

REVIEW

Open access

Artificial Intelligence for Healthcare Equity Analytics from 2017 to 2024: A Systematic Review of Bias Detection, Fairness-Aware Prediction, Disparity Monitoring, and Algorithmic Accountability in Health Systems

Khaled Mahfouz^{1*}, Rania Abdelaziz¹, Sherif Adel², Dina Mostafa¹, Tamer Nabil², Reem Saad¹

Abstract

Artificial intelligence models are increasingly used to support healthcare decision-making, resource allocation, clinical triage, population health management, and quality improvement. Without explicit equity assessment, these tools may reproduce, obscure, or scale existing disparities across racial, ethnic, sex, gender, socioeconomic, and geographic groups. This systematic review examined artificial intelligence approaches for healthcare equity analytics from 2017 to 2024. The review focused on bias detection, fairness-aware prediction, disparity monitoring, and algorithmic accountability in health systems. A PRISMA 2020-compliant search was conducted across PubMed, Scopus, Web of Science, and IEEE Xplore for peer-reviewed English-language studies published between 2017 and 2024. Eligible records were screened by two reviewers, and findings were synthesised narratively by equity task, method type, clinical application, fairness metric, and implementation context. The evidence base showed increasing attention to bias detection and fairness-aware algorithm design, especially in clinical risk prediction, diagnostic imaging, and population health management. Disparity monitoring systems and accountability structures were less frequently evaluated in deployed health system environments and were commonly described as governance recommendations, audit frameworks, or pilot-stage approaches. The technical foundations for equitable artificial intelligence in healthcare are advancing, but translation into operational health system practice remains limited. The strongest evidence concerns bias measurement, whereas evidence that fairness interventions reduce real-world health disparities remains nascent.

Keywords Artificial intelligence, Healthcare equity, Algorithmic bias, Fairness-aware machine learning, Disparity monitoring, Algorithmic accountability

*Correspondence:

Khaled Mahfouz
khaled.mahfouz@gmail.com

¹ Department of Intelligent Healthcare Systems, Faculty of Medicine, Mansoura University, Mansoura, Egypt

² Department of Clinical Informatics and AI Engineering, Faculty of Engineering, Helwan University, Cairo, Egypt

Introduction

Healthcare artificial intelligence has become central to risk prediction, diagnostic decision support, care management, and operational prioritisation, creating both opportunities and risks for health equity. Several foundational papers warned that models trained on historical clinical data can encode inequitable access, biased measurement, structural racism, and unequal care pathways rather than objective need [1, 2]. The widely discussed population health management algorithm

evaluated by Obermeyer and colleagues illustrated how cost-based proxies can underestimate the health needs of Black patients, making algorithmic bias a health system governance problem rather than only a modelling flaw [2]. This evidence has driven increasing interest in equity analytics as a systematic approach to detecting, measuring, mitigating, and monitoring AI-induced disparities [3, 4].

In this review, bias refers to systematic differences in model performance, error distribution, calibration, data representation, or downstream allocation across socially meaningful groups. Fairness refers to the use of normative and statistical criteria, including equal opportunity, equalised odds, demographic parity, calibration, and subgroup validity, to assess whether model behaviour is acceptable across populations [5, 6]. Accountability refers to the institutional capacity to document, audit, explain, govern, and revise algorithmic systems when they contribute to inequitable outcomes [7, 8]. These concepts overlap but are not interchangeable, because a model may satisfy one fairness metric while still producing inequitable clinical consequences or weak organisational accountability [5, 9].

Health systems have begun to recognise that algorithmic equity cannot be assured at model development alone. Bias may arise during problem formulation, cohort construction, label definition, missingness patterns, feature selection, threshold choice, deployment, clinician uptake, or post-deployment drift [10, 11]. Several studies and commentaries emphasised that fairness evaluation must be continuous because model performance may vary across sites, time periods, clinical workflows, and patient subgroups [12, 13]. However, the literature remains uneven, with more work on retrospective bias detection than on prospective disparity monitoring, audit enforcement, or demonstrated reductions in health inequities [14, 15].

This systematic review therefore synthesises evidence on artificial intelligence for healthcare equity analytics across four linked domains: bias detection, fairness-aware prediction, disparity monitoring, and algorithmic accountability. The review follows PRISMA 2020 principles by defining a structured search strategy, eligibility criteria, screening process, data extraction framework, and narrative synthesis plan. It focuses on peer-reviewed studies and frameworks published from 2017 to 2024 that address healthcare AI fairness, bias mitigation, or equity governance. The objective is to map how the field has moved from identifying biased algorithms toward developing accountable, equity-oriented analytics infrastructures.

Materials and Methods

Search strategy

A systematic search was conducted across PubMed, Scopus, Web of Science, and IEEE Xplore for records published from 1 January 2017 to 31 December 2024. Search strings combined terms for artificial intelligence, machine learning, bias detection, fairness-aware prediction, disparity monitoring, social determinants of health, accountability, audit, and health equity, reflecting terminology used across clinical informatics, digital health, and biomedical ethics literature. The search strategy was informed by recurring concepts in prior health AI fairness literature, including algorithmic bias, race correction, model auditing, fairness metrics, and equity-oriented governance. Targeted supplementary searching was also used to identify highly cited studies and review articles in journals relevant to health informatics, digital medicine, ethics, and health services research.

Inclusion and exclusion criteria

Studies were eligible if they were peer-reviewed, written in English, published between 2017 and 2024, and addressed artificial intelligence or machine learning in relation to healthcare equity, bias, fairness, disparity monitoring, or accountability. Eligible publications included original empirical studies, modelling studies, systematic or scoping reviews, ethical analyses, and governance frameworks when they directly addressed healthcare AI equity analytics. Studies were excluded if they focused on general AI ethics without a healthcare application, described non-AI health disparities without algorithmic relevance, or lacked explicit discussion of fairness, bias, subgroup performance, equity governance, or

accountability. Papers addressing race correction in clinical algorithms were included when they were relevant to algorithmic equity, even if the tool was not a modern machine learning model, because these studies informed the broader accountability literature.

Screening and selection

After database and supplementary searching, 1,850 records were identified, and 410 duplicates were removed before screening. A total of 1,440 titles and abstracts were screened, 1,210 records were excluded, and 230 full-text reports were assessed for eligibility, resulting in 68 studies included in the narrative synthesis. Full-text exclusions were mainly due to absence of a healthcare AI application, lack of equity or fairness analysis, purely technical fairness methods without clinical relevance, or commentary without a clear accountability or governance contribution [3, 10]. The screening process was structured for presentation in a PRISMA 2020 flow diagram, with the final included evidence base grouped by bias detection, fairness-aware prediction, disparity monitoring, and algorithmic accountability.

Data extraction

Data extraction captured publication year, geography, clinical domain, model or framework type, healthcare setting, population subgroup variables, fairness metrics, mitigation strategy, deployment status, and accountability mechanisms. Particular attention was given to whether studies examined demographic, socioeconomic, geographic, sex, gender, race, ethnicity, or intersectional disparities in model performance or care allocation. For modelling studies, extracted fields included whether fairness was assessed through pre-processing, in-processing, post-processing, calibration analysis, subgroup error analysis, adversarial learning, or threshold adjustment. For governance papers, extraction focused on auditability, transparency, stakeholder responsibility, regulatory implications, model documentation, and institutional oversight.

Risk of bias assessment

Risk of bias assessment was adapted from prediction model appraisal principles and focused on sources of inequity in cohort definition, predictor availability, outcome labelling, missing data, validation design, and subgroup analysis. Studies were considered at higher risk of equity-related bias when they relied on single-site retrospective data, used proxy labels linked to unequal access, omitted relevant protected or social variables, or reported aggregate performance without subgroup evaluation. The review also considered whether fairness metric selection was justified, whether trade-offs were discussed, and whether mitigation approaches were evaluated beyond overall predictive accuracy. Ethical and governance publications were appraised narratively according to whether they specified responsible actors, audit processes, transparency requirements, and pathways for remediation.

Synthesis methods

Because included studies varied substantially in clinical domain, model type, fairness metric, population, and implementation maturity, a meta-analysis was not appropriate. A narrative synthesis was conducted by grouping evidence into four equity analytics tasks: bias detection, fairness-aware prediction, disparity monitoring, and algorithmic accountability. Within these groups, studies were further organised by methodological approach, including subgroup performance assessment, calibration analysis, pre-processing mitigation, constrained optimisation, adversarial learning, post-processing, model documentation, and governance recommendations. The synthesis emphasised patterns in reported methods and implementation barriers rather than producing pooled effect estimates or claims of clinical effectiveness.

Results and Discussion

Study selection

The PRISMA selection process identified 68 included studies from 1,850 records after duplicate removal, title and abstract screening, and full-text assessment. Most excluded records addressed AI in healthcare without explicit attention to equity, fairness, bias detection, disparity monitoring, or accountability. Among full-text exclusions, common reasons included absence of subgroup analysis, non-healthcare technical focus, insufficient description of algorithmic governance, or lack of relevance to health system equity analytics [3, 16].

Study characteristics

The included evidence base expanded notably after 2019, corresponding with increased attention to biased population health algorithms, ethical health AI governance, and subgroup evaluation in clinical prediction and imaging [2, 10]. Studies were concentrated in high-resource settings, especially the United States and other digitally mature health systems, with fewer examples from low- and middle-income contexts [11, 16]. Clinical domains included population health management, diagnostic imaging, risk prediction, triage, resource allocation, public health, and quality monitoring [17, 18]. Across domains, many studies were retrospective, and only a smaller subset addressed implementation, ongoing audit, or organisational responsibility [15, 16].

Figure 1 presents the evidence synthesis framework linking healthcare equity problem contexts, AI fairness analytics tasks, reviewed data structures, evaluation logic, governance findings, and practical future directions.

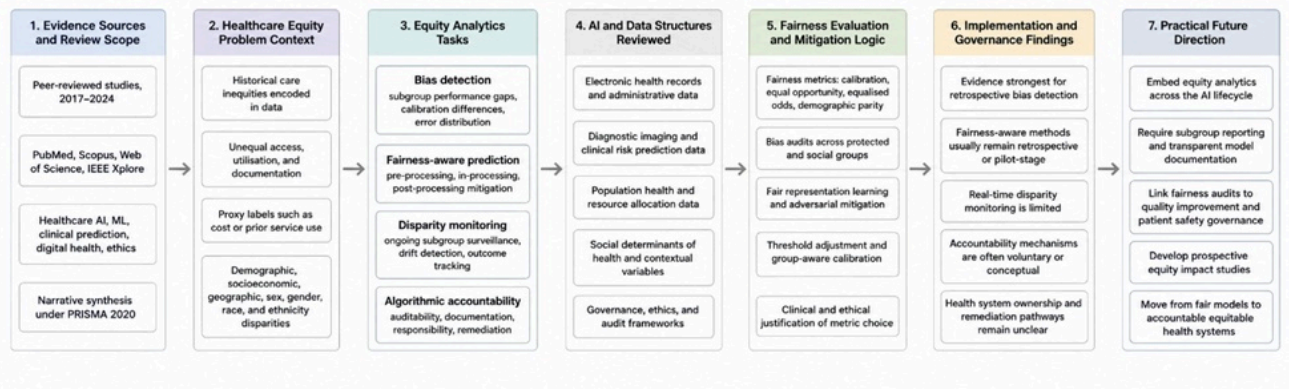


Figure 1. Evidence Synthesis Framework for Artificial Intelligence in Healthcare Equity Analytics

Bias detection approaches

Bias detection approaches most commonly examined whether model performance differed across race, ethnicity, sex, gender, age, insurance status, or other subgroup categories. Several studies reported subgroup differences in diagnostic performance, calibration, or allocation outcomes, particularly when models relied on proxies such as healthcare spending, historical utilisation, or diagnosis frequency [2, 19]. Imaging studies also demonstrated that models could encode demographic information in unexpected ways, raising concern that apparently neutral clinical data may contain latent signals associated with social identity [17, 18]. These findings support the need for systematic subgroup auditing before deployment and throughout the model life cycle [11, 20].

Common fairness metrics and audits

The most frequently discussed fairness metrics included calibration across groups, equal opportunity, equalised odds, demographic parity, predictive parity, subgroup sensitivity, subgroup specificity, and differences in false-positive or false-negative rates. Empirical and methodological studies emphasised that metric choice is not merely technical, because each metric reflects a different ethical and operational priority [5, 9]. Several authors noted that fairness criteria may conflict, especially when baseline outcome rates differ across groups due to structural inequities rather than biological

differences [1, 6]. As a result, fairness audits increasingly require clinical, statistical, ethical, and community-informed justification rather than automatic selection of a single metric [8, 15].

Table 1 provides an evidence taxonomy of the main AI approaches used for healthcare equity analytics, organised by analytical purpose, data structure, fairness logic, and operational relevance.

Table 1. Evidence Taxonomy of Artificial Intelligence Approaches for Healthcare Equity Analytics

Evidence domain	Main analytical purpose	Typical healthcare data structures	Common AI or ML approaches reviewed	Equity variables or subgroup dimensions	Fairness or bias assessment logic	Practical health system relevance
Bias detection in deployed or proposed healthcare AI	Identify whether model performance, allocation, calibration, or error patterns differ across socially meaningful groups	Electronic health records, administrative claims, imaging datasets, care management registries, risk scores, utilisation histories	Subgroup performance analysis, calibration assessment, error-rate comparison, proxy-label evaluation, retrospective audit	Race, ethnicity, sex, gender, age, insurance status, socioeconomic position, geographic location, language, disability, comorbidity burden	Compares sensitivity, specificity, false-positive rates, false-negative rates, calibration, risk score distribution, or allocation rates across groups	Helps health systems identify inequitable model behaviour before or after deployment and prevents hidden disparities from being scaled through automated decision support
Fairness-aware prediction using pre-processing methods	Reduce bias before model training by modifying data, labels, sampling, or representations	EHR tables, claims data, structured demographics, social determinants of health, historical utilisation data	Reweighting, resampling, missingness review, feature review, representation learning, cohort balancing	Protected attributes, under-represented subgroups, social risk variables, missing demographic fields	Evaluates whether modified data construction reduces subgroup performance gaps without obscuring clinically relevant risk	Supports more equitable model development by addressing bias at the level of cohort design, label construction, and data representativeness
Fairness-aware prediction using in-processing methods	Embed fairness objectives directly into model training	Clinical prediction datasets, imaging datasets, multi-site EHR data, structured and semi-structured patient records	Constrained optimisation, fairness regularisation, adversarial learning, multi-objective learning, fair representation models	Race, ethnicity, sex, age, site, insurance status, and other subgroup markers available during training or evaluation	Tests whether model optimisation can reduce unfair dependence on protected attributes while preserving clinically useful signal	Offers technically advanced fairness correction but often requires specialised expertise, clear ethical justification, and validation beyond retrospective datasets

Fairness-aware prediction using post-processing methods	Adjust model outputs after training to reduce inequitable decision thresholds or calibration differences	Risk scores, probability outputs, triage recommendations, care management flags, diagnostic decision outputs	Threshold adjustment, group-aware calibration, output review rules, decision-policy modification	Groups with unequal false-negative rates, false-positive rates, risk calibration, or allocation rates	Compares decision consequences before and after output adjustment across subgroups	Useful for modifying existing tools but raises governance questions when group-specific thresholds affect access to care resource allocation
Disparity monitoring systems	Track whether AI-supported care produces unequal outcomes, recommendations, or workflow consequences over time	Operational analytics data, quality metrics, patient safety data, care pathway data, model logs, outcome registries	Monitoring dashboards, drift detection, subgroup audit pipelines, alert thresholds, periodic recalibration review	Demographic, socioeconomic, geographic, clinical, and intersectional subgroups	Monitors changes in subgroup performance, clinical actions, access, quality indicators, and outcome disparities after deployment	Connects AI fairness to quality improvement, patient safety, health system accountability, and remains underdeveloped in the literature
Algorithmic accountability frameworks	Define how AI tools should be documented, audited, governed, and corrected when inequitable effects occur	Model documentation, audit reports, governance policies, validation records, institutional review files	Model cards, datasheets, algorithmic impact assessments, lifecycle governance, audit committees, transparency frameworks	Populations affected by high-stakes AI decisions, especially historically marginalised groups	Evaluates whether responsibility, transparency, audit timing, reporting obligations, and remediation pathways are specified	Moves fairness from technical evaluation to organisational responsibility, regulatory readiness, and institutional equity governance
Social determinants-informed equity modelling	Use contextual variables to understand, detect, or reduce inequitable prediction and allocation	Area deprivation indices, housing instability, income proxies, education, food insecurity, transportation, neighbourhood context, insurance status	Feature enrichment, confounding assessment, equity stratification, contextual risk modelling, fairness-aware feature governance	Socioeconomic position, neighbourhood deprivation, access barriers, social needs, structural vulnerability	Assesses whether social variables explain inequitable risk patterns or unintentionally encode disadvantage	Can improve contextual understanding of inequity but requires safeguards against stereotyping, surveillance, and inequitable resource restriction
Clinical task-specific	Apply equity analytics to	Readmission prediction, mortality	Risk prediction, classification,	Subgroups defined by demographic,	Measures whether AI-supported	Demonstrates where healthcare AI fairness

fairness applications	particular health system use cases	prediction, diagnostic imaging, triage, population health management, case management, public health analytics	imaging AI, patient representation learning, resource allocation algorithms	clinical, utilisation, and social characteristics	decisions distribute errors, benefits, burdens, or care access unequally	becomes operationally consequential for patients, clinicians, and administrators
-----------------------	------------------------------------	--	---	---	--	--

Fairness-aware prediction: pre-processing methods

Pre-processing methods included data reweighting, resampling, representation learning, missingness assessment, feature review, and attempts to improve cohort representativeness before model training. Several studies and reviews discussed the importance of examining how labels, predictors, and data availability reflect unequal access to care, because biased data construction can undermine fairness even before modelling begins [11, 21]. Social and demographic variables were sometimes used to diagnose inequities, but their inclusion as predictors required careful justification because they may either improve equity assessment or reinforce group-based stereotyping [13, 22]. Overall, pre-processing approaches were described as necessary but insufficient unless combined with transparent model evaluation and post-deployment monitoring [5, 10].

Fairness-aware prediction: in-processing methods

In-processing fairness methods included regularisation, constrained optimisation, adversarial learning, and multi-objective training designed to reduce performance disparities during model development. A subset of empirical work explored adversarial approaches and fair representation learning for clinical prediction, aiming to reduce dependence on protected attributes while preserving clinically useful signal [12, 23]. These approaches demonstrated the feasibility of embedding fairness objectives into model training, but studies often remained retrospective and used limited external validation [6, 12]. The review found that in-processing methods were technically promising but less commonly connected to operational governance decisions or prospective patient outcomes [14, 24].

Fairness-aware prediction: post-processing methods

Post-processing methods included threshold adjustment, group-specific calibration, output review, and decision policy modification after model development. Such approaches were attractive for health systems because they can sometimes be applied to existing predictive tools without full model redevelopment [1, 6]. However, ethical analyses cautioned that post-processing can create difficult trade-offs if thresholds differ across groups without transparent justification, stakeholder engagement, and clinical governance [8, 9]. The evidence suggested that post-processing should be framed as one component of an accountable intervention pathway rather than a standalone technical correction [7, 15].

Disparity monitoring systems

Disparity monitoring systems were less developed than bias detection studies and were often described conceptually rather than evaluated as mature operational platforms. Several governance-oriented papers argued that health systems should monitor AI outputs, clinical actions, and patient outcomes across subgroups after deployment, because pre-deployment validation cannot anticipate all workflow effects [7, 10]. Monitoring proposals commonly included subgroup performance dashboards, periodic recalibration reviews, alert thresholds for inequitable drift, and institutional committees responsible for remediation [14, 15]. Nevertheless, the review identified limited peer-reviewed evidence of real-time AI-enabled disparity monitoring embedded in routine clinical operations [22, 25].

Social determinants in equity-focused models

Social determinants of health were discussed as both essential context and a potential source of modelling risk. Studies and frameworks noted that housing instability, income, neighbourhood deprivation, education, transportation access, food insecurity, and insurance patterns may improve understanding of inequitable outcomes when used carefully [13, 21]. At the same time, these variables may encode structural disadvantage and produce harmful prioritisation if interpreted as individual risk rather than contextual inequity [5, 22]. The evidence therefore supported the use of social determinants primarily within transparent, equity-oriented modelling and audit frameworks rather than as unexamined predictors [1, 20].

Algorithmic accountability frameworks

Algorithmic accountability frameworks emphasised transparency, documentation, auditability, responsibility assignment, regulatory oversight, and remediation pathways. Several articles proposed that health systems should treat AI models as governed clinical technologies requiring lifecycle evaluation, rather than as static software tools [7, 10]. Ethical analyses highlighted the need for explicit accountability when algorithms shape triage, diagnosis, resource allocation, or care management decisions [8, 15]. However, many frameworks remained voluntary or principle-based, and relatively few specified enforceable mechanisms for reporting audit results, correcting harms, or compensating affected patients [3, 14].

Clinical tasks with fairness-aware implementations

Clinical tasks with fairness-aware implementations included population health risk stratification, diagnostic imaging, hospital risk prediction, public health modelling, and patient representation learning from electronic health records. The population health algorithm evaluated by Obermeyer and colleagues was influential because it showed how resource allocation models can disadvantage patients when cost is used as a proxy for need [2]. Diagnostic imaging studies demonstrated the importance of subgroup evaluation and revealed that AI systems may learn demographic signals that are not obvious to developers or clinicians [17, 18]. Other studies explored fair patient representations and clinical risk prediction use cases, but operational evidence remained limited [23, 24].

Barriers to real-world implementation

Common barriers to real-world implementation included fragmented data infrastructure, incomplete demographic data, inconsistent social determinants documentation, lack of external validation, legal uncertainty, and limited institutional ownership of algorithmic harms. Several studies noted that fairness methods developed in controlled retrospective datasets may not transfer cleanly to changing clinical workflows, multi-site environments, or settings with different population structures [10, 11]. Health systems also face practical challenges in deciding who should monitor AI equity, how often audits should occur, and what actions should follow when inequities are detected [7, 15]. These barriers help explain why many fairness-aware approaches remain methodological demonstrations rather than embedded health system practices [14, 25].

Impact of fairness interventions on health equity outcomes

The reviewed literature provided limited evidence that fairness interventions have reduced health disparities in patient outcomes at scale. Several studies demonstrated bias detection, subgroup performance differences, or technical mitigation, but fewer evaluated downstream effects on access, quality, treatment, morbidity, or patient experience [6, 12]. Commentaries and governance frameworks repeatedly called for prospective evaluation, ongoing monitoring, and accountability mechanisms to determine whether fairer models produce fairer care [3, 26]. Overall, the evidence suggested that healthcare AI fairness has advanced methodologically, but equity impact remains under-measured [13, 22].

Technical maturity vs. Operational reality

This review found a growing technical literature on bias detection and fairness-aware machine learning, but much weaker evidence that these methods are routinely deployed in live health systems. The distinction between technical fairness and operational equity was central across several studies, because a model may perform more similarly across groups without improving access, treatment, or outcomes [5, 9]. Health systems require governance, monitoring, workflow redesign, and accountability structures to translate model-level fairness into equitable care delivery [7, 15]. The field is therefore technically maturing but operationally incomplete [14, 25].

Bias detection precedes mitigation

More studies measured bias than corrected it, and more studies proposed mitigation than evaluated mitigation prospectively. Bias detection was often performed through subgroup performance comparison, calibration assessment, or analysis of proxy labels, as seen in studies of population health management and diagnostic imaging [2, 17]. Mitigation studies, including adversarial training and fair representation learning, were less common and were typically retrospective [12, 23]. This imbalance suggests that the first wave of healthcare AI equity analytics has been diagnostic rather than corrective [6, 11].

Multi-dimensional fairness tensions

The review found recurring tensions among fairness metrics, clinical goals, and ethical priorities. Equal opportunity, calibration, demographic parity, and equalised odds may imply different intervention thresholds or trade-offs, especially when structural inequities influence baseline disease prevalence, care access, and outcome measurement [1, 9]. Several authors argued that fairness decisions should be clinically and socially justified rather than treated as purely mathematical optimisation problems [5, 8]. This is particularly important in high-stakes settings such as triage, resource allocation, and diagnostic support, where different errors may have unequal consequences across groups [3, 15].

Disparity monitoring remains retrospective

Disparity monitoring remains largely retrospective, despite broad agreement that algorithmic equity should be evaluated continuously after deployment. Several papers proposed dashboards, subgroup audits, drift detection, recalibration, and lifecycle governance, but few described sustained real-time monitoring of AI-related disparities in operational care pathways [7, 10]. Retrospective audits are valuable for identifying inequity, but they may detect harms only after patients have already been affected [14, 22]. The evidence therefore supports a shift from episodic fairness evaluation toward embedded equity surveillance within health system analytics [25].

The accountability ecosystem

The accountability ecosystem for healthcare AI remains fragmented, with responsibility distributed across developers, clinicians, health system leaders, regulators, vendors, journal editors, and data governance bodies. Ethical and regulatory papers argued that AI tools should be auditable, transparent, and subject to clear oversight when they influence clinical or operational decisions [7, 8]. However, voluntary model documentation and internal review may be insufficient when algorithms contribute to unequal care [15]. The review indicates that accountability must include enforceable reporting, role clarity, patient-facing transparency, and remediation mechanisms [3, 14].

Social determinants as a double-edged sword

Social determinants of health can strengthen equity analytics by revealing contextual drivers of risk, access, and outcomes. At the same time, they can reproduce structural disadvantage if models treat deprivation as an individual attribute rather than a signal of system-level inequity [13, 21]. Several authors warned that uncritical inclusion of social and demographic variables may intensify surveillance, stereotyping, or discriminatory allocation if governance is weak [5, 22]. The implication is that social determinants should be used within explicit fairness frameworks, with transparency about why variables are included and how outputs will be acted upon [1, 20].

From fair models to equitable health systems

The central message of the reviewed literature is that fair models alone cannot produce equitable health systems. AI equity requires attention to data generation, clinical workflows, organisational incentives, implementation contexts, patient trust, and accountability for harm [10, 26]. Several studies showed that algorithmic bias may reflect broader inequities in healthcare utilisation, diagnostic access, and documentation, meaning that technical correction without system reform may be inadequate [2, 19]. Equity analytics should therefore be integrated into institutional quality improvement, patient safety, and health equity governance rather than treated as an isolated modelling task [15, 22].

Table 2 translates the review findings into a governance and implementation framework for evaluating, monitoring, and correcting inequitable AI effects in health systems.

Table 2. Governance, Implementation, and Future Research Framework for AI-Based Healthcare Equity Analytics

Framework dimension	Main problem identified in the review	Operational risk for health systems	Governance requirement	Implementation action	Future research priority
Problem formulation	AI tools may optimise convenience, cost, utilisation, or throughput rather than equitable clinical need	Models may reproduce historical underuse of care, unequal access, or biased proxies while appearing technically valid	Require explicit statement of equity objective, intended use, affected population, and clinical decision context	Review model purpose before development or procurement and reject proxy outcomes that conflict with equity goals	Study how alternative outcome definitions influence fairness and downstream care allocation
Data representativeness	Training datasets often under-represent marginalised populations or contain incomplete subgroup documentation	Aggregate performance may conceal poor performance for smaller or historically underserved groups	Require subgroup completeness assessment and documentation of missing demographic and social variables	Conduct representativeness audits before model validation and deployment	Develop robust methods for fairness assessment under small sample sizes and incomplete subgroup labels
Label construction	Labels such as cost, utilisation, diagnosis frequency, or prior referrals may reflect unequal access rather than true need	AI systems may allocate fewer resources to patients who historically received less care	Require clinical and equity review of labels before model training	Evaluate whether labels encode access, documentation, or payment bias	Compare fairness effects of alternative labels in clinical prediction and population health management

Fairness metric selection	Different fairness metrics can produce conflicting conclusions	Health systems may select favourable metrics while overlooking other forms of inequity	Require justification for metric choice based on clinical task, harm type, and stakeholder priorities	Report multiple subgroup metrics, including calibration and error distribution where relevant	Develop clinical guidance for choosing fairness metrics in high-stakes healthcare decisions
Mitigation strategy	Technical bias mitigation may not address workflow, access, or organisational drivers of inequity	Fairness-aware models may improve statistical parity without improving patient outcomes	Require mitigation plans that connect model behaviour to clinical action and equity goals	Pair pre-processing, in-processing, or post-processing methods with workflow review and human oversight	Conduct prospective studies testing whether mitigation changes care delivery and health outcomes
Disparity monitoring	Many studies evaluate fairness retrospectively but do not monitor AI after deployment	Model drift, changing populations, or workflow adaptation may create new inequities over time	Require post-deployment equity monitoring with predefined audit intervals and escalation rules	Build subgroup monitoring into quality improvement, patient safety, and operational analytics systems	Evaluate real-time disparity monitoring systems in multi-site health system settings
Accountability ownership	Responsibility for algorithmic harm is often unclear across developers, vendors, clinicians, and health systems	Inequitable AI effects may persist because no actor is accountable for correction	Define responsible parties for validation, monitoring, reporting, and remediation	Establish AI equity governance committees with authority to pause, revise, or retire models	Study governance models that successfully translate audit findings into corrective action
Transparency and documentation	Model assumptions, data limitations, subgroup performance, and intended use are often insufficiently documented	Clinicians and administrators may over-trust tools without understanding their limitations	Require model documentation, audit summaries, intended-use statements, and known equity limitations	Use model cards, datasheets, algorithmic impact assessments, and implementation checklists	Assess whether documentation improves clinician trust, patient understanding, and institutional accountability
Human oversight	Human review is often mentioned but not operationally defined	Clinicians may either ignore AI outputs or defer to biased recommendations without structured safeguards	Specify when and how humans can override, question, or escalate algorithmic recommendations	Create decision protocols that clarify AI's advisory role and preserve clinician responsibility	Evaluate how human-AI interaction affects disparities in real clinical workflows

Patient and community involvement	Fairness definitions are often selected by researchers without meaningful input from affected communities	AI governance may miss patient priorities such as dignity, transparency, contestability, and historical harm	Include patient, community, and equity representatives in governance and review processes	Create mechanisms for patient-facing explanation, appeal, and feedback where AI affects access or care	Develop participatory methods for defining acceptable fairness criteria and accountability standards
Regulatory readiness	Many accountability recommendations remain voluntary or principle-based	High-stakes AI tools may be deployed without enforceable equity safeguards	Require auditability, lifecycle monitoring, risk classification, and public or regulator-accessible reporting	Align institutional AI governance with emerging regulatory and ethical expectations	Study how regulation affects AI innovation, transparency, equity, and patient safety
Equity impact evaluation	Few studies demonstrate that fairness interventions reduce real-world disparities	Health systems may claim fairness improvement without evidence of patient-level benefit	Require equity impact endpoints in evaluations of high-stakes AI tools	Measure access, treatment, quality, safety, and outcome differences before and after implementation	Conduct prospective, implementation-focused equity impact studies across diverse populations

Limitations

Review limitations

This review was limited to peer-reviewed English-language publications from 2017 to 2024 and may have missed relevant grey literature, local health system audits, regulatory reports, and non-English studies. The evidence base was also shaped by terminology, because some studies relevant to health equity may not use terms such as fairness, bias, accountability, or algorithmic audit [16, 27]. The narrative synthesis approach was appropriate given heterogeneity but does not provide pooled quantitative estimates of fairness intervention effects [6, 28]. In addition, the review prioritised studies directly relevant to healthcare AI equity analytics, which may exclude broader technical fairness literature without clinical application [5].

Evidence base limitations

The included literature was limited by retrospective designs, single-site datasets, incomplete subgroup reporting, inconsistent fairness metrics, and limited prospective validation. Several studies identified bias or proposed mitigation strategies, but few evaluated whether interventions changed clinical decisions, care delivery, or patient outcomes in ways that reduced disparities [12, 17]. Fairness metric selection was sometimes insufficiently justified, creating the possibility that favourable metrics were emphasised while other inequities remained unexamined [6, 9]. These limitations indicate that the current evidence base is stronger for identifying algorithmic inequity than for demonstrating accountable, sustained equity improvement in health systems [14, 15].

Comparison with prior reviews

Prior reviews and ethical analyses have often focused on broad principles for responsible healthcare AI, including transparency, safety, fairness, accountability, and avoidance of harm. These contributions established that algorithmic systems can reproduce structural inequities when trained on biased data or deployed without subgroup evaluation [10, 26]. Other reviews focused more specifically on biased data, open science, fairness recommendations, or the ethical limits of technical fairness solutions [9, 11]. Compared with these works, the present review narrows attention to equity analytics as an applied health system function rather than treating fairness only as an abstract ethical principle [10, 29].

This review differs from prior work by synthesising four connected domains within one PRISMA-guided framework: bias detection, fairness-aware prediction, disparity monitoring, and algorithmic accountability. Some earlier publications examined algorithmic bias in specific contexts such as race correction, population health management, diagnostic imaging, or clinical risk prediction [2, 19, 17]. Others focused on fairness methods, global health implications, or AI governance without fully mapping how detection, mitigation, monitoring, and accountability interact across the model life cycle [7, 16]. By integrating these strands, this review shows that healthcare equity analytics requires both technical methods and institutional structures [15, 22].

The main contribution of this review is the mapping of the full equity analytics pipeline from identifying biased model behaviour to embedding accountability in health system governance. Bias detection studies demonstrate where inequities may arise, fairness-aware methods attempt to modify model behaviour, disparity monitoring tracks whether inequities 13 after deployment, and accountability frameworks determine who must respond when harm is detected [6, 18]. This pipeline-oriented perspective highlights a major gap in the literature: many studies stop at measurement or mitigation without showing how health systems should operationalise findings in real time [14, 25]. The review therefore extends prior work by framing healthcare AI fairness as a continuous governance process rather than a single model evaluation step [13, 21].

Research gaps

Prospective equity impact studies

A major research gap is the absence of prospective studies showing that fairness-aware models reduce disparities in real-world patient outcomes. Existing studies frequently identify subgroup differences, test fairness metrics, or demonstrate mitigation in retrospective datasets, but they rarely evaluate whether these interventions improve access, treatment, safety, quality, or patient experience across groups [6, 12]. This gap is important because technical fairness improvements may not translate into equitable care if clinical workflows, resource constraints, or organisational incentives remain unchanged [9, 10]. Future research should therefore measure equity impact after deployment and distinguish model-level fairness from health system-level disparity reduction [14, 22].

Intersectional fairness

Most healthcare AI fairness studies focus on single protected attributes, such as race, ethnicity, sex, age, or insurance status, rather than intersectional combinations of social position. This creates a risk that models may appear fair across broad categories while still performing poorly for smaller groups defined by overlapping disadvantage [17, 29]. Intersectional fairness is methodologically difficult because subgroup sizes may be small, labels may be incomplete, and statistical uncertainty may be high [5, 24]. Nevertheless, future studies should use transparent methods for reporting intersectional limitations rather than ignoring multi-axis inequities [11, 13].

Integration of community and patient perspectives

Community and patient perspectives remain under-integrated in the healthcare AI fairness literature. Many studies define fairness through statistical criteria selected by developers or researchers, but affected patients and communities may prioritise transparency, dignity, access, historical harm, or accountability differently [8, 22]. The absence of community input is especially concerning when algorithms shape care management, public health surveillance, or resource allocation

for populations already affected by structural inequity [2, 16]. Future equity analytics should therefore include participatory governance, patient-facing communication, and mechanisms for contesting or reviewing algorithmic decisions [15, 21].

Implications

For research practice

For research practice, this review implies that equity analysis should be embedded throughout the healthcare AI development pipeline rather than added after model evaluation. Study teams should examine how cohorts are defined, how outcomes are labelled, how missing data are distributed, how protected attributes are measured, and how fairness metrics align with clinical use [1, 5]. Model development should include transparent subgroup evaluation and external validation wherever feasible, especially for tools intended to operate across heterogeneous populations or institutions [10, 11]. This approach would make equity a core criterion of methodological quality rather than an optional ethical supplement [27, 28].

For clinical practice

For clinical practice, the findings indicate that health systems should not deploy AI tools without routine bias audits and transparent performance reporting across patient subgroups. Clinicians and operational leaders need clear information about where models perform less reliably, how thresholds are selected, and what safeguards exist when algorithmic recommendations may affect access or treatment [3, 7]. Diagnostic imaging and clinical prediction studies show that models can encode demographic signals and produce unequal performance even when protected attributes are not intentionally used [17, 18]. Routine audit processes should therefore be linked to clinical governance, quality improvement, and patient safety systems [15, 25].

For policy

For policy, this review supports clearer lines of accountability for AI-induced health disparities, analogous to accountability structures used for clinical negligence, patient safety events, and quality failures. Developers, vendors, health systems, and regulators should share responsibility for documenting model purpose, validating performance, monitoring inequity, and responding when harm is identified [8, 14]. Policy frameworks should also require attention to social determinants, data representativeness, and structural bias rather than treating fairness as a narrow statistical property [13, 21]. Without enforceable accountability, healthcare AI may continue to identify inequity without creating reliable obligations to correct it [19, 22].

Conclusion

Healthcare equity analytics is a rapidly advancing field within artificial intelligence and health systems research. Bias detection and fairness-aware methods are technically robust enough to reveal important subgroup differences in performance, calibration, and allocation. The strongest evidence currently supports systematic auditing of algorithms before and after implementation. These methods provide a necessary foundation for more equitable healthcare AI.

However, disparity monitoring systems and algorithmic accountability mechanisms remain under-developed. Many frameworks describe what responsible governance should include, but fewer studies show how these mechanisms operate in live health systems. Real-time equity surveillance, enforceable audit requirements, and clear remediation pathways are still uncommon. This limits the ability of health systems to respond quickly when algorithmic tools contribute to inequitable care.

The critical gap is the translation of fairness principles into measurable improvements in health equity at scale. A model that performs more fairly across groups does not automatically produce fairer access, treatment, or outcomes. Equity depends on how algorithms interact with clinical workflows, institutional priorities, resource constraints, and patient trust. Future work must therefore evaluate both technical fairness and real-world equity impact.

A concerted, cross-sector effort is needed to embed equity analytics into routine health system governance. Researchers, journals, clinicians, health systems, regulators, and communities all have roles in making healthcare AI transparent, auditable, and accountable. The next stage of the field should move beyond identifying bias toward preventing, monitoring, and correcting inequitable algorithmic effects. Only then can artificial intelligence contribute meaningfully to health equity rather than reproducing the disparities it is intended to address.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 14 Oct 2024

Revised: 31 Oct 2024

Accepted: 28 Dec 2024

Published online: 25 February 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866-72.
- Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-53.
- Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019;322(24):2377-8.
- Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health*. 2019;9(2):020318.

Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci.* 2021;4(1):123-44.

Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. *J Biomed Inform.* 2021;113:103621.

McCradden MD, Joshi S, Anderson JA, Mazwi M, Goldenberg A, Zlotnik Shaul R. Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *J Am Med Inform Assoc.* 2020;27(12):2024-7.

Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics.* 2020;46(3):205-11.

McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digit Health.* 2020;2(5):e221-e223.

Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med.* 2019;25(9):1337-40.

Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. *Patterns.* 2021;2(10):100347.

Yang J, Soltan AA, Eyre DW, Yang Y, Clifton DA. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digit Med.* 2023;6(1):55.

Ferryman K, Mackintosh M, Ghassemi M. Considering biased data as informative artifacts in AI-assisted health care. *N Engl J Med.* 2023;389(9):833-8.

Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: challenges and mitigation strategies. *NPJ Digit Med.* 2023;6(1):113.

Chin MH, Afsar-Manesh N, Bierman AS, Chang C, Colón-Rodríguez CJ, Dullabh P, et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. *JAMA Netw Open.* 2023;6(12):e2345050.

Fletcher RR, Nakeshimana A, Olubeko O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Front Artif Intell.* 2021;3:561802.

Seyyed-Kalantari L, Zhang H, McDermott MB, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med.* 2021;27(12):2176-82.

Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen LC, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health.* 2022;4(6):e406-e414.

Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *N Engl J Med.* 2020;383(9):874-82.

Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. *Nat Med.* 2020;26(1):16-7.

Kidwai-Khan F, Wang R, Skanderson M, Brandt CA, Fodeh S, Womack JA. A roadmap to artificial intelligence (AI): methods for designing and building AI ready data to promote fairness. *J Biomed Inform.* 2024;154:104654.

Dankwa-Mullan I. Health equity and ethical considerations in using artificial intelligence in public health and medicine. *Prev Chronic Dis.* 2024;21:E64.

Sivarajkumar S, Huang Y, Wang Y. Fair patient model: Mitigating bias in the patient representation learned from the electronic health records. *J Biomed Inform.* 2023;148:104544.

Silva PC, Sun H, Rodriguez-Brazzara P, Rezk M, Zhang X, Fliegenschmidt J, et al. Evaluating gender bias in ML-based clinical risk prediction models: A study on multiple use cases at different hospitals. *J Biomed Inform.* 2024;157:104692.

Carey S, Pang A, de Kamps M. Fairness in AI for healthcare. *Future Healthc J.* 2024;11(3):100177.

Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med.* 2018;378(11):981-3.

Huang Y, Guo J, Chen WH, Lin HY, Tang H, Wang F, et al. A scoping review of fair machine learning techniques when using real-world data. *J Biomed Inform.* 2024;151:104622.

Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn J Radiol.* 2024;42(1):3-15.

Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit Med.* 2020;3(1):81.