IIAR JOURNALS

Innovating Knowledge. Advancing Research.

IIAR Journal of Communication, Media, and Digital Society

Vol 1, No 1, 2026

Algorithmic Bias and Health Equity: A Conceptual Model of How Bias Enters, Propagates and Compounds in Medical AI Across the Data-to-Deployment Pipeline

IGWE Emmanuel Sunday[1], NWOGBOJI Emmanuel Usulor[2], BAMIGBOLA Paul Idowu[3]

[1,3] Veristzon Global Research Lab; [2] International Institute of AI and Research

ABSTRACT

Bias in medical artificial intelligence (AI) is most often traced to training data, and the usual remedy is to collect more or better data. This theoretical paper argues that bias can enter at every stage of the data-to-deployment pipeline, from the social context of data generation to monitoring after deployment. It also argues that bias accumulates and compounds across stages, so a correction at one stage can leave much of the disparity in place. An integrative review was combined with theory synthesis, and the population of interest was the peer-reviewed literature on algorithmic bias and health equity in medical AI. Seven stage-aligned search strings were run in PubMed/MEDLINE and supplemented by eight OpenAlex queries. Records were sampled purposively, by relevance in PubMed and by citation count in OpenAlex. Forty-nine journal articles published between 2023 and 2026 were included; each was verified against its Crossref publisher record and screened for retractions, corrections and expressions of concern. The evidence covered under-representation in dermatology and medical imaging, emerging work in genomics and clinical language models, evaluation studies and frameworks for fairness and equity. Articles were coded deductively to pipeline stages and bias mechanisms and summarised as frequencies. The evidence was concentrated in the United States (30 of 49 first authors), with one article from Africa. Propagation ($n = 35$) and entry ($n = 30$) were often addressed, but compounding was explicit in only 10 articles. The resulting Bias Cascade Model has eight stages, three entry routes, four propagation channels and three conditions under which bias compounds. The model is accompanied by an equity check for each stage and six testable propositions about how bias accumulates, compounds and can be interrupted. The African context is central: populations under-represented in training data are those on which deployed systems can be expected to perform worst. The paper closes with implications for health ministries, regulators and research institutions.

Keywords: algorithmic bias; health equity; medical AI; representation; data pipeline

Algorithmic Bias and Health Equity: A Conceptual Model of How Bias Enters, Propagates and Compounds in Medical AI Across the Data-to-Deployment Pipeline

1. INTRODUCTION

As artificial intelligence (AI) moves into clinical care, a well-documented risk follows: systems that perform well on average can perform worse for patients who already have the poorest health outcomes. A scoping review found that health algorithms often match or exceed human diagnostic precision yet can perpetuate racial and ethnic inequities (Hussain et al., 2025), and another found that the equity problems reported most often concern the characteristics of the data and the design of the model (Berdahl et al., 2023). Stratified evaluations have documented differences in how patients are diagnosed, treated and billed when algorithms are applied across demographic groups (Chen et al., 2023). Regulatory records offer little reassurance. Of 692 AI-enabled medical devices approved by the United States Food and Drug Administration, only 3.6% reported the race or ethnicity of the populations studied, and only 9.0% included a prospective post-market study (Muralidharan et al., 2024).

A common response has been to locate bias in training data and to improve the quantity and quality of data, one of the strategies most often recommended in the literature (Berdahl et al., 2023). This response is necessary but insufficient. Bias can also enter through how outcomes are measured and labelled, which variables represent patients, what a model is optimised for, where it is validated and how its outputs are used in care (Cross et al., 2024; Nazer et al., 2023; Tejani et al., 2024). Stage-specific typologies name these entry points, and recent lifecycle frameworks propose audits at each stage (Ahmad et al., 2026; Alluhaibi, 2026). What remains under-theorised is the dynamic between stages: how bias entering at one stage is inherited by the next, how it is concealed or amplified, and why a single late correction may leave the disparity in place (Soltan & Washington, 2024).

This gap matters most where the evidence base is thinnest. In a scoping review of AI and health-equity guidance, 70% of the literature originated in the United States and only 13% of articles defined health equity (Efird et al., 2026). African populations are under-represented in the genomic datasets used to build AI, so systems developed elsewhere may carry contextual bias when adopted in Africa (Gaye et al., 2026). Tools built in high-resource settings are also deployed unchanged in low-resource ones (Joseph, 2025). If bias compounds along the pipeline, these populations can be expected to bear the largest cumulative disparities.

This paper develops a conceptual model, the Bias Cascade Model (BCM), to explain how bias enters, propagates and compounds across eight stages of the medical AI pipeline. It addresses three questions. (1) At which stages, and through which routes, does bias enter medical AI? (2) Through which channels is bias carried forward or hidden along the pipeline? (3) Under what conditions do stage-level biases compound into larger inequities, and where are interventions most effective? The model draws on 49 verified peer-reviewed articles published in 2023–2026 and yields a stage-matched equity-check protocol and six testable propositions.

2. LITERATURE REVIEW

2.1 CONCEPTUAL REVIEW

Three concepts underpin the model. The first, *algorithmic bias*, denotes a systematic difference in the performance, calibration or consequences of an AI system across patient groups that is not justified by clinically relevant differences between them. Similar overall accuracy does not rule out harm when errors fall unequally. In a medication-adherence prediction model with comparable accuracy across groups, false positives were 1.50 times as common among patients from low-socioeconomic-status neighbourhoods and 2.05 times as common among Black patients (Mukhopadhyay et al., 2026). Bias may also be present when a model appears to perform well, if the reference standard used to judge it is itself distorted (Zanger-Tishler et al., 2024).

The second concept distinguishes fairness from health equity. *Fairness* refers to the formal criteria used to detect or constrain such differences. In

© IIARJournal. All rights reserved.

medical AI these are mostly group fairness metrics, which all 35 studies in one systematic review of mitigation strategies used (Gadhoumi et al., 2026). *Health equity*, by contrast, is a normative goal concerned with differences in health that are not inevitable but unnecessary and unfair (Whitehead, 1992). The two concepts are not interchangeable. Similar accuracy across groups can coexist with harmful error asymmetries whose consequences depend on the action a prediction triggers (Mukhopadhyay et al., 2026). The guidance literature also often leaves health equity undefined: only 13% of the 166 articles in one scoping review defined it (Efird et al., 2026).

The third concept is the *pipeline*. Some authors divide the medical AI lifecycle into stages from problem framing and data collection to development, validation and implementation (Nazer et al., 2023); in imaging AI, the stages run from data handling to deployment (Tejani et al., 2024). Both catalogue the biases characteristic of each stage, and reviews that map equity issues to mitigation strategies show many-to-many links between stages and remedies (Berdahl et al., 2023). In this paper, the "data-to-deployment pipeline" covers the full lifecycle, from the social context in which data are generated to monitoring after deployment. Its eight stages are treated as coupled: the output of each stage is the input to the next, so an early distortion becomes part of the "ground truth" that later stages take as given. Because most medical AI reuses data recorded for care, measurement and labelling (Stage 3) precede problem formulation (Stage 4), when an available variable is chosen as the target.

2.2 THEORETICAL REVIEW

Three theoretical lenses inform the model. The first, *fundamental cause theory*, explains why the association between socioeconomic status and mortality has persisted despite radical changes in the diseases and risk factors presumed to explain it. Socioeconomic status embodies resources, such as money, knowledge, prestige, power and beneficial social connections, that protect health whatever mechanisms are relevant at a given time (Phelan et al., 2010). Applied to AI, the theory predicts that bias originates upstream of the data, in the social processes that determine who receives care, what is recorded and how. Shipton and Vitale (2024) argue that AI in global health repeats a "politics of avoidance" seen in earlier technological interventions: it addresses proximate problems while leaving the fundamental determinants of inequity untouched. At the level of clinical practice, the history of the race coefficient in kidney-function estimation shows how assumptions rooted in racial essentialism can be encoded in clinical formulas and persist for decades (Mohottige et al., 2023).

The second lens is *sociotechnical systems theory*. It originated in a study of the longwall method of coal-getting that examined a work group in relation to both the social structure and the technological content of the work system (Trist & Bamforth, 1951); the theory analyses the technical and social components of a system jointly. In interviews with 17 leaders from 13 academic medical centres, only a minority reported considering inequity or bias in AI governance; most conceptualised these issues as characteristics of the tool, and fewer considered implications beyond the technology (Nong et al., 2024). Integrated frameworks argue that statistical diagnostics must be coupled with governance across the lifecycle (Ahmad et al., 2026). Members of the public likewise describe AI-related harms in sociotechnical terms, including epistemic injustice and exclusion from access (Townsend et al., 2025). Through this lens, deployment and clinician–AI interaction are stages where bias can enter and be amplified, not merely revealed.

The third lens, *cumulative disadvantage*, relates inequality to the passage of time: it is used to explain heterogeneity and inequality as advantages and disadvantages accumulate (Dannefer, 2003). Joseph (2025) describes how historical, representational and measurement biases are compounded by aggregation and by the unchanged transfer of tools to low-resource settings. Alluhaibi (2026) proposes a Lifecycle Equity Debt measure: a weighted aggregate of the amounts by which subgroup performance gaps exceed pre-declared tolerances at successive model checkpoints. The BCM integrates the three lenses. Fundamental cause theory explains entry, sociotechnical theory explains why propagation continues beyond the technical stages, and cumulative disadvantage explains compounding.

© IIARJournal. All rights reserved.

2.3 EMPIRICAL REVIEW

The evidence is reviewed by pipeline stage. At the data stages (Stages 1–2), under-representation is well documented. Among collections in The Cancer Imaging Archive that report demographics, American Indian and Alaska Native, Black and Hispanic patients are under-represented by 47.7%, 35.8% and 14.7%, respectively, relative to the United States cancer population (Dulaney & Virostko, 2024). A scoping review of dermatology AI found datasets dominated by Fitzpatrick skin types I–III, no study identifying Hispanic or Latino participants, and lower accuracy for darker skin where results were reported (Pérez-Lalinde & Flores, 2026). In a cohort used to model Alzheimer disease progression, 92.4% of participants were White, and sensitivity was lower for Hispanic, Black and Asian participants across several model types (Yuan et al., 2023). In genomics, the under-representation of African populations, together with legacies of mistrust, weakens model validity and amplifies bias (Gaye et al., 2026). In clinical language, large language models used to generate synthetic electronic health records showed demographic bias in every model tested, with more bias in larger models (R. Huang et al., 2025).

At measurement and formulation (Stages 3–4), devices and proxies perform unevenly. Pulse oximeters often overestimate arterial oxygen saturation in patients with darker skin, raising the risk of occult hypoxaemia (Savorgnan et al., 2026). Photoplethysmography-based wearables for atrial fibrillation perform worse on darker skin, in ways that amount to a hidden form of race correction (Merid & Volpe, 2023). Community- and insurance-based proxies of socioeconomic status misrepresent individual circumstances, and their accuracy differs by race (Makkad et al., 2025). An online database lists 42 risk calculators, 5 laboratory reference ranges and 5 devices that use race or ethnicity or perform differently by it (Visweswaran et al., 2025), and organised medicine has begun efforts to eliminate harmful race-based clinical algorithms (Cleveland Manchanda et al., 2024).

At development and evaluation (Stages 5–6), bias can be inherited and concealed. A chest radiography foundation model showed reduced performance for female and Black patients, and this bias was inherited by downstream tasks (Glocker et al., 2023). Coarse race and ethnicity labels masked underdiagnosis disparities that appeared only when subgroups were disaggregated (Bachina et al., 2023). In cardiac disease prediction, false-negative rates were higher for women in 13 of 16 experiments (Straw et al., 2024). Debiasing has shown promise: contrastive learning mitigated 88.5% of the disparities detected in pathology AI (Lin et al., 2025), and generative augmentation and reinforcement learning improved fairness (Ktena et al., 2024; Yang et al., 2023). Late corrections are less reliable. Post-processing gave mixed results for a breast cancer staging model trained on data in which more than 70% of patients were White (Soltan & Washington, 2024).

At deployment and monitoring (Stages 7–8), human and temporal factors become salient. In a laboratory reader study, automation bias accounted for 45.5% of the errors that remained when clinicians used AI support and affected clinicians at all levels of expertise (Wang et al., 2023). A deployed malnutrition model showed calibration differences between Black and White patients, which logistic recalibration reduced in a hold-out sample (Liou et al., 2024). Across 11 years of national veterans' data, fairness gaps in surgical-outcome models drifted, and routine model updates sometimes restored fairness and sometimes worsened it (Davis et al., 2025).

Not every cascade proceeds unchecked. In an observational study, the gap in diabetic eye-examination adherence between Asian and Black patients narrowed from 15.6% in 2019 to 3.5% in 2021 after autonomous AI was introduced (J. J. Huang et al., 2024). A multimodal prostate cancer model developed on racially diverse trial data showed no evidence of algorithmic bias between men of African and non-African origin (Roach et al., 2025).

Within the evidence reviewed, no account links these findings into a single causal process. Lifecycle frameworks list entry points and audit questions stage by stage, and one aggregates residual disparity across model checkpoints (Alluhaibi, 2026). None, however, formally models the channels through which bias passes between stages or the conditions under which it compounds (Nazer et al., 2023; Tejani et al., 2024). The BCM is proposed to fill this gap.

© IIARJournal. All rights reserved.

3. MATERIALS AND METHODS

3.1 RESEARCH DESIGN

The study used a theory-building conceptual design. An integrative review was combined with theory synthesis to construct a process model and derive testable propositions. No patient data were collected. The study population was the peer-reviewed literature on algorithmic bias and health equity in medical AI, and the unit of analysis was the journal article.

3.2 SEARCH STRATEGY AND SOURCES

Seven search strings (S1–S7) were aligned with parts of the pipeline: core bias and equity terms; data generation and sampling; measurement, labels and proxies; development and evaluation; deployment and feedback; global health and low- and middle-income settings; and lifecycle and sociotechnical frameworks. They were run in PubMed/MEDLINE through the E-utilities interface of the National Center for Biotechnology Information, on title and abstract fields. Searches were limited to 2023–2026, English and the journal-article publication type, and excluded preprints, editorials, comments, letters, news items and errata. The searches returned 1,529 hits. Sampling was purposive: the 40 most relevant records from each string were screened, giving 265 unique records. Eight complementary OpenAlex queries captured journals outside MEDLINE, and the 15 most-cited 2023–2026 journal articles from each were screened. All searches were run on 20 October 2026; the full strings, filters and hit counts are in Supplementary File S1.

3.3 INCLUSION AND EXCLUSION CRITERIA

Articles were included if they (a) were peer-reviewed journal articles published in 2023–2026, (b) addressed bias, fairness or equity in AI or algorithmic tools used in health care, and (c) contributed evidence or argument relevant to at least one pipeline stage. Conference papers, book chapters, preprints, editorials and comments were excluded, as were unmatched or retracted articles. Screening was by title, then abstract. Forty-nine articles met the criteria; one retracted article and several non-peer-reviewed commentaries were excluded. As a declared exception to the date window, four earlier peer-reviewed articles are cited solely to attribute the theoretical lenses and the definition of health equity (Dannefer, 2003; Phelan et al., 2010; Trist & Bamforth, 1951; Whitehead, 1992). They were verified in the same way but were not part of the coded evidence base.

3.4 VERIFICATION OF THE EVIDENCE BASE

Each included article's digital object identifier (DOI) was resolved through the Crossref application programming interface. Every bibliographic field (type, title, journal, authors, volume, issue, pages and year) was compared with the extracted record and with PubMed or, for the one article not indexed in PubMed, OpenAlex. Discrepancies were resolved in favour of the publisher record. Crossref update notices, which include Retraction Watch data, and PubMed comment-and-correction links were checked for retractions, corrections and expressions of concern. No included article had been retracted; two carried corrections, which were reviewed before citation. Every numerical finding and every claim about a cited study was checked against the source abstract or, where available, the full text.

3.5 DATA EXTRACTION AND ANALYSIS

Data were extracted into a 22-field literature matrix of bibliographic and analytic fields (Supplementary File S1). Each article was coded deductively to one or more of eight pipeline stages and to one or more of three mechanisms. *Entry* is bias introduced at a stage. *Propagation* is bias carried forward, or left undetected, into a later stage or setting. *Compounding* is bias amplified beyond simple accumulation through interaction between stages, the consequences of an output, or time. Articles reporting an interrupted cascade formed a fourth category. Inductively, recurring conditions that strengthened or weakened these mechanisms were coded as moderators, and evidence of interrupted cascades as candidate circuit breakers. The model was then built abductively by integrating the coded evidence with the three theoretical lenses.

3.6 BASES OF JUDGEMENT

An article was coded for entry when it identified where a group-differential distortion arises. It was coded for propagation when it showed or argued that a distortion persists into, is inherited by, or passes unchecked through a later stage or setting, including through gaps in evaluation, reporting or governance.

© IIARJournal. All rights reserved.

It was coded for compounding when it showed or argued that a disparity is amplified through interacting stages, the consequences of an output, or time. The codes therefore record where the literature addresses a mechanism, not only where it has been demonstrated. Moderators were conditions reported to strengthen or weaken these mechanisms. Candidate circuit breakers were reports of reduced disparity after an intervention, detection of a bias-relevant shift, or no detectable bias under a specific design choice. A proposition was retained only where at least two included articles provided evidence that motivated it. The propositions are hypotheses: the cited studies motivate them but do not test them.

3.7 VARIABLE MEASUREMENT

All analytic variables were categorical: pipeline stage (eight categories), mechanism (three mechanisms plus interrupted cascade), study design (five categories), first-author location and publication year. Stage and mechanism allowed multiple codes. The variables were summarised as frequencies and percentages (Table 1). Stage-level fidelity ($f_{g,k}$; Section 4.1.2) is a theoretical construct for future operationalisation, for example as group-specific sensitivity or calibration at each stage. It was not measured in this study.

3.8 RELIABILITY, VALIDITY AND DISCLOSURES

Validity was supported by restricting the evidence to verified peer-reviewed sources, checking every claim against its source and retaining disconfirming cases. Reliability was supported by a documented search log, explicit coding categories and a stage key (Supplementary File S1). Ethical approval was not required because no human participants were involved. *AI usage disclosure:* a generative AI assistant (Claude, Anthropic) supported database searching, Crossref metadata verification and data extraction; it also checked claims against their sources and assisted with drafting and revision. The authors reviewed and edited all outputs and take full responsibility for the content.

4. RESULTS AND DISCUSSION

4.1 RESULTS

4.1.1 Profile of the Evidence Base

Table 1 summarises the 49 included articles. Quantitative empirical studies formed the largest group (n = 23; 26 including qualitative and survey studies). Of the 49 first authors, 30 were based in the United States, and only one article came from Africa. Stage coverage was uneven: Stages 2, 6 and 7 were well covered (n = 21–23), whereas Stage 1 (n = 9) and Stage 4 (n = 10) received least attention. Propagation (n = 35) and entry (n = 30) were often addressed, but compounding was explicit in only 10 articles. Two articles reported an interrupted cascade as their main finding (J. J. Huang et al., 2024; Roach et al., 2025), and five mitigation or monitoring studies were also coded as candidate circuit breakers.

Table. 1. Characteristics of the included evidence (N = 49)

Characteristic	Category	*n* (%)
Publication year	2023	10 (20.4)
	2024	18 (36.7)
	2025	12 (24.5)
	2026	9 (18.4)
Study design	Quantitative empirical studies (model, dataset and algorithm analyses)	23 (46.9)
	Reviews (systematic, scoping, narrative, critical)	15 (30.6)
	Perspectives, opinion and conceptual analyses	6 (12.2)
	Qualitative and survey studies	3 (6.1)
	International consensus guidelines	2 (4.1)
First-author location	United States	30 (61.2)
	United Kingdom	6 (12.2)
	Middle East (United Arab Emirates, Saudi Arabia, Jordan)	5 (10.2)
	Asia (China, India)	3 (6.1)
	Continental Europe (Germany, Spain)	2 (4.1)

© IIARJournal. All rights reserved.

Characteristic	Category	*n* (%)
	Other Americas (Canada, Colombia)	2 (4.1)
	Africa (South Africa)	1 (2.0)
Pipeline stage[a]	1 Social context and data generation	9 (18.4)
	2 Data collection and sampling	23 (46.9)
	3 Measurement and labelling	13 (26.5)
	4 Problem formulation and feature engineering	10 (20.4)
	5 Model development and optimisation	17 (34.7)
	6 Validation and evaluation	22 (44.9)
	7 Deployment and clinical integration	21 (42.9)
	8 Monitoring and feedback	13 (26.5)
Mechanism[a]	Entry	30 (61.2)
	Propagation	35 (71.4)
	Compounding	10 (20.4)
	Interrupted cascade or no bias detected	2 (4.1)

Note. [a] Articles could be coded to more than one stage and mechanism, so percentages do not sum to 100. Location is the first author's country; for one article whose affiliation states no country (Janneker, 2025), the study setting was used. Source: authors' analysis of the literature matrix (Supplementary File S1).

4.1.2 The Bias Cascade Model

Figure 1 presents the BCM, which comprises eight pipeline stages, three mechanisms, moderators that strengthen or weaken those mechanisms, and candidate circuit breakers that may interrupt the cascade. Table 2 maps each stage to its entry route, the propagation or compounding risk that operates there, illustrative evidence and an equity check. Together, the eight checks form the stage-matched equity-check protocol.

© IIARJournal. All rights reserved.

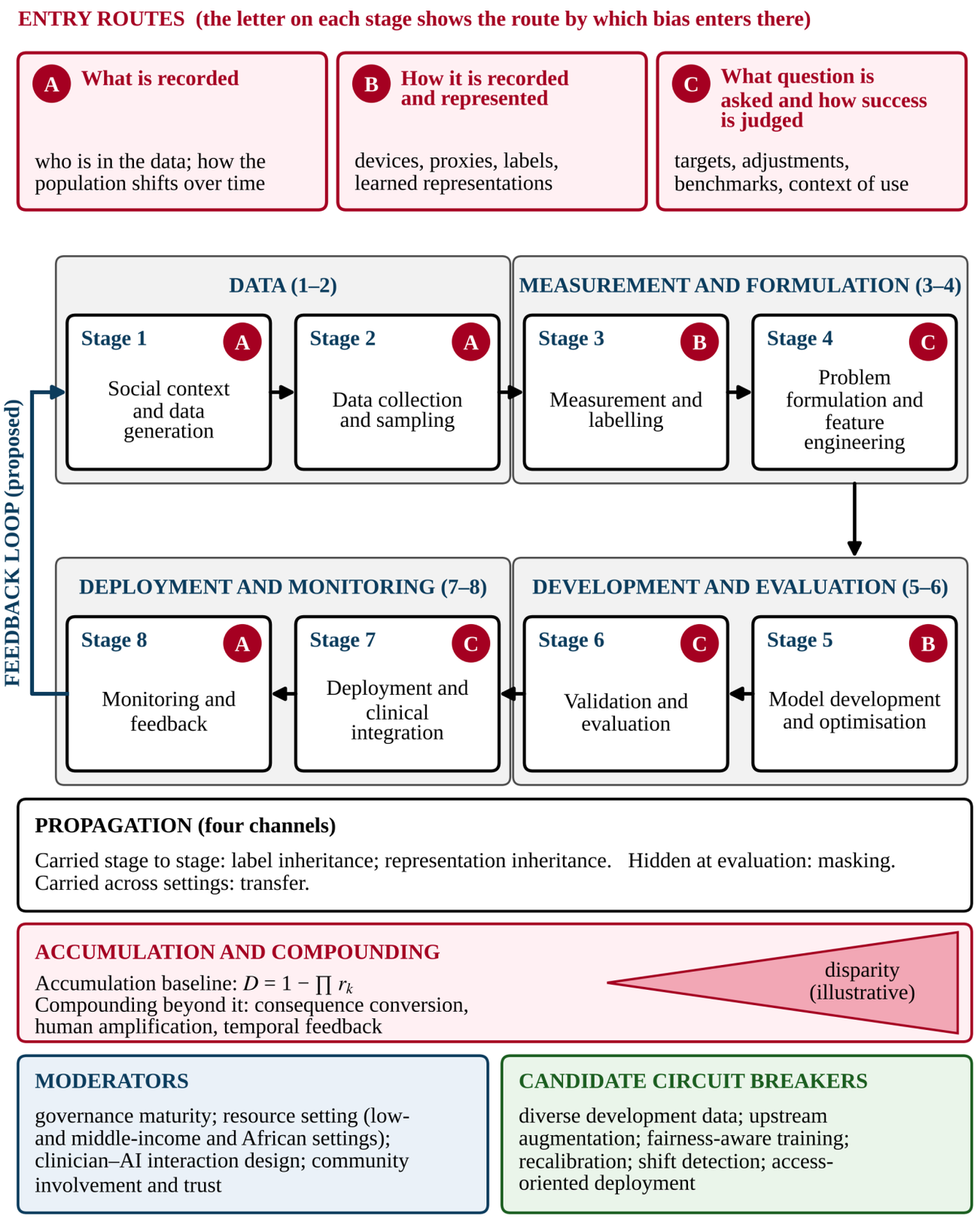


Figure. 1. The Bias Cascade Model. Bias enters each of eight pipeline stages through one of three routes (A–C), is carried forward or hidden through four propagation channels, accumulates across stages and compounds under three conditions, with a feedback loop from monitoring to data generation. Moderators condition the strength of the cascade; candidate circuit breakers may interrupt it. The disparity wedge is illustrative, not to scale.

© *IIARJournal. All rights reserved.*

Entry. The evidence points to three entry routes, labelled A–C in Figure 1 and Table 2, that together cover all eight stages. The first, route A, is *what is recorded*: who is present in the data (Stages 1–2) and how the recorded population shifts over time (Stage 8). For example, groups under-represented in cancer clinical trials are also under-represented in a major cancer imaging archive (Dulaney & Virostko, 2024). The second, route B, is *how it is recorded and represented*. It covers measurement devices, proxies and labels that work less well for some groups at Stage 3 (Makkad et al., 2025; Savorgnan et al., 2026), and the learned representations and aggregation choices of model development at Stage 5. The third, route C, is *what question is asked and how success is judged*. It covers the target outcome and the adjustments built into a formula at Stage 4 (Diao et al., 2024; Zanger-Tishler et al., 2024), the benchmarks used for evaluation at Stage 6 and the context in which the output is used at Stage 7. The direction of an entered bias is not always predictable. An electronic health record–based paediatric obesity classifier performed *better* for African American and Medicaid-enrolled children; the authors hypothesised that the most predictive diagnostic patterns were more common in those groups (Campbell et al., 2024).

Table. 2. Stages of the Bias Cascade Model: entry routes, propagation and compounding risks, illustrative evidence and stage-matched equity checks

Stage (*n* articles)	Entry route (A–C) and typical bias	Propagation or compounding risk	Illustrative evidence (2023–2026)	Stage-matched equity check
1. Social context and data generation (9)	A: historical and structural bias	Origin of the cascade	Shipton & Vitale (2024); Mohottige et al. (2023); Gaye et al. (2026)	Are the structural determinants that shape who receives care and is recorded documented, and were affected communities consulted?
2. Data collection and sampling (23)	A: representation and sampling bias	Representation inheritance	Dulaney & Virostko (2024); Pérez-Lalinde & Flores (2026); Yuan et al. (2023)	Does the dataset's composition match the intended deployment population, and is it documented to the STANDING Together recommendations (Alderman et al., 2025)?
3. Measurement and labelling (13)	B: device, label and proxy bias	Label inheritance	Savorgnan et al. (2026); Merid & Volpe (2023); Makkad et al. (2025)	Do devices, proxies and labels perform equally across groups?
4. Problem formulation and feature engineering (10)	C: target choice, race correction, proxy variables	Label inheritance into feature selection	Zanger-Tishler et al. (2024); Diao et al. (2024); Visweswaran et al. (2025)	Is the target the outcome of interest or a proxy for it, and is every race-based or proxy input justified, with its downstream consequences modelled?
5. Model development and optimisation (17)	B: aggregation, objective-function and pretrained-representation bias	Representation inheritance	Glocker et al. (2023); Weng et al. (2024); Lin et al. (2025)	Have models built on pretrained components been evaluated for subgroup bias on a representative cohort, and was a fairness-aware objective considered?
6. Validation and evaluation (22)	C: evaluation bias, benchmark mismatch	Masking	Bachina et al. (2023); Straw et al. (2024); Muralidharan et al. (2024)	Are performance and calibration reported for granular subgroups, and validated externally in the deployment population?
7. Deployment and clinical integration (21)	C: context-of-use mismatch	Transfer; consequence conversion; human amplification	Mukhopadhyay et al. (2026); Wang et al. (2023); J. J. Huang et al. (2024)	What action does the output trigger, who bears the cost of each error type, and what safeguards limit automation bias?
8. Monitoring and feedback (13)	A: dataset shift over time	Temporal feedback; fairness drift	Davis et al. (2025); Koch et al. (2024); Liou et al. (2024)	Is fairness monitored continuously, with shift detection and recalibration, and are model updates checked for new disparities?

Note. *n* = number of included articles mapped to the stage (multiple coding permitted). Entry routes (as in Figure 1): A, what is recorded (Stages 1, 2 and 8); B, how it is recorded and represented (Stages 3 and 5); C, what question is asked and how success is judged (Stages 4, 6 and 7).

Propagation. Bias is carried forward through four channels: two carry it from stage to stage, one hides it at evaluation and one carries it across settings. (a) *Label inheritance.* When a proxy label is measured

© IIARJournal. All rights reserved.

differently across groups, adding a feature can improve prediction of the proxy while worsening prediction of the true outcome; label bias is thus transmitted into feature selection (Zanger-Tishler et al., 2024). (b) *Representation inheritance.* Biases in a pretrained chest radiography foundation model carried through to disease-detection models built on its frozen features (Glocker et al., 2023). General-purpose embeddings retain signals about sensitive attributes, which Weng et al. (2024) argue should not be stripped out; each downstream model must therefore be evaluated for bias. Synthetic health records generated by large language models distort demographic distributions, a concern where such records are proposed as training data (R. Huang et al., 2025). (c) *Masking.* Evaluation choices determine whether propagated bias is visible. Coarse demographic labels hide disparities (Bachina et al., 2023), and most AI-enabled devices approved or cleared in the United States report no demographic breakdown (Muralidharan et al., 2024; Patel, 2026). (d) *Transfer.* When a tool built in a high-resource setting is deployed unchanged in a low-resource one, the bias embedded at entry travels with it (Joseph, 2025).

Accumulation and compounding. Distortions accumulate across stages even when the stages do not interact. As an accumulation baseline, let $f_{g,k}$ denote the fidelity of stage k for group g, that is, the proportion of clinically relevant signal the stage preserves. End-to-end fidelity is the product of stage-level fidelities (Equation 1). The relative disparity D between a disadvantaged group d and an advantaged group a follows from Equation 2, where $r_k = f_{d,k} / f_{a,k}$.

$$F_g = \prod_{k=1}^{8} f_{g,k} \tag{1}$$

$$D = 1 - \frac{F_d}{F_a} = 1 - \prod_{k=1}^{8} r_k \tag{2}$$

Suppose four stages each retain 95% of the relative signal for the disadvantaged group ($r_k = 0.95$) and the other stages are unbiased. Then $D = 1 - 0.95^4 \approx 0.19$. Fully correcting any one biased stage reduces D only to $1 - 0.95^3 \approx 0.14$, a reduction of about 23%. Most of the disparity persists because three biased stages are untouched, consistent with the post-processing evidence (Soltan & Washington, 2024). For small distortions this baseline is close to additive; the numbers are illustrative, not estimates.

Compounding occurs when stages interact, so that end-to-end disparity exceeds this baseline. Three conditions produce it. The first is *consequence conversion*, in which an error asymmetry becomes harm through the action the output triggers. The same false-positive disparity has different consequences depending on what the prediction prompts (Mukhopadhyay et al., 2026). A single change to lung-function reference equations may reclassify ventilatory impairment for an estimated 12.5 million people, with downstream effects on disability, occupational and military decisions (Diao et al., 2024). The second is *human amplification*, in which clinicians adopt model errors through automation bias (Abdelwanis et al., 2024; Wang et al., 2023). The third is *temporal feedback*. Performance can deteriorate after deployment when the real-world population differs from the training cohort (Cross et al., 2024), and fairness drifts over time, with model updates either restoring or worsening it (Davis et al., 2025). The BCM further proposes that outputs that shape subsequent care and records can feed bias back into future training data.

Moderators. Four moderators condition the strength of the cascade. The first is governance maturity. Equity was inconsistently considered in AI governance at academic medical centres, and disparities in information technology resources between health systems were repeatedly identified as a threat to equity (Nong et al., 2024). The second is resource setting. Populations in low-resource settings are often under-represented in training data, and their health systems frequently lack the capacity to audit AI tools (Joseph, 2025). The third is the design of clinician–AI interaction (Wang et al., 2023). The fourth is community involvement and trust, particularly where legacies of research exploitation have produced mistrust (Gaye et al., 2026) and where members of the public seek participatory engagement in AI design and regulation (Townsend et al., 2025).

Candidate circuit breakers. These were identified at several stages: development on racially diverse trial data, after which no bias was detected (Roach et al., 2025); upstream data augmentation (Ktena et al., 2024); fairness-aware training (Lin et al., 2025; Yang et al., 2023); post-deployment recalibration (Liou et

© IIARJournal. All rights reserved.

al., 2024); distribution-shift detection (Koch et al., 2024); and point-of-care deployment associated with wider access (J. J. Huang et al., 2024).

4.1.3 Propositions

Six propositions follow from the model (Table 3). Each is a hypothesis that states a directional relationship, identifies the included studies that motivate it and suggests how it could be tested.

Table. 3. Propositions (hypotheses for testing) derived from the Bias Cascade Model

Proposition	Hypothesised relationship	Motivating evidence (what it shows)	Suggested empirical test
P1. Compounding beyond accumulation	Where consequence conversion, human amplification or temporal feedback operate, end-to-end disparity exceeds the accumulation baseline, so correcting a single stage reduces it by less than that stage's share of stage-level bias.	Soltan & Washington (2024): a single post-processing correction did not consistently reduce disparity; Alluhaibi (2026): disparity can be tracked across lifecycle checkpoints	End-to-end audits comparing observed disparity with the accumulation baseline after sequential stage-specific corrections
P2. Upstream leverage	Bias entering at Stages 1–4 is less recoverable by downstream correction than bias of similar size entering at Stages 5–6, because it alters the reference standard.	Zanger-Tishler et al. (2024): a biased proxy label distorts the reference standard and feature choice; Ktena et al. (2024): upstream data augmentation improved fairness; Soltan & Washington (2024): late correction fell short	Simulation and benchmark studies injecting matched bias at different stages
P3. Masking	The probability that a propagated disparity is detected falls as demographic labels become coarser and subgroup reporting less complete.	Bachina et al. (2023): coarse labels masked disparities; Straw et al. (2024): sex disparities were missed without stratified reporting; Muralidharan et al. (2024): most approvals reported no demographic data, and race or ethnicity reporting was rare	Re-analysis of published evaluations with disaggregated labels
P4. Consequence conversion	The harm produced by a given error asymmetry increases with the consequence of the action the output triggers and with clinicians' reliance on the output.	Mukhopadhyay et al. (2026): harm depends on the action triggered; Diao et al. (2024): one equation change carries different consequences across decisions; Wang et al. (2023): automation bias converts AI errors into clinician errors	Decision-analytic modelling and clinician-in-the-loop experiments across use cases
P5. Temporal compounding	Without continuous fairness monitoring, fairness gaps widen over time in deployed models, and reuse of model outputs as data feeds bias back into early stages.	Davis et al. (2025): fairness drifts and updates can worsen it; Koch et al. (2024): post-market shifts can be detected; R. Huang et al. (2025): records generated by large language models and proposed for training are demographically biased	Longitudinal fairness monitoring of deployed models; studies of synthetic-data reuse
P6. Transfer amplification	The disparity of a transferred model increases with the difference between development and deployment populations, and is reduced by local validation and community governance.	Joseph (2025): tools transferred unchanged carry embedded bias; Gaye et al. (2026): validity concerns in under-represented African populations and calls for population-specific validation and local governance	Multi-site external validation of imported models in African and other low- and middle-income settings

4.2 DISCUSSION

The findings answer the three research questions. On the first, bias can enter at all eight stages through three routes. On the second, it is carried forward or hidden through four channels. On the third, stage-level biases accumulate even when stages act independently and compound beyond that baseline under the three compounding conditions. The model therefore locates the greatest leverage for intervention at the point of entry, at disaggregated evaluation and in monitoring.

The evidence supports all three mechanisms, but unevenly. Entry and propagation are well documented, whereas compounding is explicit in only a fifth of the literature, consistent with the field's focus on single-stage audits. In one systematic review, 60% of mitigation strategies were applied during development and silent evaluation (Gadhoumi et al., 2026). The BCM predicts that, by this point, much of the disparity has already been established upstream, while much of the harm has yet to materialise downstream. The same review's finding that 85% of studies reported effective bias reduction should therefore be interpreted cautiously: a reduction measured at the model output is not a reduction in end-to-end inequity.

The BCM extends previous work in three ways. First, stage typologies (Nazer et al., 2023; Tejani et al., 2024) identify where bias enters and organise

© IIARJournal. All rights reserved.

mitigation stage by stage. Tejani et al. (2024) note that bias can persist across the lifecycle and feed back from deployment into development, but neither typology formally specifies the channels that couple stages, a gap the BCM addresses. Second, Alluhaibi's (2026) Lifecycle Equity Debt is an accounting summary rather than a causal model. The BCM proposes how disparities combine: they accumulate multiplicatively when stages are independent and compound under consequence conversion, human amplification and temporal feedback. Third, standards such as STANDING Together for datasets (Alderman et al., 2025) and FUTURE-AI for the lifecycle (Lekadir et al., 2025) can operationalise many checks in Table 2, and the BCM indicates where they have most leverage. The two interrupted cascades (J. J. Huang et al., 2024; Roach et al., 2025) show that bias need not cascade unchecked.

The findings accord with each theoretical lens. Consistent with fundamental cause theory, the least-studied stages, social context and problem formulation, are those at which structural inequities first become data. Race-based clinical algorithms also remain in use decades after race coefficients were first encoded in clinical equations (Mohottige et al., 2023; Visweswaran et al., 2025). As sociotechnical theory would predict, much of the propagation and compounding evidence concerns deployment, governance and human interaction rather than the model itself. In keeping with cumulative disadvantage, longitudinal evidence shows that fairness gaps can widen over time when models are maintained without fairness monitoring (Davis et al., 2025).

Implications for African health systems. Only one of the 49 articles originated in Africa. The two articles centred on African health systems and genomic data were perspectives rather than empirical evaluations (Gaye et al., 2026; Janneker, 2025), and the only empirical study of people of African origin analysed phase III trial participants (Roach et al., 2025). Populations least represented in training data are therefore those on which imported systems can be expected to perform worst, often where governance and post-deployment monitoring capacity are weakest.

The implications differ by actor. For *health ministries*, the model implies requiring local validation in the intended deployment population before procurement. This validation depends on what Janneker (2025) calls for in South Africa: interoperable health data, a unified digital health infrastructure and bias-aware models trained on local datasets. For *regulators*, the model implies moving from point-in-time clearance to total-product-lifecycle oversight. That oversight would include demographic transparency and continuous post-market monitoring (Patel, 2026), standardised demographic reporting (Muralidharan et al., 2024) and periodic auditing for newly arising bias (Lekadir et al., 2025). For *research institutions*, it implies leading African data generation under culturally grounded governance that addresses legacies of mistrust (Gaye et al., 2026), and documenting datasets to recognised standards (Alderman et al., 2025).

Limitations and future research. The model is conceptual, and its propositions have not been tested. The evidence came from PubMed and OpenAlex and was sampled by relevance and citation count; Scopus, Web of Science and CINAHL were not searched, so relevant work may have been missed. Restricting the evidence to 2023–2026 ensured currency but excluded earlier empirical studies. The accumulation baseline (Equations 1–2) is illustrative. Future research should operationalise stage-level fidelity, estimate the size of compounding through end-to-end audits of deployed systems, and test the propositions in African and other low- and middle-income settings.

5. CONCLUSION AND RECOMMENDATION

Algorithmic bias in medical AI can enter at every stage of the data-to-deployment pipeline. It is carried forward through identifiable channels, and it accumulates and compounds, so a correction at one stage can leave much of the end-to-end disparity in place. The Bias Cascade Model integrates three theoretical perspectives with 49 verified recent studies and yields six testable propositions and a stage-matched equity-check protocol.

Five recommendations follow. First, developers should apply the equity checks in Table 2 at every stage, not only during model training. Second,

© IIARJournal. All rights reserved.

evaluations should be disaggregated by granular demographic group and conducted in the intended deployment population. Third, harm should be assessed in the clinical context of use, including the actions that outputs trigger and safeguards against automation bias. Fourth, regulators and health systems should require continuous fairness monitoring and recalibration after deployment. Fifth, health ministries, funders and research institutions in Africa should invest in locally governed data generation and validation, so that the populations at greatest risk from the cascade are no longer the least represented at its first stage.

6. CONTRIBUTION TO EXISTING KNOWLEDGE

Theoretically, the paper recasts algorithmic bias in medicine as a cascade of entry, propagation and compounding, grounded in three theories. Methodologically, it offers an accumulation baseline against which compounding can be measured, together with six testable propositions. Practically, it provides a stage-matched equity-check protocol and implications for health ministries, regulators and research institutions, with particular attention to African populations that the evidence has largely left out.

7. REFERENCES

Abdelwanis, M., Alarafati, H. K., Tammam, M. M. S., & Simsekler, M. C. E. (2024). Exploring the risks of automation bias in healthcare artificial intelligence applications: A Bowtie analysis. *Journal of Safety Science and Resilience, 5*(4), 460–469. https://doi.org/10.1016/j.jnlssr.2024.06.001

Ahmad, A., Vallès, Y., & Idaghdour, Y. (2026). Bias in AI systems: Integrating formal and socio-technical approaches. *Frontiers in Big Data, 8*, 1686452. https://doi.org/10.3389/fdata.2025.1686452

Alderman, J. E., Palmer, J., Laws, E., McCradden, M. D., Ordish, J., Ghassemi, M., Pfohl, S. R., Rostamzadeh, N., Cole-Lewis, H., Glocker, B., Calvert, M., Pollard, T. J., Gill, J., Gath, J., Adebajo, A., Beng, J., Leung, C. H., Kuku, S., Farmer, L.-A., . . . Liu, X. (2025). Tackling algorithmic bias and promoting transparency in health datasets: The STANDING Together consensus recommendations. *The Lancet Digital Health, 7*(1), e64–e88. https://doi.org/10.1016/S2589-7500(24)00224-3

Alluhaibi, R. (2026). Algorithmic bias in AI-based diagnostic models: A systematic review and lifecycle framework for fairness auditing and governance in healthcare. *Healthcare, 14*(19), 3293. https://doi.org/10.3390/healthcare14193293

Bachina, P., Garin, S. P., Kulkarni, P., Kanhere, A., Sulam, J., Parekh, V. S., & Yi, P. H. (2023). Coarse race and ethnicity labels mask granular underdiagnosis disparities in deep learning models for chest radiograph diagnosis. *Radiology, 309*(2), e231693. https://doi.org/10.1148/radiol.231693

Berdahl, C. T., Baker, L., Mann, S., Osoba, O., & Girosi, F. (2023). Strategies to improve the impact of artificial intelligence on health equity: Scoping review. *JMIR AI, 2*, e42936. https://doi.org/10.2196/42936

Campbell, E. A., Bose, S., & Masino, A. J. (2024). Conceptualizing bias in EHR data: A case study in performance disparities by demographic subgroups for a pediatric obesity incidence classifier. *PLOS Digital Health, 3*(10), e0000642. https://doi.org/10.1371/journal.pdig.0000642

Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering, 7*(6), 719–742. https://doi.org/10.1038/s41551-023-01056-8

Cleveland Manchanda, E. C., Aikens, B., De Maio, F., Jordan, W., Brown, J. T., Sivashanker, K., & Maybank, A. (2024). Efforts in organized medicine to eliminate harmful race-based clinical algorithms. *JAMA Network Open, 7*(3), e241121. https://doi.org/10.1001/jamanetworkopen.2024.1121

Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for

© IIARJournal. All rights reserved.

clinical decision-making. *PLOS Digital Health, 3*(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651

Dannefer, D. (2003). Cumulative advantage/disadvantage and the life course: Cross-fertilizing age and social science theory. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 58*(6), S327–S337. https://doi.org/10.1093/geronb/58.6.s327

Davis, S. E., Dorn, C., Park, D. J., & Matheny, M. E. (2025). Emerging algorithmic bias: Fairness drift as the next dimension of model maintenance and sustainability. *Journal of the American Medical Informatics Association, 32*(5), 845–854. https://doi.org/10.1093/jamia/ocaf039

Diao, J. A., He, Y., Khazanchi, R., Nguemeni Tiako, M. J., Witonsky, J. I., Pierson, E., Rajpurkar, P., Elhawary, J. R., Melas-Kyriazi, L., Yen, A., Martin, A. R., Levy, S., Patel, C. J., Farhat, M., Borrell, L. N., Cho, M. H., Silverman, E. K., Burchard, E. G., & Manrai, A. K. (2024). Implications of race adjustment in lung-function equations. *New England Journal of Medicine, 390*(22), 2083–2097. https://doi.org/10.1056/NEJMsa2311809

Dulaney, A., & Virostko, J. (2024). Disparities in the demographic composition of The Cancer Imaging Archive. *Radiology: Imaging Cancer, 6*(1), e230100. https://doi.org/10.1148/rycan.230100

Efird, C. R., Park, C., Mishra, P., Barr, A., Twinamatsiko, A. J., & King, C. J. (2026). Recommended guidelines concerning artificial intelligence and health equity: A scoping review. *Health Affairs Scholar, 4*(10), qxag236. https://doi.org/10.1093/haschl/qxag236

Gadhoumi, K., Senior, R., Wei, S., Ledbetter, L., Green, M. D., Bonner, B. D., Mosley, Y., Seidler, K., Green, K., Lee, D., Hong, C., Guilamo Ramos, V., & Cary, M. P., Jr. (2026). Strategies for mitigating artificial intelligence bias in healthcare: A systematic review. *JAMIA Open, 9*(3), ooag081. https://doi.org/10.1093/jamiaopen/ooag081

Gaye, A., Bonham, V. L., & Mersha, T. B. (2026). Trust, mistrust, and the promise of AI in genomics for African populations. *The American Journal of Human Genetics, 113*(7), 1367–1371. https://doi.org/10.1016/j.ajhg.2026.05.009

Glocker, B., Jones, C., Roschewitz, M., & Winzeck, S. (2023). Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence, 5*(6), e230060. https://doi.org/10.1148/ryai.230060

Huang, J. J., Channa, R., Wolf, R. M., Dong, Y., Liang, M., Wang, J., Abramoff, M. D., & Liu, T. Y. A. (2024). Autonomous artificial intelligence for diabetic eye disease increases access and health equity in underserved populations. *npj Digital Medicine, 7*(1), 196. https://doi.org/10.1038/s41746-024-01197-3

Huang, R., Wu, H., Yuan, Y., Xu, Y., Qian, H., Zhang, C., Wei, X., Lu, S., Zhang, X., Kan, J., Wan, C., & Liu, Y. (2025). Evaluation and bias analysis of large language models in generating synthetic electronic health records: Comparative study. *Journal of Medical Internet Research, 27*, e65317. https://doi.org/10.2196/65317

Hussain, S. A., Bresnahan, M., & Zhuang, J. (2025). The bias algorithm: How AI in healthcare exacerbates ethnic and racial disparities – a scoping review. *Ethnicity & Health, 30*(2), 197–214. https://doi.org/10.1080/13557858.2024.2422848

Janneker, W. (2025). The role of AI in transforming healthcare in South Africa. *South African Medical Journal, 115*(5b), e3672. https://doi.org/10.7196/SAMJ.2025.v115i5b.3672

Joseph, J. (2025). Algorithmic bias in public health AI: A silent threat to equity in low-resource settings. *Frontiers in Public Health, 13*, 1643180. https://doi.org/10.3389/fpubh.2025.1643180

Koch, L. M., Baumgartner, C. F., & Berens, P. (2024). Distribution shift detection for the postmarket surveillance of medical AI algorithms: A retrospective simulation study.

© IIARJournal. All rights reserved.

npj Digital Medicine, 7(1), 120. https://doi.org/10.1038/s41746-024-01085-w

Ktena, I., Wiles, O., Albuquerque, I., Rebuffi, S.-A., Tanno, R., Roy, A. G., Azizi, S., Belgrave, D., Kohli, P., Cemgil, T., Karthikesalingam, A., & Gowal, S. (2024). Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine, 30*(4), 1166–1173. https://doi.org/10.1038/s41591-024-02838-6

Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., Weicken, E., Asselbergs, F. W., Prior, F., Collins, G. S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J. A., Kushibar, K., Riklund, K., Marias, K., . . . Starmans, M. P. A. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ, 388*, e081554. https://doi.org/10.1136/bmj-2024-081554

Lin, S.-Y., Tsai, P.-C., Su, F.-Y., Chen, C.-Y., Li, F., Zhao, J., Ho, Y. Y., Lee, T.-L. M., Healey, E., Lin, P.-J., Kao, T.-W., Vremenko, D., Roetzer-Pejrimovsky, T., Sholl, L., Dillon, D., Lin, N. U., Meredith, D., Ligon, K. L., Lo, Y.-C., . . . Yu, K.-H. (2025). Contrastive learning enhances fairness in pathology artificial intelligence systems. *Cell Reports Medicine, 6*(12), 102527. https://doi.org/10.1016/j.xcrm.2025.102527

Liou, L., Scott, E., Parchure, P., Ouyang, Y., Egorova, N., Freeman, R., Hofer, I. S., Nadkarni, G. N., Timsina, P., Kia, A., & Levin, M. A. (2024). Assessing calibration and bias of a deployed machine learning malnutrition prediction model within a large healthcare system. *npj Digital Medicine, 7*(1), 149. https://doi.org/10.1038/s41746-024-01141-5

Makkad, H., Saini, A., Manning, E. R., Duan, Q., Colegate, S., & Brokamp, C. (2025). Racial fairness of individual- and community-level proxies of socioeconomic status among birthing parent–child dyads. *Journal of Racial and Ethnic Health Disparities, 12*(4), 2293–2298. https://doi.org/10.1007/s40615-024-02050-9

Merid, B., & Volpe, V. (2023). Race correction and algorithmic bias in atrial fibrillation wearable technologies. *Health Equity, 7*(1), 817–824. https://doi.org/10.1089/heq.2023.0034

Mohottige, D., Olabisi, O., & Boulware, L. E. (2023). Use of race in kidney function estimation: Lessons learned and the path toward health justice. *Annual Review of Medicine, 74*(1), 385–400. https://doi.org/10.1146/annurev-med-042921-124419

Mukhopadhyay, A., Zhao, Y., Chunara, R., Kronish, I. M., Lawrence, S., Blecker, S., & Adhikari, S. (2026). The importance of clinical context in evaluating algorithmic fairness: Insights from a medication adherence prediction algorithm. *Journal of the American Medical Informatics Association, 33*(10), 1866–1874. https://doi.org/10.1093/jamia/ocag097

Muralidharan, V., Adewale, B. A., Huang, C. J., Nta, M. T., Ademiju, P. O., Pathmarajah, P., Hang, M. K., Adesanya, O., Abdullateef, R. O., Babatunde, A. O., Ajibade, A., Onyeka, S., Cai, Z. R., Daneshjou, R., & Olatunji, T. (2024). A scoping review of reporting gaps in FDA-approved AI medical devices. *npj Digital Medicine, 7*(1), 273. https://doi.org/10.1038/s41746-024-01270-x

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health, 2*(6), e0000278. https://doi.org/10.1371/journal.pdig.0000278

Nong, P., Hamasha, R., & Platt, J. (2024). Equity and AI governance at academic medical centers. *The American Journal of Managed Care, 30*(Spec. No. 6), SP468–SP472. https://doi.org/10.37765/ajmc.2024.89555

Patel, N. N. (2026). Bias, external validation, and real-world implementation of artificial intelligence models in cardiovascular medicine. *The American Journal of*

© IIARJournal. All rights reserved.

Cardiology, 273, 44–50. https://doi.org/10.1016/j.amjcard.2026.06.005

Pérez-Lalinde, S., & Flores, F. (2026). Demographic reporting and the absence of Hispanic/Latino representation in dermatology artificial intelligence: A scoping review. *Skin Research and Technology, 32*(9), e70385. https://doi.org/10.1111/srt.70385

Phelan, J. C., Link, B. G., & Tehranifar, P. (2010). Social conditions as fundamental causes of health inequalities: Theory, evidence, and policy implications. *Journal of Health and Social Behavior, 51*(1, Suppl.), S28–S40. https://doi.org/10.1177/0022146510383498

Roach, M., III, Zhang, J., Mohamad, O., van der Wal, D., Simko, J. P., DeVries, S., Huang, H.-C., Joun, S., Schaeffer, E. M., Morgan, T. M., Keim-Malpass, J., Chen, E., Yamashita, R., Monson, J. M., Naz, F., Wallace, J., Bahary, J.-P., Wilke, D., Batra, S., . . . Feng, F. Y. (2025). Assessing algorithmic fairness with a multimodal artificial intelligence model in men of African and non-African origin on NRG Oncology prostate cancer phase III trials. *JCO Clinical Cancer Informatics, 9*, e2400284. https://doi.org/10.1200/CCI-24-00284

Savorgnan, F., Prabhu, J., Pilla, P., Flores, S., Loomba, R. S., & Acosta, S. (2026). Pulse oximetry, skin pigmentation, and occult hypoxemia in pediatric cardiac critical care. *Pediatric Cardiology*. Advance online publication. https://doi.org/10.1007/s00246-026-04292-2

Shipton, L., & Vitale, L. (2024). Artificial intelligence and the politics of avoidance in global health. *Social Science & Medicine, 359*, 117274. https://doi.org/10.1016/j.socscimed.2024.117274

Soltan, A., & Washington, P. (2024). Challenges in reducing bias using post-processing fairness for breast cancer stage classification with deep learning. *Algorithms, 17*(4), 141. https://doi.org/10.3390/a17040141

Straw, I., Rees, G., & Nachev, P. (2024). Sex-based performance disparities in machine learning algorithms for cardiac disease prediction: Exploratory study. *Journal of Medical Internet Research, 26*, e46936. https://doi.org/10.2196/46936

Tejani, A. S., Ng, Y. S., Xi, Y., & Rayan, J. C. (2024). Understanding and mitigating bias in imaging artificial intelligence. *RadioGraphics, 44*(5), e230067. https://doi.org/10.1148/rg.230067

Townsend, B. A., Hodge, V. J., Richardson, H., Calinescu, R., & Arvind, T. T. (2025). Cautious optimism: Public voices on medical AI and sociotechnical harm. *Frontiers in Digital Health, 7*, 1625747. https://doi.org/10.3389/fdgth.2025.1625747

Trist, E. L., & Bamforth, K. W. (1951). Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. *Human Relations, 4*(1), 3–38. https://doi.org/10.1177/001872675100400101

Visweswaran, S., Sadhu, E. M., Morris, M. M., Vis, A. R., & Samayamuthu, M. J. (2025). Online database of clinical algorithms with race and ethnicity. *Scientific Reports, 15*(1), 10913. https://doi.org/10.1038/s41598-025-94152-5

Wang, D. Y., Ding, J., Sun, A. L., Liu, S. G., Jiang, D., Li, N., & Yu, J. K. (2023). Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias. *Journal of the American Medical Informatics Association, 30*(10), 1684–1692. https://doi.org/10.1093/jamia/ocad118

Weng, W.-H., Sellergen, A., Kiraly, A. P., D'Amour, A., Park, J., Pilgrim, R., Pfohl, S., Lau, C., Natarajan, V., Azizi, S., Karthikesalingam, A., Cole-Lewis, H., Matias, Y., Corrado, G. S., Webster, D. R., Shetty, S., Prabhakara, S., Eswaran, K., Celi, L. A. G., & Liu, Y. (2024). An intentional approach to managing bias in general purpose embedding models. *The Lancet Digital Health, 6*(2), e126–e130. https://doi.org/10.1016/S2589-7500(23)00227-3

© IIARJournal. All rights reserved.

Whitehead, M. (1992). The concepts and principles of equity and health. *International Journal of Health Services, 22*(3), 429–445. https://doi.org/10.2190/986L-LHQ6-2VTE-YRRN

Yang, J., Soltan, A. A. S., Eyre, D. W., & Clifton, D. A. (2023). Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nature Machine Intelligence, 5*(8), 884–894. https://doi.org/10.1038/s42256-023-00697-3

Yuan, C., Linn, K. A., & Hubbard, R. A. (2023). Algorithmic fairness of machine learning models for Alzheimer disease progression. *JAMA Network Open, 6*(11), e2342203. https://doi.org/10.1001/jamanetworkopen.2023.42203

Zanger-Tishler, M., Nyarko, J., & Goel, S. (2024). Risk scores, label bias, and everything but the kitchen sink. *Science Advances, 10*(13), eadi8411. https://doi.org/10.1126/sciadv.adi8411

© IIARJournal. All rights reserved.