International Journal of Public Health and Nursing (IJPHN)

ISSN: 3070-202X (ONLINE)

VOLUME 2 ISSUE 1 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

International Journal of Public Health and Nursing (IJPHN)

Volume 2 Issue 1, Year 2026
ISSN: 3070-202X (Online)
DOI: https://doi.org/10.54536/ijphn.v2i1.6988
https://journals.e-palli.com/home/index.php/ijphn

Explainable and Bias-Aware AI Models for Clinical Decision Support in U.S. Healthcare Systems

Sudip Sharma[1*], Kevin Leziga Giami[2], Uche Stanley Chukwuemeka[3]

Article Information

Received: December 08, 2025

Accepted: February 09, 2026

Published: May 22, 2026

Keywords

Algorithmic Bias and Fairness, Bias Mitigation and Equity Auditing, Clinical Decision Support Systems (CDSS), Explainable Artificial Intelligence (XAI), U.S. Healthcare Deployment and Governance

ABSTRACT

Although AI-enabled clinical decision support systems (CDSS) are becoming more prevalent in U.S. healthcare, inequities and opaque models pose a threat to patient safety and clinician trust. This scoping review mapped evidence on explainable and bias-aware clinical AI systems to inform their equitable deployment. Following the PRISMA-ScR guidelines and a PCC framework, we searched MEDLINE/PubMed and Embase for English peer-reviewed studies published between 2015 and 2025. The full texts of eligible studies were charted across eight domains, including data modality, CDSS use case, model approach, documentation bias, explanation technique, implications for trust and outcomes and mitigation and governance actions. Of the 464 records identified, 18 studies met the inclusion criteria. The evidence spanned imaging (predominantly chest radiography), EHR-based risk prediction and emergency department operational and safety models. The evidence was largely retrospective in nature. Explainability was most defensible when used as a safety audit to support reviewable rationales and detect shortcut learning, typically via feature attribution, perturbation tests, or visualisation. Bias reflected demographic signal leakage, temporal or label leakage, proxy targets, subgroup error disparities, and site confounding factors, which can inflate apparent performance. The most common mitigation strategies combined reweighting or fairness-aware selection, data augmentation, and setting-specific recalibration, but post-deployment monitoring was inconsistently reported. A trustworthy CDSS requires explicit equity objectives, multi-site evaluation, standardised documentation, and continuous surveillance for drift and emergent disparities.

INTRODUCTION

Clinical decision support is entering a new phase as U.S. healthcare systems retain and generate large volumes of digital data from monitoring devices, imaging, operational workflows and electronic health records. These 'big data' environments have enabled machine learning approaches that can match the performance of experts in certain classification and prediction tasks, while providing scalable support for care coordination, prognosis, and diagnosis (Beam & Kohane, 2018). Advances in the acquisition of digitised data and computing power have widened the range of clinical tasks that AI can support (Yu *et al.*, 2018). However, data only matter when they are translated into actionable inferences, so model evaluation and choice are as important as data volume (Obermeyer & Emanuel, 2016). In the AI era, decision support systems are designed for interactive use by clinicians, making human–system interaction foundational to effectiveness and safety (Shortliffe & Sepúlveda, 2018). This vision aligns with 'high-performance medicine', in which AI augments health systems and clinicians rather than replacing clinical judgement (Topol, 2019).

Despite accelerating innovation, achieving a lasting clinical impact remains uncommon because successful translation requires more than just high headline accuracy. Successful systems must be able to integrate into sociotechnical pathways, be evaluated using measures that reflect clinical utility and downstream outcomes and generalise across sites (Kelly *et al.*, 2019). Experience with CDSS shows that the balance of harms and benefits also depends on implementation choices and the behavioural effects of alerts within real workflows (Sutton *et al.*, 2020). The feasibility of autonomous diagnostic testing for diabetic retinopathy in primary care has been demonstrated (Abràmoff *et al.*, 2018). Similarly, a real-world integration effort for sepsis decision support highlights the need for infrastructure, end-user training and stakeholder alignment in order to make model outputs usable at the bedside (Sendak *et al.*, 2020).

In this context, the legitimacy of AI-enabled CDSS depends on explainability and awareness of bias. A lack of explainability gives rise to legal, clinical and ethical concerns, as clinicians are required to interpret uncertainty, judge plausibility and retain accountability for patient-facing decisions (Amann *et al.*, 2020). Furthermore, explainability is task-dependent; clinicians require actionable and contextual explanations that align with the stakes of the decision (Tonekaboni *et al.*, 2019). Similarly, bias awareness is important because deployment and design choices can perpetuate inequity, even when overall performance appears strong (Yu *et al.*, 2018). Looking ahead, responsible progress requires technical development be aligned with professional and governance norms and that workflows be integrated (Briganti & Le Moine, 2020).

Against this backdrop, a focused synthesis is needed to

[1] Morgan State University, Department of Computer Science, Baltimore, Maryland, USA
[2] Modinfra Technologies Ltd, England
[3] Prairie View A&M University, USA
[*] Corresponding author's e-mail: sudipsharmarr@gmail.com

clarify the meanings of 'explainable' and 'bias-aware' in applied clinical decision support systems (CDSS) research, and how these concepts are operationalized in deployed or deployment-ready systems. As the field encompasses a variety of workflows and modalities, clinicians and implementers require a clear overview of the types of explanations being employed, the sources of bias being measured, and the methods used to evaluate mitigation in practice (Shortliffe & Sepúlveda, 2018). Evidence-based research can also help to set realistic expectations regarding where additional clinical evaluation or monitoring is required prior to scaling up and what machine learning can deliver (Beam & Kohane, 2018). This scoping review therefore surveys applied studies to describe current approaches, highlight gaps in evaluation and reporting, and identify priorities for more accountable adoption in US healthcare systems nationwide.

Scope of Review

This scoping review maps peer-reviewed evidence from 2015 to 2025 on explainable, bias-aware AI used for clinical decision support within the US healthcare system or US-relevant datasets. The review covers imaging (e.g. ultrasound, chest radiographs, and MRI scans), electronic health records (EHRs) and claims data, as well as emergency department operations. The focus is on how models are interpreted (saliency/attribution, SHAP-style feature contributions, visualisation and causal probing) and how bias is detected and managed. Documented bias pathways include temporal leakage, proxy learning of protected attributes, site confounding and subgroup error disparities by race/ethnicity, sex and socioeconomic position. Mitigation strategies and governance practices are synthesised to protect patients, calibrate clinician trust, and guide the accountable deployment of AI across diverse populations, workflows and hospitals.

Aim and Specific Objectives of Review

The aim is to map and critically appraise the implementation and evaluation of explainability and bias-aware methods in AI-enabled CDSS, and to derive practical implications for the equitable and transparent deployment of these methods.

Specific Objectives

i. Characterize the evidence base by summarising CDSS use-cases, clinical settings, and data modalities (imaging, EHR/ED operations), and the AI approaches used.
ii. Synthesize explainability practice by cataloguing XAI techniques reported and the purposes they serve (e.g., auditing shortcuts, supporting clinician interpretation, surfacing feature reliance).
iii. Interrogate equity and governance by classifying documented bias sources/impacts and summarising mitigation or governance strategies tested (e.g., reweighing, augmentation, fairness-aware/interpretable model selection) and their reported implications for trust and outcomes.

MATERIALS AND METHODS

Design and Reporting Standards

This study employed a scoping review design based on the Population–Concept–Context (PCC) framework to map the scope, features, and practical governance implications of bias-aware and explainable AI models utilized for clinical decision-making support within U.S. healthcare systems. Reporting was guided by the PRISMA extension for scoping reviews (PRISMA-ScR), which includes selection decisions in a flow diagram and transparent documentation of screening. No protocol registration was undertaken. In line with the purpose of scoping reviews (mapping evidence rather than estimating pooled effects), no formal risk-of-bias appraisal was performed. Methodological safeguards included a multi-step screening process; explicit, PCC-based eligibility criteria; and standardized data charting aligned to the study's predefined extraction domains.

Eligibility Criteria

The eligibility criteria were defined a priori using PCC to ensure the inclusion of empirical, real-world clinical AI/CDSS studies that explicitly addressed interpretability (XAI) and documented bias (or equity-related performance differences) that are relevant to the deployment of these systems in U.S. healthcare. Studies were required to report a component on explainability (e.g. saliency/visual explanations, SHAP/LIME-type feature attribution, counterfactuals, model simplification or other approaches to interpretability that are clinician-facing) and document sources of bias and/or impacts on subgroups (e.g. disparities in error by sex, age, race/ethnicity, insurance status or deprivation) with or without mitigation. Table 1 provides a breakdown of the inclusion and exclusion criteria.

Table 1: Eligibility Criteria according to the PCC Framework

Item	Inclusion Criteria	Exclusion Criteria
Population	Humans receiving care in clinical settings or represented in clinical datasets relevant to U.S. healthcare delivery (e.g., EHR cohorts; radiology/clinical registries)	Animal/in vitro studies; purely synthetic populations without clinical linkage

Concept (Inervention)	AI/ML-based CDSS models for diagnosis, prognosis, risk stratification, operations, triage, or treatment planning that include explicit XAI and documented bias (bias sources and/or subgroup impacts), with or without mitigation	AI models without any explainability component; XAI-only methodological papers without clinical CDSS application; fairness-only papers without interpretability; tools not intended for decision support
Context	U.S. healthcare systems, U.S.-derived clinical datasets, or U.S. multi-site/registry contexts (including models evaluated for U.S. clinical deployment)	Non-U.S.-only contexts with no relevance to U.S. systems or datasets
Study Designs	Peer-reviewed empirical studies (e.g., cohort/ retrospective analyses, cross-sectional evaluations, model development + validation with clinical data, implementation/effectiveness evaluations)	Narrative reviews, editorials, letters, protocols without results, conference abstracts without full text
Outcomes	Any of: predictive performance + calibration; explanation utility/inspection outputs; subgroup errors/fairness metrics; bias mechanisms (e.g., proxy learning, leakage, site confounding); mitigation effects; implications for clinician trust/workflow/patient outcomes	Studies reporting only technical metrics without clinical relevance; non-clinical benchmark-only evaluations
Publication Type	Peer-reviewed journal articles	Preprints, theses, reports, guidelines, books/ chapters, grey literature
Language	English	Non-English
Timeframe	2010-2025	Studies published outside this timeframe

Information Sources

Searches were restricted to peer-reviewed literature indexed in two bibliographic databases that were selected for their comprehensive coverage of health services research, clinical medicine and biomedical informatics: MEDLINE/PubMed and Embase (Ovid). No grey literature sources were searched. All retrieved records were exported to a reference manager for de-duplication, followed by screening of titles and abstracts, and then full texts, using the eligibility criteria. Disagreements were resolved through consensus.

Search Strategy

Database-specific strategies were employed that combined controlled free-text terms and vocabulary reflecting the following PCC elements: healthcare settings and clinical decision support; interpretability/ explainability; AI/ML models; and bias/fairness/equity. Priority was given to sensitivity in order to capture heterogeneous modalities (EHR/tabular, multimodal, imaging, and text) and varied explanation approaches (model simplification, saliency, counterfactuals and feature attribution). Filters were applied to limit the results to English-language studies relevant to humans and conducted between 2015 and 2025. Table 2 summarises the search strings used.

Data Extraction and Synthesis

Data charting used a standardized extraction template aligned to eight domains: modality and settings; approach and model; governance testing and mitigation; outcomes and implications for trust; included studies; clinical CDSS use cases; data; XAI reporting and bias documentation. The charting form was refined and piloted to ensure the consistent capture of explainability methods (e.g. interpretability audits, visualization/saliency probing and SHAP-based feature attribution) and bias mechanisms

Table 2: Search String

Database	Search string
MEDLINE/PubMed	((“Clinical Decision Support Systems”[Mesh] OR “decision support”[tiab] OR CDSS[tiab] OR “clinical decision support”[tiab]) AND (“Artificial Intelligence”[Mesh] OR “Machine Learning”[Mesh] OR “Deep Learning”[Mesh] OR artificial intelligence[tiab] OR machine learning[tiab] OR deep learning[tiab]) AND (explainab*[tiab] OR interpretab*[tiab] OR “explainable AI”[tiab] OR XAI[tiab] OR SHAP[tiab] OR LIME[tiab] OR saliency[tiab] OR “saliency map*”[tiab] OR “counterfactual”[tiab] OR “model simplification”[tiab]) AND (bias[tiab] OR fairness[tiab] OR “algorithmic bias”[tiab] OR disparity[tiab] OR inequity[tiab] OR “healthcare disparities”[Mesh])) AND (english[lang]) AND (“2015/01/01”[dp] : “2025/12/31”[dp])

Embase (Ovid)	('clinical decision support system'/exp OR 'clinical decision support':ti,ab,kw OR CDSS:ti,ab,kw OR 'decision support':ti,ab,kw) AND ('artificial intelligence'/exp OR 'machine learning'/exp OR 'deep learning'/exp OR 'neural network'/exp OR artificial intelligence:ti,ab,kw OR machine learning:ti,ab,kw OR deep learning:ti,ab,kw) AND (explainab*:ti,ab,kw OR interpretab*:ti,ab,kw OR 'explainable ai':ti,ab,kw OR XAI:ti,ab,kw OR SHAP:ti,ab,kw OR LIME:ti,ab,kw OR saliency:ti,ab,kw OR 'counterfactual':ti,ab,kw) AND (bias:ti,ab,kw OR fairness:ti,ab,kw OR 'algorithmic bias':ti,ab,kw OR disparity:ti,ab,kw OR inequity:ti,ab,kw) AND [english]/lim AND [2015-2025]/py AND [article]/lim

(e.g. subgroup error disparities, leakage risks, proxy learning, and leakage risks), as reflected in the results synthesis. The synthesis was mapping-oriented and descriptive rather than effect-estimating. A numerical summary characterized the studies according to clinical task category (e.g. operational safety/CDSS, EHR risk prediction, diagnostic imaging), explanation family, mitigation approach, documented bias type and model class. A narrative synthesis organized the findings into three higher-order themes: (1) explainability as a mechanism for clinical alignment and safety auditing, (2) bias pathways with implications for disparities and misleading performance claims, and (3) patterns of mitigation and governance supporting trustworthy deployment.

RESULT AND DISCUSSIONS

Screening and Selection.

The database search yielded 464 records. After removing 240 duplicates, 224 unique records remained to be screened against the eligibility criteria based on their titles and abstracts. At this stage, 139 records were excluded due to incorrect titles/abstracts, leaving 85 articles for full-text assessment. A further 67 articles were excluded at this stage (Reason 1: narrative/systematic reviews, editorials, letters, protocols without results and conference abstracts without full text = 34; Reason 2: AI models without an explainability component, XAI-only methodological papers without a clinical CDSS application, fairness-only papers without interpretability or tools not intended for decision support = 13; Reason 3: studies reporting only technical metrics without clinical relevance or non-clinical benchmark-only evaluations = 20). Finally, 18 studies that m*et al*l the criteria were synthesised (see Fig. 1).

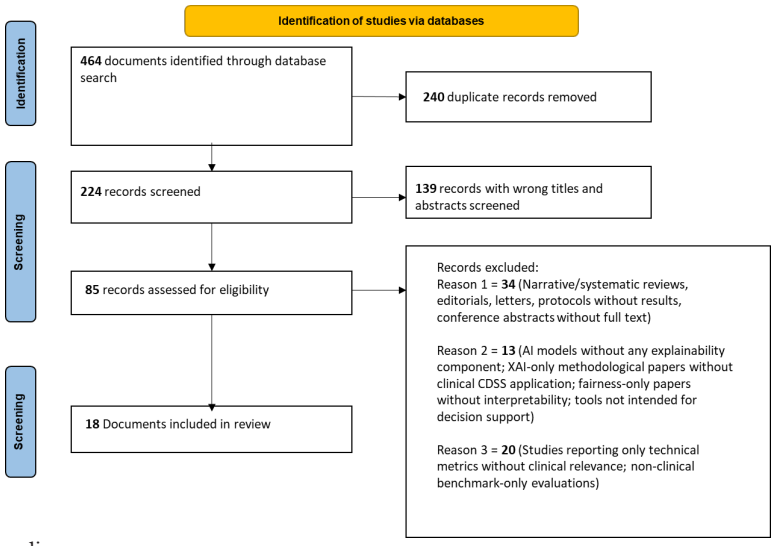

Figure 1: Prisma flow diagram

Summary of Study Characteristics

Eighteen peer-reviewed empirical studies (2018–2025) spanning imaging, electronic health records (EHR) and operational clinical decision support systems (CDSS) in U.S.-relevant care settings were included. The imaging studies primarily analysed multi-site chest radiographs, frequently employing MIMIC-CXR and external validation on CheXpert, with one study incorporating neuroimaging from ADNI. The EHR/registry studies covered topics such as ICU mortality, perioperative pain, wait-time prediction and ED admission, surgical readmission and agitation forecasting, using health system cohorts or NSQIP. The models ranged from deep neural networks to gradient-boosted trees (CatBoost and XGBoost) and simpler baselines such as random forests and logistic regression. XAI approaches included pertubation tests, SHAP and visualisation/saliency probing. Bias assessment reported fairness metrics, demographic proxy signals, temporal label leakage, site effects and subgroup error rates. Most studies were retrospective, with only a few describing live workflow integration, user training or post-deployment monitoring plans.

Thematic Analysis

Theme 1: Explainability as a clinical alignment and safety-audit mechanism

Explanability was most useful across the included studies when it helped people test whether a model's reasoning was consistent with clinical judgement and local workflow. DeGrave *et al.* (2021) demonstrate that chest X-ray models may appear robust yet rely on shortcuts, underscoring the importance of using explanations as a safety check prior to implementation. Zech *et al.* (2018) demonstrate that performance can deteriorate when a model is transferred between hospitals, meaning that explainability must be paired with testing across sites. Adleberg *et al.* (2022) add that models can pick up sensitive patient information from radiographs, meaning explanations could support audits to detect hidden proxy risks. Gichoya *et al.* (2022) reinforce this concern by demonstrating that models can infer race from images in ways that are not easily detectable by clinicians. McLeod *et al.* (2025) highlight that AI and humans may focus on different visual cues, meaning explanations should reveal mismatches rather than just 'looking convincing'. In ICU prediction, interpretability has helped examine whether demographic patterns are shaping mortality risk in unfair or unclear ways (Meng *et al.*, 2022). In quality and operational settings, explanations were used to demonstrate the factors influencing predictions in a format that could be reviewed and discussed by teams. Wang *et al.* (2025) used explanation outputs to inform the review of emergency department delay predictions. Zander *et al.* (2025) used explanations to summarise why readmission risk increased after surgery, providing support for improvement and feedback. In general, the concept of explainability proved most efficacious as a pragmatic instrument for verifying plausibility, identifying potential shortcuts, and promoting accountable clinical utilization.

Theme 2: Bias pathways that drive disparities and misleading "success" in real-world CDSS

Bias has been shown to arise at multiple points in the clinical AI pipeline, with direct implications for patient harm and disparities. Underdiagnosis bias is particularly consequential because it shifts the burden of error towards underserved groups, delaying care. Seyyed-Kalantari *et al.* (2021) demonstrate higher false-negative rates in intersectional and underserved subgroups for chest radiograph tasks, highlighting that bias can manifest as missed diagnoses as well as score differences. System and site artifacts also create bias-like failures by making models brittle across hospitals. Zech *et al.* (2018) demonstrate that pneumonia models generalize inconsistently across institutions, which is consistent with confounding factors being site-specific patterns rather than a robust disease signal. Another type of bias is workflow-invalid 'leakage', which inflates apparent performance while undermining clinical usefulness. Ramadan *et al.* (2025) demonstrate that predicting the same admission using ICD codes can result in temporal label leakage, producing seemingly strong results that cannot support real-time decision-making. When it comes to bias, standard metrics can hide the ways in which it interacts with local context at scale. Wang *et al.* (2024) demonstrate that disparities in readmission errors tied to income and race vary across hospitals, implying that fairness is not a fixed property of a model, but rather of a model in a specific setting. Bias is also evident in operational decision support, where disparities impact access and throughput. Davoudi *et al.* (2023) report differences in fairness across race, deprivation, age and insurance in postoperative pain prediction, indicating that inequity can be embedded even in routine risk tools. Finally, the absence of explicit fairness reporting is itself a governance gap in high-volume CDSS, which is a problem that must be addressed. Wong *et al.* (2025) exemplify this risk by presenting a multi-site ED model in which operational value is emphasized while equity evaluation is not given sufficient consideration.

Theme 3: Mitigation and governance approaches that make equity and transparency deployable

The mitigation strategies identified in the evidence base can be categorized as follows: (i) fairness-aware model design; (ii) data/representation interventions; and (iii) setting-specific recalibration and monitoring. Fairness-aware design involves treating equity as a selectable constraint rather than a post-hoc score, as exemplified by certain frameworks. Liu *et al.* (2024) operationalize this concept through fairness-aware, interpretable model selection that targets intersectional bias whilst maintaining utility. This approach is more aligned with the way in which health systems make deployment trade-offs in practice. Data-centric mitigation is most clearly demonstrated in imaging, where shortcut learning can be reduced by altering how models learn representations. Wang *et al.* (2024) demonstrate that augmentation can reduce demographic detectability and disparity in errors. This indicates that bias mitigation can be achieved at an earlier stage rather than only by adjusting thresholds. When disparities are driven by subgroup sample imbalance, mitigation can be framed as improving the performance of the worst-performing subgroup without eroding the performance of the dominant subgroup. A quasi-Pareto approach was used by Yao *et al.* (2024) to reduce subgroup AUC gaps in thyroid ultrasound prediction, with the performance-equity negotiation that must be overseen by governance being made explicit. In operational risk modelling, mitigation must account for local context and maintenance over time. Wang *et al.* (2024) emphasize the importance of recalibration and retraining alongside routine bias audits, implying that governance must be continuous rather than a "one-time validation". Some studies demonstrate how monitoring can be incorporated into CDSS design alongside interpretability and fairness checks. Li *et al.* (2025) merge SHAP-supported interpretability with fairness evaluation in SLE flare prediction, demonstrating a review-ready

Table 3: Data Extraction Table (8 themes/columns) – Included Studies (n=18)

Included study (year)	Clinical CDSS use case	Data & modality/ setting	AI model/ approach	XAI / interpretability reported	Bias documented (source + impact)	Mitigation / governance tested	Key implications (trust/outcomes)
Adleberg *et al*., 2022	Demographic inference from chest radiographs (risk signal for downstream CDSS bias)	55,174 radiographs (MIMIC-CXR) + external validation (CheXpert; multihospital urban system)	Deep learning classifiers	Visualization techniques to show anatomical regions of interest	Hidden demographic signal (gender/age/ethnicity/ insurance) learnable from CXR; risk of unintended proxy use	Uses demographic-prediction + visualization as auditing tools; emphasizes training diversity checks	Imaging can encode protected attributes; governance should require bias auditing before deployment
Davoudi *et al*., 2023	Acute postoperative pain risk prediction (perioperative CDSS)	EHR (UF Health/Shands); 14,263 orthopedic surgery patients	CatBoost	Focus on fairness metrics; interpretable feature contributions reported (study-level)	Bias across age, race, area deprivation index (ADI), insurance; not observed for sex/language/health literacy (as reported)	Reweighing protected attributes; evaluates multiple fairness metrics	High AUROC can mask subgroup harm; fairness testing should be routine pre-deployment
DeGrave *et al*., 2021	COVID-19 detection from chest radiographs	Multi-hospital CXR training/evaluation design	Deep learning	XAI analyses used to reveal reliance on shortcuts/confounders	Shortcut learning/ confounding from data collection and site effects; external testing alone may not detect it	Governance lesson: treat XAI as a pre-deployment requirement; improve dataset construction and evaluation	Apparent accuracy can be unsafe; XAI supports appropriate clinician trust calibration
Gichoya *et al*., 2022	Race recognition risk in medical imaging (bias pathway in imaging CDSS)	Private + public datasets across imaging modalities (CXR/CT/mammogram/ hand, etc.)	Deep learning	Stratification + image perturbation/cropping/ noise analyses; saliency-style probing	Models infer race with high accuracy; downstream risk of inequitable decisions via proxy learning	No definitive mitigation; highlights the difficulty of preventing protected-attribute inference	Strong governance needed: models may encode race even when clinicians cannot perceive it
Liu *et al*., 2024 (FAIM)	Hospital admission prediction with fairness + interpretability	MIMIC-IV-ED + SGH-ED	Fairness-aware interpretable modelling framework (FAIM)	Interpretable model family + interface to select models with improved fairness-utility balance	Targets intersectional bias (race × sex)	FAIM model selection improves fairness compared with common mitigation baselines while preserving performance	Practical pathway for clinician-in-the-loop fairness decisions; supports accountable deployment

Wang H.E. *et al.*, 2024	Seyyed-Kalantari *et al.*, 2021	Saxena *et al.*, 2025	Ramadan *et al.*, 2025	Meng *et al.*, 2022	McLeod *et al.*, 2025
30-day readmission prediction (quality/ operations CDSS)	Chest X-ray pathology classification	Healthcare access prediction in underserved communities	Same-admission outcome prediction validity (mortality) and leakage risks	In-hospital mortality prediction (ICU CDSS)	Imaging diagnosis: human vs AI visual bias (decision quality)
10.6 million discharges (Maryland & Florida; 2016–2019)	Multiple large CXR datasets; multi-source evaluation	Community-level access and health data (details limited in excerpt)	MIMIC-IV; 180,640 patients; evaluation of feature timing realism	MIMIC-IV ICU data	Empirical framing in diagnostic imaging (study-specific datasets not summarized in excerpt)
LACE, HOSPITAL, CMS model (as-is vs retrained/recalibrated)	State-of-the-art vision models	Fair and interpretable deep learning approach	LR / RF / XGBoost using ICD codes (audit focus)	Deep learning mortality models	AI + human perception comparison
Fairness metrics with interpretation of operational meaning; model comparisons	Model-behaviour diagnostics; interpretable subgroup error characterization	Interpretable modelling (SHAP-style reporting noted in record-level description)	Variable/feature importance used to audit non-actionable predictors	Interpretability methods (feature importance/ attribution) to interrogate reliance on attributes	Discusses interpretability needs; evaluates distinct bias patterns affecting humans vs AI
Racial + income-related disparities in error rates; heterogeneity by hospital context	Underdiagnosis bias: higher false negatives in underserved and intersectional groups	Socioeconomic / demographic bias in access predictions targeted	Label leakage / temporal bias: ICD codes not available at prediction time inflate apparent performance	Representation bias + unequal reliance on demographic attributes; disparate errors possible across groups	Distinct visual biases in humans and AI; risk of misaligned learning signals
Recalibration / retraining; recommends routine bias audits tailored to site context	Primarily bias characterization; governance warning for clinical deployment	Fairness-aware learning + optimization; equity-oriented access planning	Governance: restrict to time-available variables; leakage checks as standard validation	Proposes a combined workflow: interpretability + fairness metrics to detect disparities	Emphasizes interpretability + bias controls as system requirements
Fairness results must be clinically interpreted; it reduces risk of inequitable quality interventions	Direct patient-harm pathway via delayed diagnosis; fairness must be a safety metric	Supports policy-linked equity planning; illustrates fairness + interpretability for resource allocation	Prevents illusory CDSS performance that would fail in workflow; protects clinician trust	Trust requires transparency about drivers; fairness audits complement interpretability	Misaligned visual reasoning can lead to inconsistent decisions; it validates the need for XAI+bias checks

Zander *et al.*, 2025	Li *et al.*, 2025 (FLAME)	Yao *et al.*, 2024	Wong *et al.*, 2025	Wang *et al.*, 2024 (Drop the shortcuts)	Wang *et al.*, 2025 (ED wait)
30-day readmission after open ventral hernia repair	SLE flare prediction (3-month risk CDSS)	Thyroid nodule risk prediction (ultrasound)	Predict agitation events in emergency department (safety CDSS)	Radiology finding detection + Alzheimer's detection (fairness improvement)	Prolonged ED wait time prediction (operational CDSS)
NSQIP 2018–2021; 59,482 patients	28,433 SLE patients; EHR + 675 contextual SDoH variables	Large thyroid ultrasound dataset with external validation	3,048,780 ED visits across 9 sites (2015–2022)	MIMIC-CXR 194,359; CheXpert external 134,300; ADNI 1,195 MRIs	173,856 ED visits; ESI level 3 cohort
XGBoost	XGBoost + logistic regression; causal structure learning	QP-Net: multi-task + domain adaptation	Predictive model using 50 predictors	Deep learning with augmentation	XGBoost
SHAP (groupwise influential features)	SHAP + causal structure learning for interpretation	Not primarily XAI; fairness-focused performance balancing	Top predictors reported (feature salience)	Model probing/visualization to identify shortcut reliance; demographic detectability tests	SHAP global + local explanations
Fairness assessed across gender/ethnicity/race; minimal performance differences reported	Fairness assessed (equality of opportunity/FNR across race/ethnicity); no significant bias detected (as reported)	Subgroup AUC disparities driven by sample-size imbalance	Fairness/bias not explicitly detailed in the document excerpt	Demographic shortcut learning (race/age/sex inference) linked to disparity in errors	Disparities in errors across sex, race/ethnicity, insurance
Fairness evaluation + SHAP reporting	Fairness monitoring + interpretable predictor reporting	Quasi-Pareto improvement reduces subgroup disparity while preserving dominant-group performance	Not specified	Image augmentation reduced demographic detectability and reduced error disparities; external validation used	Bias characterization with governance recommendation to pair performance checks with fairness monitoring
Template for trustworthy surgical CDSS: performance + fairness + interpretable drivers	Demonstrates combined interpretability + fairness assessment for equity-sensitive disease management	Fairness strategy for imaging CDSS where subgroup underperformance is clinically consequential	Multi-site performance supports operational value; lack of explicit fairness reporting remains a governance gap	Concrete mitigation with external validation; improves generalizability and equity, supporting clinician confidence	Explainability supports workflow adoption; fairness prevents operational inequity affecting patients

Zech *et al.*, 2018	Pneumonia screening (CXR CAD) across hospitals	158,323 CXRs across NIH, Mount Sinai, Indiana University	CNN imaging classifiers	Not explicit XAI; demonstrates confounding via cross-site generalization tests	Dataset shift / site confounding: models learn hospital system & prevalence differences; weaker external generalization	Prevalence balancing experiments; recommends multi-site evaluation beyond internal testing	External validity is a safety requirement; shortcuts can distort clinical decisions and erode trust

strategy for ongoing disease administration.

Discussion

Summary of Key Findings

Explanability was most useful across studies when treated as a safety audit. Methods such as SHAP or visual probes helped to establish whether predictions supported rationales that could be reviewed and relied on clinically plausible drivers. Documented bias clustered into four recurring mechanisms: subgroup error disparities, which often affected underserved groups; demographic proxy signals embedded in the data; site or system confounding, which undermined generalization; and workflow-invalid label leakage, which inflated apparent performance. Evidence for mitigation favoured practical, system-adaptable strategies, such as data augmentation to reduce shortcut learning, recalibration/retaining to address local context and reweighting or fairness-aware model selection to improve subgroup performance. However, many studies lacked consistent reporting of human–AI interaction, equity trade-offs, and monitoring. This indicates that trustworthy CDSSs require standardized documentation, continuous post-deployment surveillance and explicit fairness objectives.

Comparison with global literature

The findings suggest that explainability should be used to check the alignment of decision-making processes and real workflows. This emphasis is consistent with clinician-centred work, which argues that explanations are useful only when they are contextual and support practical reasoning at the point of care (Tonekaboni *et al.*, 2019). They also coincide with the legal and ethical perspective that explanations must be actionable and arguable, and not just translate technical information (Wachter *et al.*, 2017). Nevertheless, our findings also underscore the shortcomings of post-hoc explanations as an alternative to safety, indicating the concern that contemporary XAI approaches could be a false sense of security in a clinical environment (Ghassemi *et al.*, 2021). This contends the hypothesis that interpretable models can be the best to use in making high-stakes choices when possible (Rudin, 2019).

The synthetic evidence of disparities created by proxy targets and structural inequity in data, and the synthesis of bias mechanism in this study agrees with that. The most common case is the application of cost prediction as a proxy of illness, which has created a lot of racial discrimination because spending is based on unequal access to care (Obermeyer *et al.*, 2019). This analogy underscores the significance of making distributive justice an explicit element of assessment and goals since various equitable options emphasise disparate forms of adversity (Rajkomar *et al.*, 2018). It further supports the conceptualization of algorithmic bias as a patient safety concern, which needs to be systematically assessed beyond the headline accuracy (Challen *et al.*, 2019). This makes stewardship critical because there will be no quantification of fairness; it will require governance judgment of values and context (Panch *et al.*, 2019). Also, critiques of race correction warn that integrating race in clinical algorithms might continue to make biased assumptions in daily care practises (Vyas *et al.*, 2020).

The findings of the study are in agreement with the global consensus regarding mitigation and governance in the development of a trustworthy CDSS. The development of a trustworthy CDSS is a process that follows a lifecycle management rather than a single validation process. Translation roadmaps reiterate the significance of working with non-disciplinary stakeholders and unintended post-deployment and post-monitoring drift (Wiens *et al.*, 2019). Equally, risk management models demand organizations to repeat controls and operationalize transparency, fairness and accountability (Tabassi, 2023). The principle of priority in human factors and viability of the real world at an early-stage of evaluation is also aligned with our focus on deployment realism (Vasey *et al.*, 2022). Reporting standards complement this strategy because they require comprehensive explanations of the error analysis and human-AI interaction (Liu *et al.*, 2020). Documentation tools enhance accountability because they ensure that the subgroup performance and intended use become visible to the decision-makers (Mitchell *et al.*, 2019). This is supplemented by dataset documentation that reveals provenance and possible biases that may result in inequalities in the future (Gebru *et al.*, 2021).

Implications for Policy and Practice

Policy: It is proposed that explainable and bias-conscious AI-enabled clinical decision support systems (CDSS) be implemented as commissioned equity and safety standards throughout the US healthcare delivery. This would entail compulsory pre-deployment reporting

on calibration assessment, subgroup performance and documentation of intended use, human-AI interaction and contraindications. These systems would have regular bias audits, which would extend beyond the question of overall accuracy into equity-relevant distributions of errors and impacts of resource allocation. The audits would also entail the explicit choice of fairness objectives that are in line with distributive justice principles. The protection of equity should be in the form of strict governance or banning of proxy targets (e.g. cost-based outcomes) where they entrench structural inequities. Race-variable scrutiny and the non-indiscriminate practice of race correction also should be required. Standardization of provision of explanations should be done by use of clinician-facing model information artifacts (e.g. fact labels) which define uncertainty, failure modes and monitoring responsibilities. Also, documentation of datasets and models should be compulsory in order to make sure that there is transparency in terms of representativeness, missing data, and subgroup performance. There should be model cards and datasheets that may serve as governance tools that are auditable. Evaluation reporting standards must be established, such as trial report and protocol reporting requirements that would reflect workflow integration, error analysis and impact of the user. Lastly, an organized system-level AI risk management process must be established with a requirement to conduct regular post-deployment monitoring of drift and emerging inequities. This framework must comprise accountability among vendors and health systems, escalation channels, and protection mechanisms to ensure enhancements in efficiency do not widen disparities.

Practice: Implementing an 'audit-design-deploy-monitor' model is essential. This model has minimum requirements for:(i) prospective workflow mapping; (ii) selecting explanations that are appropriate for the modality and user task (e.g. feature attribution summaries, saliency localization, or counterfactual action guidance); (iii) fairness testing across relevant subgroups before any live use. Deploying clinician-centred explanation interfaces that prioritize concise, actionable outputs supporting safe overrides and contestability over persuasive narratives is crucial. To ensure appropriate clinical reliance, teams should undergo training in interpreting limitations, subgroup performance differences and uncertainty and use structured communication tools (e.g. model fact labels). Conversely, it is advisable to refrain from deploying black-box models in high-stakes pathways where comparable performance can be achieved with interpretable alternatives, and to avoid explanations that can be exploited or are unstable. Deployment must be customized to the clinical context, including documentation practices, operational constraints and local population characteristics. Periodic threshold reviews and recalibration are required where equity trade-offs are explicit. Outcomes should be monitored using clinical response patterns, alert burden, subgroup error profiles, model performance and calibration. Drift detection and incident reporting should feed into iterative improvement cycles. Training multidisciplinary teams (including nursing leadership) in bias-aware triage, escalation protocols and explanation review is recommended to strengthen adoption, safety and equity.

4.4 Limitations of Review

This scoping review only used two databases and limited its inclusion criteria to English, peer-reviewed articles, meaning that relevant studies may have been overlooked. No formal risk-of-bias appraisal or meta-analysis was conducted. Heterogeneity in CDSS tasks, XAI reporting and fairness metrics limited comparability. Most of the evidence was retrospective, with limited post-deployment evaluation.

CONCLUSION

This scoping review mapped 18 empirical studies relevant to the U.S. on bias-aware and explainable AI for clinical decision support. Across modalities, explainability was most effective when employed as an accountability and a safety tool, guiding appriopriate clinical reliance and supporting the auditing of model reasoning, rather than as a mere transparency layer. Bias was repeatedly linked to workflow realities and structural data, including learning based on demographic proxies, disparities in errors among subgroups, site confounding and temporal/label leakage, which can produce misleading 'success'. Although mitigation approaches showed promise, they were unevenly evaluated due to the limited standardisation of human-AI interaction reporting, fairness objectives and post-deployment monitoring. Future work should prioritise prospective, multi-site evaluations with explicit equity targets, continous surveillance for drift and emergent disparities and robust documentation, in order to enable the trustworthy deployment of AI in U.S. healthcare systems.

REFERENCES

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., Folk, J. C., & IDx-DR Study Group. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine, 1, Article 39*. https://doi.org/10.1038/s41746-018-0040-6

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 9525–9536). Curran Associates, Inc.

Adleberg, J., Wardeh, A., Doo, F. X., Marinelli, B., Cook, T. S., Mendelson, D. S., & Kagen, A. (2022). Predicting patient demographics from chest radiographs with deep learning. *Journal of the American College of Radiology, 19*(10), 1151–1161. https://doi.org/10.1016/j.jacr.2022.06.008

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q Consortium. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary

perspective. *BMC Medical Informatics and Decision Making, 20*(1), Article 310. https://doi.org/10.1186/s12911-020-01332-6

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA, 319*(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

Briganti, G., & Le Moine, O. (2020). Artificial intelligence in medicine: Today and tomorrow. *Frontiers in Medicine, 7, Article 27*. https://doi.org/10.3389/fmed.2020.00027

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety, 28*(3), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J.-C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group, & SPIRIT-AI and CONSORT-AI Consensus Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine, 26*(9), 1351–1363. https://doi.org/10.1038/s41591-020-1037-7

Davoudi, A., Sajdeya, R., Ison, R., Hagen, J., Rashidi, P., Price, C. C., & Tighe, P. J. (2023). Fairness in the prediction of acute postoperative pain using machine learning models. *Frontiers in Digital Health, 4,* Article 970281. https://doi.org/10.3389/fdgth.2022.970281

DeGrave, A. J., Janizek, J. D., & Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence, 3*, 610–619. https://doi.org/10.1038/s42256-021-00338-7

Fletcher, R. R., Nakeshimana, A., & Olubeko, O. (2021). Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Frontiers in Artificial Intelligence, 3,* Article 561802. https://doi.org/10.3389/frai.2020.561802

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

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., Pyrros, 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

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine, 17*(1), Article 195. https://doi.org/10.1186/s12916-019-1426-2

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

Li, Y., Yao, L., Lee, Y. A., Huang, Y., Merkel, P. A., Vina, E., Yeh, Y.-Y., Li, Y., Allen, J. M., Bian, J., & Guo, J. (2025). A fair machine learning model to predict flares of systemic lupus erythematosus. *JAMIA Open, 8*(4), ooaf072. https://doi.org/10.1093/jamiaopen/ooaf072

Liu, M., Ning, Y., Ke, Y., Shang, Y., Chakraborty, B., Ong, M. E. H., Vaughan, R., & Liu, N. (2024). FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare. *Patterns, 5*(10), 101059. https://doi.org/10.1016/j.patter.2024.101059

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & CONSORT-AI/SPIRIT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine, 26*(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 4768–4777). Curran Associates, Inc. https://doi.org/10.48550/arXiv.1705.07874

McLeod, G. A., Stanley, E. A. M., Rosenal, T., Li, M., Kirby, P. A., Karwowska, M., & Forkert, N. D. (2026). Distinct visual biases affect humans and artificial intelligence in medical imaging diagnoses. *npj Digital Medicine, 9, Article 62*. https://doi.org/10.1038/s41746-025-02226-5

Meng, C., Trinh, L., Xu, N., Enouen, J., & Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Scientific Reports, 12, Article 7166*. https://doi.org/10.1038/s41598-022-11012-2

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In

Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C., Dhiman, P., Beam, A. L., van Calster, B., Celi, L. A., Denaxas, S., Denniston, A. K., Ghassemi, M., Heinze, G., Kengne, A. P., Maier-Hein, L., Liu, X., Logullo, P., McCradden, M. D., Liu, N., ... van Smeden, M. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ, 388,* e082505. https://doi.org/10.1136/bmj-2024-082505

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. *The New England Journal of Medicine, 375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

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

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: Implications for health systems. *Journal of Global Health, 9*(2), 010318. https://doi.org/10.7189/jogh.09.020318

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine, 169*(12), 866–872. https://doi.org/10.7326/M18-1990

Ramadan, B., Liu, M., Burkhart, M. C., Parker, W. F., & Beaulieu-Jones, B. K. (2025). Diagnostic codes in AI prediction models and label leakage of same-admission clinical outcomes. *JAMA Network Open, 8*(12), e2550454. https://doi.org/10.1001/jamanetworkopen.2025.50454

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saxena, A., Sharma, S., Johari, P. K., Pandey, A., & Kumar, S. (2025). A fair and interpretable deep learning approach for healthcare access prediction in underserved communities. *Discover Artificial Intelligence, 5, Article 185*. https://doi.org/10.1007/s44163-025-00425-3

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *In Advances in Neural Information Processing Systems (Vol. 28, pp. 2503–2511).* Curran Associates, Inc.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 618–626). IEEE. https://doi.org/10.1109/ICCV.2017.74

Sendak, M. P., Gao, M., Brajer, N., & Balu, S. (2020). Presenting machine learning model information to clinical end users with model facts labels. *npj Digital Medicine, 3, Article 41*. https://doi.org/10.1038/s41746-020-0253-3

Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., Futoma, J., Gao, M., Nichols, M., Revoir, M., Yashar, F., Miller, C., Kester, K., Sandhu, S., Corey, K., Brajer, N., Tan, C., Lin, A., Brown, T., Engelbosch, S., Anstrom, K., Elish, M. C., ... O'Brien, C. (2020). Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Medical Informatics, 8*(7), e15182. https://doi.org/10.2196/15182

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine, 27,* 2176–2182. https://doi.org/10.1038/s41591-021-01595-0

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 180–186). Association for Computing Machinery. https://doi.org/10.1145/3375627.3375830

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA, 320*(21), 2199–2200. https://doi.org/10.1001/jama.2018.17163

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digital Medicine, 3, Article 17*. https://doi.org/10.1038/s41746-020-0221-y

Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In Proceedings of the Machine Learning for Healthcare Conference (pp. 359–380). PMLR. https://doi.org/10.48550/arXiv.1905.05134

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25,* 44–56. https://doi.org/10.1038/s41591-018-0300-7

Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., ... DECIDE-

AI expert group. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine, 28,* 924–933. https://doi.org/10.1038/s41591-022-01772-9

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in plain sight—Reconsidering the use of race correction in clinical algorithms. *The New England Journal of Medicine, 383*(9), 874–882. https://doi.org/10.1056/NEJMms2004740

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887. https://doi.org/10.2139/ssrn.3063289

Wang, H. E., Weiner, J. P., Saria, S., & Kharrazi, H. (2024). Evaluating algorithmic bias in 30-day hospital readmission models: Retrospective analysis. *Journal of Medical Internet Research, 26,* e47125. https://doi.org/10.2196/47125

Wang, H., Sambamoorthi, N., Hoot, N., Bryant, D., & Sambamoorthi, U. (2025). Evaluating fairness of machine learning prediction of prolonged wait times in emergency department with interpretable eXtreme gradient boosting. *PLOS Digital Health, 4*(3), e0000751. https://doi.org/10.1371/journal.pdig.0000751

Wang, R., Kuo, P. C., Chen, L. C., Seastedt, K. P., Gichoya, J. W., & Celi, L. A. (2024). Drop the shortcuts: Image augmentation improves fairness and decreases AI detection of race and other demographics from medical images. *EBioMedicine, 102,* 105047. https://doi.org/10.1016/j.ebiom.2024.105047

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine, 25*(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Wong, A. H., Sapre, A. V., Wang, K., Nath, B., Shah, M. N., Kumar, A., Faustino, E. V. S., Desai, S., Hu, J., Robinson, A. L., Meng, Y., Tong, J., Bernstein, S. L., Yonkers, K. A., Melnick, E. R., Dziura, J. D., & Taylor, R. A. (2025). Predicting agitation events in the emergency department through artificial intelligence. *JAMA Network Open, 8*(5), e258927. https://doi.org/10.1001/jamanetworkopen.2025.8927

Yao, S., Dai, F., Sun, P., Zhang, W., Qian, B., & Lu, H. (2024). Enhancing the fairness of AI prediction models by quasi-Pareto improvement among heterogeneous thyroid nodule population. *Nature Communications, 15, Article 1958*. https://doi.org/10.1038/s41467-024-44906-y

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering, 2,* 719–731. https://doi.org/10.1038/s41551-018-0305-z

Zander, T., Kendall, M. A., Wolansky, R. L., Grimsley, E. A., Parikh, R., Sujka, J., & Kuo, P. C. (2025). Fairness of machine learning readmission predictions following open ventral hernia repair. *Surgical Endoscopy, 39,* 5035–5045. https://doi.org/10.1007/s00464-025-11927-7

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine, 15*(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683