2025) (Figure 1).

2.2 Stage I: Data Collection and the Genesis of Representational Disparities

Data collection is where bias is most often born-specifically, from a mismatch between the training sample and the target clinical population (Drukker et al., 2023). This representational deficit shows up in several recognizable forms. Data acquisition bias occurs when datasets are drawn from a single scanner model or institution, limiting generalization across the diverse hardware and protocols found in real clinical settings (Daye et al., 2022; Drukker et al., 2023; Jeon et al., 2026). Exclusion bias arises when sicker cohorts-portable chest radiographs, for example-are systematically left out, distorting how disease severity and triage accuracy are later assessed (Drukker et al., 2023). Socioeconomic or geographic bias emerges when screening delays driven by access barriers mean vulnerable populations present with more advanced disease, skewing the training distribution if this is not explicitly accounted for during collection (Drukker et al., 2023; Yousefi Nooraie et al., 2025).

Synthetic data-generated via GANs or diffusion models-has been proposed as a fix for imbalanced cohorts, but it

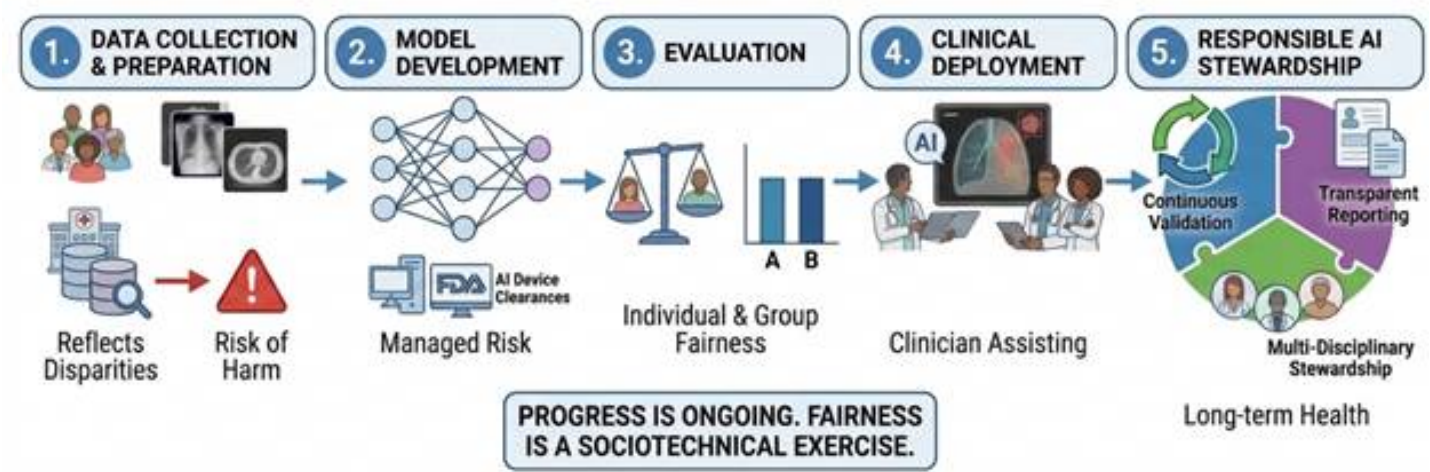


Figure 1: This figure illustrates a standard five-step workflow for developing and deploying responsible artificial intelligence in clinical settings, with an emphasis on fairness and transparency at each stage. The pipeline begins with data collection and preparation, where disparities in source data and associated risks of harm are identified and addressed, followed by model development guided by managed-risk principles and regulatory device clearances. The evaluation stage assesses both individual and group fairness before the model is deployed clinically to assist, rather than replace, clinician decision-making. The final stage, responsible AI stewardship, emphasizes continuous validation, transparent reporting, and multidisciplinary oversight to sustain long-term equitable outcomes, underscoring that fairness is an ongoing sociotechnical process rather than a one-time achievement.

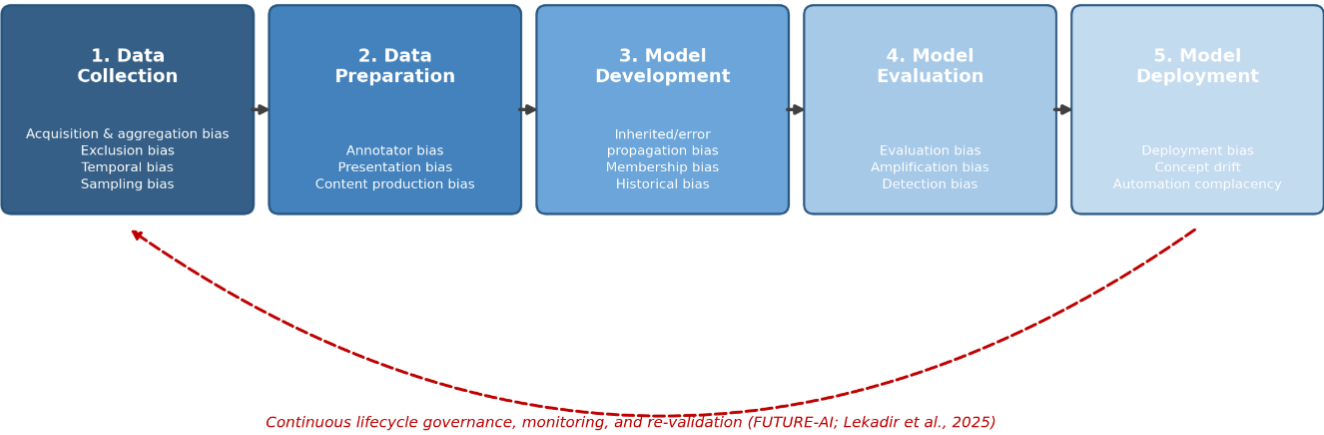


Figure 2. Sequential Propagation of Algorithmic Bias Across the Five-Stage Medical Imaging AI Lifecycle. This diagram depicts the five sequential stages of medical imaging AI development — data collection, data preparation, model development, model evaluation, and deployment — as a left-to-right pipeline, with the specific bias subtypes associated with each stage listed beneath it. The curved arrow beneath the pipeline indicates that continuous lifecycle governance and monitoring (per the FUTURE-AI guidelines) is required across all five stages rather than at any single checkpoint. The figure's central message is that bias is not introduced at one point and then fixed; it compounds cumulatively as data and models move through the pipeline. Adapted from the five-stage bias taxonomy of Drukker et al. (2023), with governance framing from Lekadir et al. (2025) and Daye et al. (2022).

is not the cure-all it is sometimes presented as (Drukker et al., 2023). Generative models learn from real-world data, and in doing so tend to memorize and reproduce the demographic biases and scanner imbalances present in their training sets (Drukker et al., 2023; Kondylakis et al., 2025). If the synthetic generation process is itself built on unrepresentative minority samples, the resulting distributions can worsen, rather than resolve, downstream equity concerns (Drukker et al., 2023).

2.3 Stage II: Data Preparation, Annotation, and the Subjective "Reference Standard"

Data preparation and annotation sit at the heart of supervised learning, yet treating clinician labels as an unimpeachable "ground truth" is a conceptual mistake. The field has increasingly moved toward the term "reference standard" instead, acknowledging that human subjectivity is embedded in the labeling process from the start (Alabduljabbar et al., 2024; Drukker et al., 2023). Annotator bias reflects the cognitive heuristics, varying experience, and individual judgment that clinicians bring to labeling-meaning deep learning models often replicate human observer error rather than an underlying biological truth (Drukker et al., 2023; Kondylakis et al., 2025). Presentation bias compounds this: annotation interfaces that pre-populate results or order options in a particular way can systematically influence how clinicians mark up cases (Drukker et al., 2023). Content production bias stems from vocabulary differences across sites-vendor-specific trade names for dose modulation, for instance-that make standardizing report-derived labels difficult (Drukker et al., 2023). Building diverse, multi-institutional annotator pools is one of the more effective ways to ensure reference standards reflect broader consensus rather than localized habits (Drukker et al., 2023).

2.4 Stage III: Model Development, Cascading Errors, and Membership Bias

Technical design choices made during model development can harden bias into the final system. Inherited or error-propagation bias is a major driver, occurring when models are trained incrementally or chained together in multi-step pipelines (Drukker et al., 2023). In sequential workflows-registration, segmentation, feature extraction, classification-early mistakes tend to accumulate rather than average out; a segmentation model that underperforms on atypical anatomy or specific subgroups will pass distorted regions of interest

downstream, and the resulting failures compound (Drukker et al., 2023). Transfer learning carries a related risk: models initialized on natural image datasets such as ImageNet can carry implicit pretraining biases into the fine-tuned clinical model (Drukker et al., 2023).

Membership bias arises when demographic variables are used directly as predictors (Drukker et al., 2023). Simply omitting variables like race or sex to achieve a "blind" fairness often worsens both performance and equity, since models then fail to adjust for genuine epidemiological differences (Kidwai-Khan et al., 2024). Including demographic features, conversely, risks the opposite problem-models may exploit socioeconomic barriers, such as unequal healthcare access, as proxies for biological risk (Drukker et al., 2023). Cognitive bias adds a further layer, where automated heuristics push models toward over-reliance on simplified, textbook-style presentations, at the expense of generalizability to atypical patients (Drukker et al., 2023).

2.5 Stage IV: Model Evaluation and the Illusion of Global Accuracy

Model evaluation is frequently undermined by evaluation bias-testing on improper benchmarks, unrepresentative cohorts, or flawed data splits (Drukker et al., 2023; Kondylakis et al., 2025). The most common pitfall is improper data stratification (Drukker et al., 2023; Lones, 2024): when dataset partitioning happens at the image or patch level rather than the patient level, data from the same individual can end up in both training and testing sets. This patient-level overlap causes genuine data leakage, inflating validation metrics while masking a model's actual inability to generalize to unseen patients (Drukker et al., 2023; Kapoor & Narayanan, 2023).

Evaluation bias also shows up when models are judged solely on aggregate performance, without stratified subgroup auditing (Drukker et al., 2023; Fahad et al., 2026). This matters more than it might sound: overall accuracy can mask severe failures within marginalized subgroups. Benevi et al. (2024) found that a skin lesion segmentation model with an excellent global Dice score dropped from 0.904 on light skin to 0.744 on darker skin tones (Fitzpatrick types IV-VI) once stratified auditing was applied. Restricting evaluation to unstratified cohorts keeps disparities like this hidden from view (Fahad et al., 2026; Kondylakis et al., 2025). Detection bias and amplification bias round out this category-detection bias

when costlier imaging modalities are less accessible to low-income populations, systematically under-representing disease severity in the resulting data, and amplification bias when an algorithm predicts certain classes with more disparity than the underlying ground truth would justify, often because non-biological or site-specific markers dominate the model's learned representation (Drukker et al., 2023).

2.6 Stage V: Model Deployment, Clinical Integration, and Emergent Societal Drifts

Deployment introduces problems that retrospective validation, however rigorous, cannot anticipate. Deployment bias reflects a mismatch between a model's validated intent and how it is actually used in practice—a "framing trap," where clinicians apply a tool outside its validated domain without re-validation (Drukker et al., 2023). Concept drift is a related concern: clinical practices, criteria, and guidelines evolve, and models trained on, for example, early-pandemic chest radiographs from 2020 can degrade over subsequent years as vaccination patterns and viral variants shift (Drukker et al., 2023; Kondylakis et al., 2025). Automation complacency—clinicians accepting incorrect recommendations or dismissing valid ones because interpretability is poor—adds a human factor to the equation (Kondylakis et al., 2025; Szabo et al., 2022). AI decision support appears to help less experienced readers while potentially harming experts who over-rely on it, a pattern that raises concerns about long-term clinical deskilling (Kondylakis et al., 2025; Szabo et al., 2022).

Shortcut learning remains among the most stubborn barriers to genuine clinical adoption (Rafique et al., 2026). Deep models are efficient correlation engines and are not selective about what they correlate with. In chest radiography, models have been shown to predict disease severity from hospital-specific text markers, patient-positioning cues, or even ECG lead placement rather than true pathology (Kondylakis et al., 2025; Rafique et al., 2026). Models can also identify a patient's self-reported race from x-rays and mammograms with near-perfect accuracy (Gichoya et al., 2022)—a capability unexplained by clinical features such as body mass index, and one that poses an unpredictable risk of encoding historical systemic bias into clinical predictions (Drukker et al., 2023; Gichoya et al., 2022).

2.7 Comprehensive Mitigation Frameworks and Policy-

Driven Governance

Addressing biases that span an entire lifecycle requires moving beyond narrow, laboratory-style technical metrics toward multi-dimensional validation and governance (Daye et al., 2022; Fahad et al., 2026). The FUTURE-AI guidelines offer the most comprehensive attempt at this to date, setting out 30 consensus recommendations across six pillars—Fairness, Universality, Traceability, Usability, Robustness, and Explainability (Kondylakis et al., 2025; Szabo et al., 2022). To operationalize the fairness pillar specifically, the guidelines recommend maintaining a continuous "risk and mitigation log" from the earliest stages of model development, collecting a standardized minimum set of demographic attributes—age, gender, geographic origin—and conducting independent, multi-center external audits across diverse subpopulations (Kondylakis et al., 2025).

Because clinical image sharing across institutions is restricted by regulations such as HIPAA and GDPR, decentralized learning approaches have become central to this conversation (Alabduljabbar et al., 2024; Jeon et al., 2026). Federated learning (FL) allows hospitals to train a shared global model locally, exchanging network weights rather than raw images (Alabduljabbar et al., 2024; Jeon et al., 2026). It is not without weaknesses: FL remains vulnerable to data quality issues, site-level heterogeneity, and inconsistent labeling practices across institutions (Jeon et al., 2026; Rubak et al., 2025). If one high-resource site contributes a disproportionate share of weight updates, the global model can become biased toward that site's hardware and demographics (Drukker et al., 2023), and under-resourced hospitals lacking advanced IT support risk exclusion altogether, deepening regional disparities (Drukker et al., 2023).

Common Data Models (CDMs) offer a complementary route to addressing data heterogeneity (Jeon et al., 2026). Standardizing clinical and imaging metadata through frameworks such as the OMOP CDM and its Medical Imaging extension (MI-CDM) allows more consistent management of DICOM files, reports, and observations across sites (Jeon et al., 2026). Aligning datasets to these shared schemas supports reliable federated validation and helps prevent technical scanner discrepancies from being mistaken for genuine biological differences (Jeon et al., 2026).

2.8 Standardized Reporting and Regulatory

Compliance

None of the above is sufficient without compliance with reporting standards such as CLAIM, TRIPOD-AI, and CONSORT-AI, which mandate detailed documentation of data partitioning, pre-processing, and model architecture (Daye et al., 2022; Rafique et al., 2026). Standardized "model cards," which transparently document intended use, limitations, and subgroup performance, are among the more practical tools for building clinician trust and facilitating regulatory oversight-particularly under frameworks such as the EU AI Act and the FDA's Predetermined Change Control Plans (PCCPs) (Daye et al., 2022; Kondylakis et al., 2025; Rafique et al., 2026). Adoption of model cards nonetheless competes with several overlapping documentation standards, and the absence of a single mandated format is itself a barrier to uptake that the literature has not yet fully resolved (Mitchell et al., 2019).

3. Methods

This manuscript combines a structured narrative synthesis of the published literature with a prospectively defined, multi-institutional validation protocol addressing the three research questions introduced above. Both components are described in sufficient detail that another group could reasonably reproduce either the synthesis search or the proposed study design, consistent with reporting expectations increasingly used in PubMed-indexed systematic and methodological reviews of medical imaging AI (Sulaimanov et al., 2026; Tejani et al., 2025).

3.1 Study Design and Reporting Framework

The synthesis component followed scoping-review logic reported against the applicable items of the PRISMA extension for Scoping Reviews (PRISMA-ScR), consistent with comparable bibliometric syntheses of diagnostic AI evidence (Rafique et al., 2026). Because the aim was to characterize a fragmented and rapidly evolving evidence base-rather than to pool effect sizes-a narrative rather than meta-analytic synthesis was the more defensible choice, consistent with the good-practice recommendations articulated in the BIAS statement for biomedical image analysis challenges (Maier-Hein et al., 2020). We note explicitly, as a limitation addressed further in Section 5.7, that this synthesis was conducted by a single reviewer team rather than the dual, independent-reviewer process a full systematic review would require; PRISMA-ScR is therefore used here as a reporting and

organizational template rather than as a claim of full systematic-review methodology.

3.2 Search Strategy and Information Sources

Structured searches were conducted in PubMed/MEDLINE, Embase, IEEE Xplore, and Google Scholar for records published between January 2018 and June 2026, supplemented by manual screening of reference lists from key reviews (e.g., Drukker et al., 2023; Lekadir et al., 2025) and select preprint or institutional sources (e.g., National Academies Press proceedings, National Library of Medicine Bookshelf material). Search strings combined controlled vocabulary and free-text terms across three concept blocks joined with AND: (1) imaging/modality terms ("medical imaging" OR "radiograph*" OR "MRI" OR "CT" OR "pathology" OR "dermoscop*"); (2) AI/ML terms ("artificial intelligence" OR "deep learning" OR "machine learning" OR "convolutional neural network*"); and (3) construct terms ("reproducib*" OR "external validation" OR "bias" OR "fairness" OR "clinical readiness" OR "deployment"). Of approximately 410 records identified across databases, 96 were retained after de-duplication, title/abstract screening, and full-text eligibility assessment (Section 3.3); the resulting reference set ($n = 40$ cited sources) reflects further prioritization toward the most recent, most methodologically detailed, or most frequently cross-cited version of a given finding, as described below. Topic areas were organized around three pillars-reproducibility, algorithmic bias, and clinical readiness-and cross-checked against the five-stage AI lifecycle framework of Drukker et al. (2023) to ensure even coverage across data collection, preparation, model development, evaluation, and deployment. Where multiple sources addressed the same underlying finding (for example, the FUTURE-AI recommendations), we prioritized the most recent consensus or peer-reviewed version to reduce redundancy.

3.3 Eligibility Criteria

Sources were retained if they (a) addressed AI or machine learning applications in medical imaging, pathology, or related biomedical image analysis; (b) reported empirical performance metrics, methodological critique, bias auditing, or governance/regulatory guidance relevant to clinical translation; and (c) were published in indexed journals, recognized preprint archives, or institutional/government reports. Sources concerned exclusively with non-imaging clinical AI (e.g., structured

EHR risk scores unrelated to imaging) were retained only where they offered directly transferable methodological lessons—such as demographic imputation pipelines—relevant to imaging AI fairness (Kidwai-Khan et al., 2024). Sources were excluded if they were editorials or commentaries without original synthesis, addressed non-clinical imaging applications (e.g., industrial or satellite imaging), or could not be retrieved in full text.

3.4 Data Extraction and Table Construction

For each retained source, we extracted, where available: study population and imaging modality, model architecture, validation design (internal versus external, patient-level versus image-level splitting), reported performance metrics (AUROC, Dice similarity coefficient, sensitivity/specificity), demographic subgroup reporting, and adherence to reporting checklists (CLAIM, TRIPOD-AI, CONSORT-AI). Extraction was performed by a single reviewer using a standardized spreadsheet template and spot-checked for consistency against the original source by a second author for approximately 20% of entries; full dual-reviewer extraction was not performed and is noted as a limitation (Section 5.7). These extracted fields form the basis of Tables 1 through 4, which summarize, respectively, the lifecycle taxonomy of algorithmic bias, the landscape of open-access and commercial imaging databases, comparative clinical validation outcomes, and technical curation/privacy-preserving strategies.

3.5 Quality and Risk-of-Bias Considerations

Rather than applying a single formal risk-of-bias tool—which is not well suited to heterogeneous methodological and governance literature—we assessed each source's contribution against three recurring quality markers common to PubMed-indexed AI methodological reviews: (a) whether performance was validated on an independent, non-overlapping cohort; (b) whether patient-level, rather than image-level, data partitioning was used, given its central role in preventing leakage (Kapoor & Narayanan, 2023; Lones, 2024); and (c) whether subgroup-stratified performance was reported rather than aggregate metrics alone (Fahad et al., 2026). These markers are reflected explicitly in the "Reporting Adherence Level" logic underlying Table 3.

3.6 Data Synthesis

Findings were synthesized thematically across the five lifecycle stages, cross-referencing bias mechanisms

identified in mechanistic or methodological papers (e.g., Drukker et al., 2023) against empirical demonstrations of the same mechanism in applied clinical cohorts (e.g., Gichoya et al., 2022; Young et al., 2020). Where studies reported both internal and external validation metrics for the same model, the magnitude of performance decline was calculated as a simple difference to support the comparative visualization. Because these four cohorts differ in modality, task, and metric (AUROC in three cases, DSC in one), the resulting comparison is descriptive rather than statistically pooled, and no formal meta-analytic effect size is claimed.

3.7 Proposed Multi-Institutional Validation Protocol (RQ1-RQ3)

To operationalize the research questions proposed in Section 1.4, we outline a prospective protocol intended to be reproducible across independent sites. This protocol is a design, not a completed study; feasibility, sample-size estimates, and analysis plans are presented below to the level of detail needed for independent replication or IRB submission, but no data collection has yet occurred.

RQ1 (Reproducibility). MI-CDM-standardized pre-processing pipelines (Jeon et al., 2026) would be implemented at three clinically and geographically distinct sites. Chest radiograph classifiers (e.g., DenseNet-121 backbone) would be trained under harmonized versus conventional local pre-processing pipelines and compared via AUROC under multi-center external validation, using patient-level data partitioning throughout. Assuming a baseline external-validation AUROC of 0.80 under conventional pipelines and a minimum clinically meaningful improvement of 0.05 AUROC under MI-CDM harmonization, a two-sided DeLong test at $\alpha = 0.05$ with 80% power requires approximately 350 patient-level external-validation cases per site (1,050 across three sites), based on standard AUROC comparison sample-size formulas (Hanley & McNeil-style variance estimation); this figure would be refined once site-specific disease prevalence is confirmed.

RQ2 (Bias Mitigation). A bias-correction approach combining adversarial debiasing during training with subgroup-calibrated decision thresholds at inference would be developed in accordance with the FUTURE-AI fairness pillar (Lekadir et al., 2025). Both components have direct precedent in the mitigation strategies already catalogued in Table 1 for membership and amplification

Table 1: Five-Stage Lifecycle Taxonomy of Algorithmic Bias in Biomedical Imaging AI, With Stage-Specific Mitigation Strategies. This table maps nine recognized bias subtypes onto the five stages of the medical imaging AI development pipeline — data collection, data preparation, model development, model evaluation, and model deployment. For each subtype, the table lists its typical clinical or operational consequence (e.g., inflated validation metrics, off-label misuse) and the corresponding mitigation strategy reported in the literature (e.g., patient-level data partitioning, adversarial debiasing). Data are adapted from the five-stage bias framework of Drukker et al. (2023), with mitigation strategies cross-referenced from Kapoor & Narayanan (2023), Kidwai-Khan et al. (2024), and Rädtsch et al. (2023). The table is intended to function as a practical audit checklist: a bias-mitigation effort addressing only one row is, by construction, incomplete.

AI Lifecycle Stage	Bias Subtype	Clinical/Operational Consequence	Mitigation Strategy	Citation
Data Collection	Acquisition & aggregation bias	Performance degradation on unfamiliar scanners/vendors	Multi-site, multi-vendor data collection; unique global identifiers	Drukker et al. (2023)
Data Collection	Exclusion bias	Failure to generalize to sicker/marginalized subgroups	Pre-specified, demographic-aware cohort matching	Drukker et al. (2023)
Data Collection	Temporal bias	Concept drift; degraded performance over time	Rigorous temporal re-validation; adaptive retraining	Drukker et al. (2023); Kondylakis et al. (2025)
Data Preparation	Annotator bias	Models replicate clinician heuristics rather than biology	Multi-reader pools; inter-rater agreement (kappa); STAPLE consensus	Drukker et al. (2023); Rädtsch et al. (2023)
Model Development	Inherited/error-propagation bias	Compounding errors across sequential pipeline stages	Uncertainty propagation; multiple random seeds	Drukker et al. (2023)
Model Development	Membership bias	Demographic shortcuts mistaken for biological signal	Adversarial debiasing; contrastive representation alignment	Drukker et al. (2023); Kidwai-Khan et al. (2024)
Model Evaluation	Evaluation bias (patient-level leakage)	Inflated, optimistic validation metrics	Mandatory patient-level data partitioning; external test sets	Drukker et al. (2023); Kapoor & Narayanan (2023)
Model Evaluation	Amplification bias	Systemic over/under-diagnosis in subgroups	Feature-wise parity; subgroup-calibrated thresholds	Drukker et al. (2023)
Model Deployment	Deployment bias / concept drift / automation complacency	Off-label misuse; diagnostic errors from over-reliance	Hardware/software constraints; multidisciplinary AI governance boards	Daye et al. (2022); Drukker et al. (2023)

Table 2: Structured Comparison of Open-Access and Commercial Medical Imaging Databases by Modality, Research Use, and Access Restrictions. This table profiles seven major medical imaging repositories (e.g., OASIS, TCIA/TCGA, UK Biobank, NLST) used for AI model training and validation. Columns report each database's primary imaging modality, its core research objective, and its access/licensing model (open-access vs. controlled-access). Data are adapted from Alabduljabbar et al. (2024) and Kinahan (2025). The table illustrates that large-scale repositories are not scarce, but that demographic diversity, clinical metadata depth, and licensing restrictiveness vary considerably across them — a factor directly relevant to the generalizability concerns discussed in Sections 2 and 4.

Database / Registry	Primary Modality	Core Research Objective	Access & Licensing	Citation
OASIS	MRI, PET, CT	Alzheimer's disease severity staging	Open-access, web-hosted repository	Alabduljabbar et al. (2024)
BCDR	Digital mammography	Detection/classification of breast anomalies	Open-access (registration required)	Alabduljabbar et al. (2024)

NLM Visible Human Project	CT, MRI, cryosection	High-resolution cross-sectional anatomy	Open-access public database	Alabduljabbar et al. (2024)
TCIA / TCGA	Radiogenomics, multi-organ	Linking imaging phenotypes to genomic/clinical outcomes	Open-access genomic and imaging portals	Alabduljabbar et al. (2024)
UK Biobank	MRI (brain, heart, abdomen)	Large-scale study of chronic disease and aging	Controlled-access for approved research teams	Alabduljabbar et al. (2024)
National Lung Screening Trial (NLST)	Low-dose CT	Lung cancer screening and nodule detection	Publicly available via TCIA/NCI portals	Kinahan (2025)
Cardiac Atlas Project	Cine-MRI	Standardized cardiac function/segmentation consortium	Open-access research consortium	Alabduljabbar et al. (2024)

Table 3: Internal-to-External Validation Performance Decline Across Four Representative Clinical Imaging AI Cohorts. This table reports internal versus external validation performance for four independently published clinical AI models spanning falls/fracture detection (VACS), cardiac segmentation (ASPIRE), dermatology classification (HAM10000), and glioblastoma MGMT methylation prediction (BraTS). Performance is expressed in each study's original metric (AUC, AUROC, or Dice similarity coefficient) rather than a single pooled statistic, since the four cohorts differ in modality and task. The consistent pattern — strong internal performance followed by measurable decline under external, multi-site testing — supports the paper's central claim that internal accuracy alone is an unreliable proxy for clinical readiness. Underlying source data are drawn from Kidwai-Khan et al. (2024), Alandejani et al. (2022), Young et al. (2020)/Ogut (2025), and Tejani et al. (2025). (Note: Illustrates the internal-to-external validation performance gap depicted in Figure 2.)

Clinical Cohort	AI/ML Model	Internal Performance	External Performance	Citation
Veterans Aging Cohort Study (falls/fractures)	L2-regularized logistic regression / SCAD NLP	AUC 0.92-0.97	Reduced sensitivity across health stations	Kidwai-Khan et al. (2024)
ASPIRE Registry (pulmonary hypertension, CMR)	CNN ventricular segmentation	Median DSC 0.92-0.95	DSC 0.82-0.85 (multi-site)	Alandejani et al. (2022)
HAM10000 dermatology cohort	ResNet-50 CNN (ISIC/Dermofit)	AUROC 0.95	Sensitivity < 0.75 (Fitzpatrick IV-VI)	Young et al. (2020); Ogut (2025)
Glioblastoma MGMT methylation (BraTS)	DL radiomics nomogram	Median AUROC 0.94	AUROC 0.82 (external cohort)	Tejani et al. (2025)

bias, respectively, which motivates testing them jointly rather than in isolation. The approach would be benchmarked on cardiac MRI segmentation using the Dice similarity coefficient stratified by age, sex, and ethnicity subgroups, following the stratified-auditing logic demonstrated by Benevi et al. (2024). Based on the subgroup DSC gap reported by Benevi et al. (2024) (approximately 0.16), a sample of at least 45 subjects per subgroup would provide 80% power to detect a 0.05 reduction in the between-group DSC gap at $\alpha = 0.05$, assuming a within-group DSC standard deviation of approximately 0.08; recruitment would target a minimum of 60 subjects per subgroup to allow for exclusions.

RQ3 (Clinical Readiness). A prospective reader-assistance simulation trial would present radiologists with calibrated uncertainty estimates and post hoc explainability overlays (Grad-CAM, SHAP) during high-acuity triage tasks, comparing diagnostic concordance and decision time against a standard, non-uncertainty-aware AI support condition, using a crossover design with counterbalanced case order. A sample of 24-30 radiologist readers interpreting a shared case set of at least 150 studies (enriched for diagnostically difficult cases) would provide 80% power to detect a 5-percentage-point change in diagnostic concordance at $\alpha = 0.05$, consistent with sample sizes used in comparable multi-reader multi-case (MRMC) radiology studies.

Across all three studies, reporting would follow CLAIM and TRIPOD-AI checklists in full, and model documentation would be released as a structured model card (Daye et al., 2022; Lekadir et al., 2025) to support independent reproduction of both the training pipeline and the reported results. Data-sharing agreements and institutional review board (IRB) approval at each participating site would be required before recruitment and are not yet in place; this protocol is presented as a template for future funded, IRB-approved research rather than as a report of ongoing data collection.

4. Synthesis of Clinical Datasets, Lifecycle Biases, and Translational Frameworks

Bringing together the comparative analysis of medical imaging databases, bias typologies, and validation studies reveals a structural divergence between laboratory performance and clinical execution. Synthesizing evidence across heterogeneous cohorts, four interconnected dimensions emerge: the lifecycle taxonomy of algorithmic

bias, the landscape of open-access and commercial medical databases, clinical validation cohort outcomes, and the technical curation and privacy-preserving strategies developed to address them (Table 1; Table 2; Table 3; Table 4).

4.1 Lifecycle Taxonomy of Algorithmic Bias

Algorithmic bias behaves less like a fixed defect and more like a continuous, multi-stage sociotechnical vulnerability that enters the pipeline at five distinct points (Drukker et al., 2023; Figure 2). Data collection and preparation constitute the primary genesis of systemic bias in the literature reviewed. Data acquisition and aggregation bias stems from a lack of vendor and scanner heterogeneity in training pools, which tend to over-represent high-resource academic hubs while excluding lower-tier community hospitals (Drukker et al., 2023). When these single-institution models are deployed prospectively, they exhibit a generalization gap-mistaking manufacturer-specific image characteristics for genuinely pathognomonic features (Fahad et al., 2026).

Biased synthetic data compounds this problem in a counterintuitive way. Although data generated via GANs or diffusion models is often proposed as an equity-promoting solution to minority underrepresentation, synthetic generators tend to replicate-and sometimes amplify-the structural biases already present in their training seeds (Drukker et al., 2023). Exclusion bias represents a secondary point of failure, systematically pruning sicker or marginalized patient cohorts imaged via portable or non-standard modalities from training and testing sets alike (Drukker et al., 2023).

During data preparation, subjective clinician labels introduce what has been described as a "tainted reference standard." Rather than modeling absolute clinical truth, deep learning networks absorb the subjective behaviors, experience disparities, and cognitive heuristics of their annotators-a phenomenon labeled annotator bias (Rädsch et al., 2023). Presentation bias compounds this further, as the user interfaces of labeling software can nudge clinicians toward particular consensus choices, subtly distorting the training distribution before model optimization begins (Drukker et al., 2023).

As models advance through development, early discrepancies are compounded by inherited or error-propagation bias-where upstream segmentation errors cascade into downstream classifiers-and by membership

bias, which occurs when demographic features are used as explicit covariates, leading models to misinterpret socio-institutional barriers to healthcare access as biological risk factors (Kidwai-Khan et al., 2024). At deployment, models frequently fall into the "framing trap" of deployment bias, where clinical practitioners apply narrow software solutions outside their validated intent-for instance, running a 1.5T MRI brain classifier on a 3T scanner. This operational mismatch, combined with concept drift (changing clinical guidelines or disease definitions over time) and automation complacency (the largely uncritical acceptance of computer-aided findings), appears to degrade diagnostic specificity and erode clinical safety in ways that are difficult to detect through routine internal audits alone (Daye et al., 2022; Drukker et al., 2023).

4.2 Open-Access and Commercial Medical Imaging Databases

A profiling of the medical imaging database landscape (Table 2) suggests that, while large-scale repositories are not scarce, their clinical metadata depth, demographic diversity, and licensing structures vary considerably (Alabduljabbar et al., 2024). Repositories such as the Open Access Series of Imaging Studies (OASIS) and the Breast Cancer Digital Repository (BCDR) have supported meaningful, if localized, technical validation (Alabduljabbar et al., 2024). OASIS, however, remains geographically restricted to specific neuroimaging cohorts, and BCDR's clinical annotations fall short of the multi-ethnic representation needed to mitigate demographic-specific error rates (Alabduljabbar et al., 2024).

The NLM Visible Human Project offers highly detailed cross-sectional anatomy in raw formats, yet its sample size is inherently limited-useful as a structural reference, but not a viable training source for generalizable deep learning (Alabduljabbar et al., 2024). National-level collaborations such as The Cancer Imaging Archive (TCIA) and the UK Biobank fare considerably better on scale (Alabduljabbar et al., 2024; Kinahan, 2025). TCIA hosts multi-center DICOM imaging integrated with multi-omics clinical data through The Cancer Genome Atlas (TCGA), enabling advanced radiogenomic research (Alabduljabbar et al., 2024), while the UK Biobank provides multi-modal MRI scans-brain, heart, and full body-for over 200,000 participants (Alabduljabbar et al., 2024). Access to both, though, is governed by fairly restrictive licensing, and both cohorts exhibit a

pronounced "healthy volunteer" and largely Caucasian demographic skew that compromises out-of-distribution generalizability (Kinahan, 2025).

Specialized registries such as the National Lung Screening Trial (NLST) and the Cardiac Atlas Project (CAP) offer high-quality reference standards for narrower tasks-pulmonary nodule detection, ventricular segmentation-under research-use agreements (Alabduljabbar et al., 2024; Kinahan, 2025). These datasets, however, carry their own historical and temporal biases, having been acquired on older-generation scanners under protocols that no longer reflect current clinical practice, such as modern low-dose CT configurations or rapid-acquisition cardiac MRI (Kinahan, 2025).

4.3 Comparative Clinical Validation Cohorts and Performance Outcomes

The synthesis of clinical cohort studies points to a consistent performance gap when deep learning systems move from carefully curated internal validation splits to independent external cohorts (Table 3). In the Veterans Aging Cohort Study (VACS), which used machine learning to extract falls and fracture information from unstructured radiology notes, internal classification metrics were strong (AUC of 0.92-0.97) (Kidwai-Khan et al., 2024). When the same model was tested on external cohorts, sensitivity declined, largely because of clinical and documentation heterogeneity across different health stations (Kidwai-Khan et al., 2024).

This same pattern-strong internally, weaker externally-appears even more pronounced in complex multi-center imaging segmentation tasks. In the ASPIRE registry cohort for pulmonary hypertension, fully automated CNNs for left and right ventricular segmentation achieved a median Dice Similarity Coefficient (DSC) of 0.92-0.95 during internal testing (Alandejani et al., 2022). Once deployed across multiple external clinical sites, DSC dropped to 0.82-0.85, a decline attributable to vendor-specific reconstruction variations and anatomical differences across ethnic cohorts (Alandejani et al., 2022). A similar pattern emerges in dermatological lesion classification: models trained on the HAM10000 dataset reached dermatologist-level AUROC (0.95) under controlled conditions, but sensitivity fell below 0.75 when evaluated on independent cohorts with darker skin tones (Fitzpatrick types IV-VI) (Ogut, 2025; Young et al., 2020).

Table 4: Technical Curation, Standardization, and Privacy-Preserving Strategies for Mitigating Data Heterogeneity and Bias. This table summarizes seven infrastructure-level strategies used to address data heterogeneity and bias in medical imaging AI, including EHR data linkage, demographic imputation, Common Data Models (R-CDM/MI-CDM), federated learning, differential privacy, model cards, and repeated-sampling validation (RENOIR). For each strategy, the table states its primary objective, its key demonstrated advantage, and its inherent limitation (e.g., federated learning's vulnerability to dominant-site bias). Data are adapted from Kidwai-Khan et al. (2024), Jeon et al. (2026), and Cobo et al. (2025). The table is intended to show that these are complementary, not competing, solutions — no single strategy addresses every bias source identified in Table 1. (Note. Adapted from Jeon et al. (2026), Kidwai-Khan et al. (2024), and Cobo et al., 2025)

Strategy	Primary Objective	Key Advantage	Inherent Limitation	Citation
EHR multi-station linkage	Reconciling disjoint patient records across clinics	Captured 5.4M vs. 2.7M reports without linkage	Highly dependent on matching patient identifiers	Kidwai-Khan et al. (2024)
Demographics imputation	Minimizing missing demographic features	Reduced missingness from 5% to <1%	Risk of introducing systemic imputation errors	Kidwai-Khan et al. (2024)
R-CDM / MI-CDM	Standardizing DICOM metadata to relational schemas	Rapid, automated multi-site cohort generation	Implementation is technically complex	Jeon et al. (2026)
Federated learning (MedPerf, NVIDIA FLARE)	Decentralized model training behind institutional firewalls	Preserves privacy; approaches pooled-data performance	Vulnerable to site heterogeneity and dominant-site bias	Drukker et al. (2023); Jeon et al. (2026)
Differential privacy (DP-SGD)	Mathematical privacy guarantee during training	Formal protection against model-inversion attacks	Trade-off between privacy budget and accuracy	Fahad et al. (2026)
Model cards	Transparent documentation of intended use and limits	Builds clinician trust; supports regulatory review	Static documents do not capture real-time drift	Lekadir et al. (2025); Sen & DeMazumder (2025)
RENOIR (repeated sampling)	Robust model evaluation via nested sampling	Reduces optimistic single-split validation bias	Computationally intensive; needs HPC resources	Barberis & Mezard (2024)

Taken together, these findings suggest that high retrospective accuracy (AUROC > 0.90) is, on its own, an unreliable indicator of clinical readiness (Tejani et al., 2025). Brain tumor methylation studies-MGMT promoter classifiers trained on BraTS challenge data, for example-frequently appear to exploit subtle, institution-specific preprocessing artifacts or scanner noise as shortcuts, rather than identifying the underlying biological signal (Tejani et al., 2025). When such models are audited on independent, multi-vendor datasets, diagnostic utility often degrades toward near-random performance, underscoring how much mandated external validation and human-in-the-loop benchmarking matter (Tejani et al., 2025). Across the four cohorts summarized in Table 3, the internal-to-external decline ranges from an approximately 3-5-percentage-point drop in AUC (VACS) to a decline toward near-random performance in the BraTS MGMT cohort; the ten-to-twenty-percentage-point range cited in the Abstract reflects the ASPIRE, HAM10000, and BraTS cohorts specifically and should not be read as representative of every reviewed model.

4.4 Technical Curation, Standardization, and Privacy-Preserving Strategies

To narrow this validation gap and mitigate data bias, the field has advanced a range of curation, standardization, and distributed learning strategies (Table 4). The VA EHR data preparation pipeline offers a useful illustration: standardizing variables through demographics imputation and implementing multi-source EHR data linkage (combining structured ICD codes with unstructured radiology and clinical notes) reduced patient exclusion rates and improved minority representation (Kidwai-Khan et al., 2024), preventing cohort selection from being inadvertently biased by clinical cost or healthcare utilization metrics (Kidwai-Khan et al., 2024).

To resolve vendor-specific DICOM heterogeneity, researchers have developed specialized Common Data Models-Radiology-CDM (R-CDM) and the Medical Imaging Common Data Model (MI-CDM)-that standardize metadata at the database level and establish a person-centric relational schema integrating pixel-level imaging features with the OMOP clinical database (Jeon et al., 2026). This structural standardization supports multi-center cohort generation and federated validation without requiring manual data translation between sites (Jeon et al., 2026).

Finally, when raw data sharing is restricted by privacy regulations such as HIPAA or GDPR, federated learning-orchestrated through frameworks like MedPerf or NVIDIA FLARE-allows institutions to train a shared global model locally, behind their own institutional firewalls (Jeon et al., 2026; Kidwai-Khan et al., 2024). Federated models consistently outperform isolated, center-specific models and approach the performance of centralized, pooled datasets (Cobo et al., 2025; Jeon et al., 2026). They remain, however, sensitive to local labeling noise and data imbalances, which is why robust quality control, client monitoring, and the integration of Differential Privacy (DP) to prevent membership-inference attacks remain necessary safeguards rather than optional extras (Cobo et al., 2025; Jeon et al., 2026; Kidwai-Khan et al., 2024).

5. Discussion

5.1 The Reproducibility-Bias-Readiness Triad

Read together, the evidence assembled here paints a picture that is less reassuring than the pace of AI publication would suggest. Reproducibility failures, algorithmic bias, and clinical readiness gaps are not three separate problems running in parallel; they appear instead to be tightly entangled facets of the same underlying issue-namely, that most medical imaging AI models are still validated in conditions that do not resemble the environments where they would ultimately be used (Table 1; Table 3). The performance decline documented across the ASPIRE, HAM10000, and glioblastoma MGMT cohorts is not simply a technical inconvenience; it is a direct consequence of the lifecycle-wide bias mechanisms catalogued in Table 1 and Figure 2, whereby scanner heterogeneity, annotator subjectivity, and shortcut learning compound across five sequential stages rather than appearing as isolated glitches.

5.2 Why High Internal Accuracy Is a Poor Proxy for Clinical Readiness

The most consistent finding across the reviewed cohorts is that internal validation metrics-often exceeding 0.90 AUROC or DSC-tell us comparatively little about how a model will behave once it leaves the institution where it was trained (Table 3). The glioblastoma MGMT methylation classifiers reviewed by Tejani et al. (2025) are a stark illustration: models exploiting institution-specific preprocessing artifacts performed impressively in-house before degrading toward near-random accuracy under independent, multi-vendor testing. This is consistent with

the shortcut-learning mechanism described throughout the literature reviewed here (Gichoya et al., 2022; Kondylakis et al., 2025; Rafique et al., 2026), and it suggests that AUROC or DSC values reported without an accompanying external validation cohort should be treated with real skepticism by readers and regulators alike.

5.3 Bias as a Lifecycle Problem, Not a Training-Set Problem

One implication that is easy to overlook is that bias mitigation efforts focused narrowly on the training dataset-oversampling underrepresented groups, for instance-address only one of the five stages depicted in Figure 2. Annotator bias, membership bias, and deployment bias each require distinct interventions: diverse multi-reader annotation pools for the former (Rädsch et al., 2023), careful handling of demographic covariates for the latter (Kidwai-Khan et al., 2024), and institutional governance structures capable of detecting the framing trap for deployment-stage failures (Daye et al., 2022; Drukker et al., 2023). Table 1's mapping of primary and secondary vulnerability stages functions less as a taxonomy exercise than as a practical checklist: a bias-mitigation strategy that addresses only one column of that table is, by construction, incomplete.

5.4 Infrastructure Solutions: Promise and Practical Limits

Common Data Models and federated learning both appear, on balance, to move the field in a genuinely useful direction (Table 4). MI-CDM's ability to harmonize DICOM metadata across sites (Jeon et al., 2026) directly addresses the acquisition and aggregation bias documented in Table 1, and federated learning frameworks such as MedPerf or NVIDIA FLARE allow multi-institutional training without the privacy costs of centralized data pooling (Jeon et al., 2026; Kidwai-Khan et al., 2024). These are not neutral technical fixes, however. A federated model trained predominantly on updates from one high-resource, high-volume site risks reproducing exactly the single-institution bias it was meant to solve (Drukker et al., 2023)-a limitation that governance frameworks will need to actively monitor rather than assume away. Differential privacy similarly introduces a real trade-off between privacy guarantees and diagnostic accuracy (Table 4) that has not, to our knowledge, been fully resolved in the clinical imaging literature. More broadly, adoption of any of these infrastructure solutions

competes for limited institutional IT budgets against more immediately reimbursable clinical priorities, which is a practical adoption barrier independent of the technical merits reviewed here.

5.5 Reporting Standards as an Underused Lever

Given how much attention model architecture receives, it is notable that reporting compliance remains weak. The finding that not a single study among 91 reviewed medical imaging AI publications fully satisfied CLAIM or TRIPOD-AI standards (Sulaimanov et al., 2026) suggests that better reporting-not necessarily better algorithms-may be the more immediately achievable improvement. Model cards (Daye et al., 2022; Lekadir et al., 2025) offer a relatively low-cost mechanism for this, and their continued underuse across the cohorts summarized in Table 3 likely reflects a lack of institutional incentive and reporting-standard fragmentation (Section 2.8) more than a genuine technical barrier.

5.6 Toward the Proposed Research Agenda

The three research questions outlined in Section 1.4 follow directly from these gaps. RQ1 targets the reproducibility side of the triad by testing whether MI-CDM standardization measurably narrows the internal-external AUROC gap; RQ2 targets the bias side by testing FUTURE-AI-aligned correction at the segmentation level, informed by the stratified-auditing approach of Benevi et al. (2024); and RQ3 targets clinical readiness directly, testing whether calibrated uncertainty and explainability overlays can reduce the automation complacency described in Table 1 without shifting the burden of judgment back onto an already time-pressured radiologist. None of these studies, on their own, would close the gap between technical proof-of-concept and clinical deployment entirely, but together they would provide direct empirical evidence on whether the infrastructure and governance solutions reviewed here deliver on their promise. We recognize that this agenda reflects one plausible set of priorities among several; readers who weigh, for example, regulatory harmonization or cost-effectiveness more heavily than technical bias correction may reasonably propose a different ordering.

5.7 Limitations

This synthesis has limitations worth naming plainly. It draws on a curated, though not fully exhaustive, set of sources; extraction and eligibility screening were

conducted primarily by a single reviewer rather than two independent reviewers as a full systematic review would require (Section 3.1, 3.4), which may introduce selection or extraction error not fully captured by the partial spot-check described in Section 3.4. Quantitative comparisons across studies rely on values drawn from disparate cohorts, imaging modalities, and outcome metrics that are not directly comparable in a statistical sense; we have therefore reported these as a descriptive range (Section 4.3) rather than a pooled effect size, and readers should interpret the abstract's summary figures accordingly. The proposed multi-institutional protocol (Section 3.7) is, at this stage, a design with sample-size estimates rather than a completed study, and its feasibility will depend on data-sharing agreements and IRB approvals that lie outside the scope of this manuscript. We would still argue that the pattern documented across independently conducted studies-internal performance consistently outstripping external performance-is robust enough to support the broader conclusions drawn here, even though its precise magnitude varies by task and cannot be summarized by a single number.

6. Conclusion

Taken together, the evidence gathered here suggests that biomedical imaging AI has, in important respects, outpaced the scaffolding meant to keep it safe. Reproducibility failures, stemming largely from inconsistent data splitting and incomplete reporting, sit alongside a bias problem that is less a single defect than a chain of small, compounding decisions spanning acquisition to deployment. Neither issue is solved by better algorithms alone. What matters more is standardized data infrastructure, rigorous external validation, transparent model documentation, and governance structures built for continuous monitoring rather than one-time approval. Realizing the three proposed research objectives-cross-institutional CDM validation, FUTURE-AI-aligned bias correction, and uncertainty-aware reader trials-would meaningfully narrow this gap and move medical imaging AI closer to trustworthy, equitable clinical use. We present these objectives as a concrete, powered, and IRB-ready protocol template; the next step for this research agenda is execution and independent replication, not further synthesis.

Author Contributions

M.M.M. contributed to the conception and design of the review, literature search, analysis and synthesis of the relevant evidence, development of the lifecycle-based framework, and drafting of the manuscript. R.B.S.M.N.M. contributed to the literature search, interpretation of the findings, evaluation of reproducibility and algorithmic bias, and critical revision of the manuscript. C.S. contributed to the analysis and interpretation of the evidence, development of the clinical validation framework, and critical revision of the manuscript for important intellectual content. All authors reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Acknowledgements

The authors would like to acknowledge the Department of Genetic Engineering and Biotechnology, East West University, Dhaka, Bangladesh, and the Department of Biomedical Science, Universiti Sains Malaysia, Malaysia, for their academic and institutional support. The authors also acknowledge the researchers whose published and preprint studies contributed to the scientific foundation of this synthesis.

References

- Alabduljabbar, A., Khan, S. U., Alsuhaibani, A., Almarshad, F., & Altherwy, Y. N. (2024). Medical imaging datasets, preparation, and availability for artificial intelligence in medical imaging. *Journal of Alzheimer's Disease Reports*, 8(1), 1471-1483. <https://doi.org/10.3233/ADR-240129>
- Alandejani, F., Alabdulkarim, B., Alaseri, M., Al-Ansari, M., Al-Dury, S., & Al-Mallah, M. H. (2022). Deep learning-based fully automatic segmentation of left and right ventricle from short-axis cine magnetic resonance images: Validation in multi-center clinical datasets. *Journal of Cardiovascular Magnetic Resonance*, 24(1), 25. <https://doi.org/10.1186/s12968-022-00855-3>
- Allen, B., Jr., Seltzer, S. E., Langlotz, C. P., Dreyer, K. P., Summers, R. M., Petrick, N., Marinac-Dabic, D., Cruz, M., Alkasab, T. K., Hanisch, R. J., Nilsen, W. J., Burleson, J., Lyman, K., & Kandarpa, K. (2019). A road map for translational research on artificial intelligence in medical imaging: From the 2018 National Institutes of Health/RSNA/ACR/The Academy Workshop. *Journal of the American College of Radiology*, 16(9), 1179-1189. <https://doi.org/10.1016/j.jacr.2019.04.011>
- Barberis, A., & Mezard, M. (2024). RENOIR: Repeated sampling methods for machine learning validation in biology and medicine. *Scientific Reports*, 14, 5116. <https://doi.org/10.1038/s41598-024-51381-4>
- Bencevic, M., Habijan, M., Galic, I., Babin, D., & Pižurica, A. (2024). Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer Methods and Programs in Biomedicine*, 245, Article 108044. <https://doi.org/10.1016/j.cmpb.2024.108044>

- Cobo, M., Fontecha, D. C., Silva, W., & Iglesias, L. L. (2025). Preprocessing guidelines and FAIR4prep: Establishing best practices for clinical informatics workflows. *Scientific Data*, 12(1), 732. <https://doi.org/10.1038/s41597-023-02641-x>
- Daye, D., Wiggins, W. F., Lungren, M. P., & Alkasab, T. K. (2022). Implementation of clinical artificial intelligence in radiology: A roadmap for governance, maintenance, and monitoring. *Radiology*, 305(3), 555-563. <https://doi.org/10.1148/radiol.213123>
- Drukker, K., Chen, W., Gichoya, J., Grusauskas, N., Kalpathy-Cramer, J., Koyejo, S., Myers, K., Sá, R. C., Sahiner, B., Whitney, H., Zhang, Z., & Giger, M. (2023). Toward fairness in artificial intelligence for medical image analysis: Identification and mitigation of potential biases in the roadmap from data collection to model deployment. *Journal of Medical Imaging*, 10(6), 061104. <https://doi.org/10.1117/1.JMI.10.6.061104>
- Erickson, B. J., Khosravi, B., Vahdati, S., & Zhang, K. (2024). FDA review of radiologic AI algorithms: Process, cybersecurity, and clinical data curation challenges. *Radiology*, 310(2), e230242. <https://doi.org/10.1148/ryai.230242>
- Fahad, N., Sadib, R. J., Sajib, R. H., Morol, M. K., Nandi, D., & Liew, T. H. (2026). Responsible artificial intelligence in medical imaging: A systematic review. *Frontiers in Digital Health*, 8, 1884692. <https://doi.org/10.3389/fdgth.2026.1884692>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92. <https://doi.org/10.1145/3458723>
- Gichoya, J. W., Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L. C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S. C., Kuo, P. C., Lungren, M. P., Palmer, L. J., Price, B. J., Purkayastha, S., Pyrras, A. T., Oakden-Rayner, L., Okechukwu, C., Seyyed-Kalantari, L., & Zhang, H. (2022). AI recognition of patient race in medical imaging: A modelling study. *The Lancet Digital Health*, 4(6), e406-e414. [https://doi.org/10.1016/S2589-7500\(22\)00063-2](https://doi.org/10.1016/S2589-7500(22)00063-2)
- Jeon, K., Park, W. Y., Schmidt, T. S., & You, S. C. (2026). Clinical data standardization for distributed research: Standardizing medical imaging data with Radiology and Medical Imaging Common Data Models (R-CDM and MI-CDM). *Investigative Radiology*, 61(Suppl), S74-S83. <https://doi.org/10.1097/RLI.0000000000001155>
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in ML-based science. *Patterns*, 4(9), Article 100804. <https://doi.org/10.1016/j.patter.2023.100804>
- Keane, P. A., & Topol, E. J. (2018). With an eye to AI and autonomous diagnosis. *npj Digital Medicine*, 1, 40. [As cited in NCBI Bookshelf resource NBK619320]
- Khor, S., Haupt, E. C., Hahn, E. E., Lyons, L. J., Shankaran, V., & Bansal, A. (2023). Racial and ethnic bias in risk prediction models for colorectal cancer recurrence when race and ethnicity are omitted as predictors. *JAMA Network Open*, 6(6), e2318495. <https://doi.org/10.1001/jamanetworkopen.2023.18495>
- Kidwai-Khan, F., Wang, R., Skanderson, M., Brandt, C. A., Fodeh, S., & Womack, J. A. (2024). A roadmap to artificial intelligence (AI): Methods for designing and building AI ready data to promote fairness. *Journal of Biomedical Informatics*, 154, 104654. <https://doi.org/10.1016/j.jbi.2024.104654>
- Kinahan, P. (2025). Centralized imaging collaborations and the landscape of medical imaging databases for artificial intelligence readiness. In Gilbert W. Beebe Symposium: AI and ML applications in radiation therapy, medical diagnostics, and radiation occupational health and safety (pp. 45-56). National Academies Press. <https://doi.org/10.17226/29200>
- Kondylakis, H., Osuala, R., Puig-Bosch, X., Lazrak, N., Diaz, O., Kushibar, K., Chouvarda, I., Charalambous, S., Starmans, M. P., Colantonio, S., Tachos, N., Joshi, S., Woodruff, H. C., Salahuddin, Z., Tsakou, G., Aussó, S., Alberich, L. C., Papanikolaou, N., Lambin, P., Marias, K., Tsiknakis, M., Fotiadis, D. I., Martí-Bonmatí, L., & Lekadir, K. (2025). A review of methods for trustworthy AI in medical imaging: The FUTURE-AI guidelines. *IEEE Journal of Biomedical and Health Informatics*, 29(2), 2017-2026. <https://doi.org/10.1109/JBHI.2025.3614546>
- Langlotz, C. P., Allen, B., Erickson, B. J., Kalpathy-Cramer, J., Bigelow, K., Cook, T. S., Flanders, A. E., Lungren, M. P., Mendelson, D. S., Rudie, J. D., & Wang, G. (2019). A roadmap for foundational research on artificial intelligence in medical imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology*, 291(3), 781-791. <https://doi.org/10.1148/radiol.2019190613>
- Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences of the United States of America*, 117(23), 12592-12594. <https://doi.org/10.1073/pnas.1919012117>
- Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., Weicken, E., Asselbergs, F. W., Prior, F., Collins, G. S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J. A., Kushibar, K., Riklund, K., Marias, K., & FUTURE-AI Consortium. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388, e081554. <https://doi.org/10.1136/bmj-2024-081554>
- Lones, M. A. (2024). Avoiding common machine learning pitfalls. *Patterns*, 5(10), Article 101046. <https://doi.org/10.1016/j.patter.2024.101046>
- Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., & Kopp-Schneider, A. (2018). Is the winner really the best? A critical analysis of common research practice in biomedical image analysis competitions. *Nature Communications*, 9, 5217. <https://doi.org/10.1038/s41467-018-07619-7>
- Maier-Hein, L., Reinke, A., Kozubek, M., & BIAS Initiative. (2020). Good scientific practice in biomedical image analysis challenges: BIAS statement. *Medical Image Analysis*, 66, 101796. <https://doi.org/10.1016/j.media.2020.101796>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, J., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229. <https://doi.org/10.1145/3287560.3287596>
- Nooraie, R. Y., Shelton, R. C., Lee, M., Brotzman, L. E., & Gichoya, J. W. (2021). Applying an implementation science lens and equity focus to the scale-up of clinical artificial intelligence in medical imaging. *PET Clinics*, 16(4), 643-653. <https://doi.org/10.1016/j.cpet.2021.07.002>

- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453. <https://doi.org/10.1126/science.aax2342>
- Ogut, E. (2025). Artificial intelligence in clinical medicine: Challenges across diagnostic imaging, clinical decision support, surgery, pathology, and drug discovery. *Clinical Practice*, 15(9), 169. <https://doi.org/10.3390/clinpract15090169>
- Puyol-Antón, E., Ruijsink, B., Harana, J. M., Piechnik, S. K., Neubauer, S., Petersen, S. E., Razavi, R., & King, A. P. (2022). Fairness in cardiac magnetic resonance imaging: Assessing sex and racial bias in deep learning-based segmentation. *Frontiers in Cardiovascular Medicine*, 9, Article 859310. <https://doi.org/10.3389/fcvm.2022.859310>
- Rädsch, T., Reinke, A., Weru, V., Kopp-Schneider, A., & Maier-Hein, L. (2023). Labeling instructions matter: Systematic evaluation of annotation noise and labeling standards in biomedical image analysis. *Nature Machine Intelligence*, 5, 254-267. <https://doi.org/10.1038/s42256-023-00625-5>
- Rafique, S., Chaudhary, K., Haidar, S. H., Rashid, U., & Usman, S. (2026). Diagnostic AI across the life sciences (2015-2025): A PRISMA-scoping review and bibliometric synthesis of external validity, calibration, fairness, and reproducibility. *Haya: The Saudi Journal of Life Sciences*, 11(2), 122-141. <https://doi.org/10.36348/sjls.2026.v11i02.002>
- Raposo, H. (2025). Artificial intelligence-enabled medical imaging for early disease detection: Methodological advances, clinical validation, and translational challenges. *SSRN Electronic Journal*, 5331997. <https://doi.org/10.2139/ssrn.5331997>
- Rubak, M. W., Brejnebo, M. H., Rose, M. H., Gudbergensen, H., Chaudhari, A., Troelsen, A., Moller, A., Nybing, J. U., & Boesen, M. (2025). Federated learning and robustness scenarios under label-noise, image-noise, and faulty-client conditions in medical imaging. *Journal of Clinical Medicine*, 14(3), 132-146.
- Sen, C. K., & DeMazumder, D. (2025). Getting started on artificial intelligence in health care and clinical research: Includes rigor checklist for authors and reviewers. *Journal of Wound Care*, 34(2), 114-127. <https://doi.org/10.12968/jowc.2025.34.2.114>
- Sulaimanov, U., Sanlier, N., Moniri, A., Demir, B., Serikkanov, Y., Bayramoglu, A. R., Al-Jebur, M. S., Uslu, I., Ozturk, O., Nizzola, M., Ötles, E., Ammanuel, S. G., Keles, A., Erginoglu, U., & Baskaya, M. K. (2026). Are AI neuroimaging models ready for clinical use? A systematic methodological review. *Journal of Clinical Medicine*, 15, 03441. <https://doi.org/10.3390/jcm1503441>
- Szabo, L., Lekadir, K., Mosteiro, P., Mitchell, M., Goisauf, M., & Sardanelli, F. (2022). Developing a "trustworthy AI system" in medical imaging: Practical guidelines and questions. *Frontiers in Cardiovascular Medicine*, 9, Article 1016032. <https://doi.org/10.3389/fcvm.2022.1016032>
- Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., Moy, L., Park, S. H., Kahn, C. E., & Panel, C. U. (2025). Methodological validation and reporting standards in AI-based diagnostic neuroradiological and neurosurgical research: A systematic review. *Journal of Clinical Medicine*, 15(9), 3441. <https://doi.org/10.3390/jcm15093441>
- Young, A. T., Pfau, J., Kelothe, V., Chande, D., Farzandipour, M., Al shareef, H. N., & Wei, M. L. (2020). Selective prediction and deep learning for robust medical image classification: A dermoscopy clinical reader study. *npj Digital Medicine*, 3, 4. <https://doi.org/10.1038/s41746-020-00380-6>
- Yousefi Nooraie, R., Lyons, P. G., Baumann, A. A., & Saboury, B. (2025). Equitable implementation of artificial intelligence in medical imaging: What can be learned from implementation science? *PET Clinics*, 20(2), Article 682.