World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Review Article)

From Lab to Clinic: Addressing Bias and Generalizability in AI Diagnostic Systems

Johnson Gbenga Oyeniyi *

Department of Computing and Informatics, Bournemouth University Poole, Dorset, Bournemouth, United Kingdom.

World Journal of Advanced Research and Reviews, 2025, 28(03), 2134-2179

Publication history: Received on 16 November 2025; revised on 26 December 2025; accepted on 29 December 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.28.3.4249

Abstract

Artificial intelligence (AI) diagnostic systems demonstrate exceptional performance in controlled laboratory settings yet consistently fail to translate into equitable and reliable clinical tools. This thesis identifies and analyzes the structural roots of this translation gap, arguing that the pervasive challenges of algorithmic bias and poor generalizability are not isolated technical failures but predictable outcomes of a development paradigm that prioritizes narrow accuracy metrics over robust, equitable performance.

Through a systematic analysis of evidence across medical specialties, this research demonstrates how models trained on geographically concentrated and demographically homogeneous data systematically underperform for marginalized populations and fail when deployed in new contexts. The compounding of bias (differential performance across groups) and poor generalizability (performance degradation across settings) creates an "equity paradox" wherein AI tools perform best for populations with the least need and worst for those who could benefit most from improved diagnostic access.

This thesis reveals how current regulatory frameworks, economic incentives, and organizational structures actively reinforce these problematic practices. It moves beyond technical mitigation strategies to propose a fundamental reorientation of the AI development lifecycle that centres equity and generalizability as non-negotiable requirements. The proposed framework includes proactive data diversity, mandatory multi-site and intersectional validation, fairness-aware optimization, and robust governance structures.

The findings necessitate a paradigm shift from accuracy-focused to equity-centred AI development, with implications for researchers, regulators, healthcare institutions, and policymakers. Ultimately, this thesis contends that the technical capacity for building equitable AI diagnostics exists; what is required is the collective commitment to treat equity not as an aspirational goal but as a fundamental criterion for clinical deployment.

Keywords: Medical Artificial Intelligence; Algorithmic Bias; Generalizability; Health Equity; FDA Regulation; Machine Learning; Diagnostic Systems; Clinical Translation; Healthcare Disparities; Responsible AI

1. Introduction

1.1. The Promise of AI in Medical Diagnostics

The integration of artificial intelligence into medical diagnostics represents one of the most promising advances in modern healthcare, poised to redefine the standards of accuracy, efficiency, and accessibility in medicine. Deep learning algorithms now demonstrate exceptional, and at times superhuman, performance in detecting diseases from medical images, analyzing pathology slides, and identifying subtle patterns in complex clinical data (LeCun et al., 2015; Topol, 2019). This rapid progress is not merely theoretical; it is being rapidly codified into clinical practice. By mid-2024, the

* Corresponding author: Johnson Gbenga Oyeniyi

U.S. Food and Drug Administration (FDA) had cleared nearly 950 AI-enabled medical devices, with approximately 100 new approvals annually, predominantly in high-stakes fields like radiology, cardiology, and neurology (FDA, 2024). The market trajectory reflects this optimism, with valuations projecting explosive growth from $13.7 billion in 2024 to over $255 billion by 2033. This rapid proliferation signals a pivotal shift, suggesting that AI diagnostics has transitioned from an experimental technology to an emerging clinical reality.

1.2. The Paradox: Laboratory Success vs. Clinical Failure

Yet, this compelling narrative of technological triumph obscures a fundamental and deeply troubling paradox: despite exceptional performance in controlled laboratory settings, AI diagnostic systems consistently struggle to translate into equitable, robust, and widespread clinical use (Kelly et al., 2019). Models that achieve near-perfect accuracy on curated test sets frequently fail when deployed in different hospitals, with different patient populations, or across different geographic regions (Zech et al., 2018). This performance degradation is not random; it follows predictable and systematic patterns.

More troubling than simple performance drop is the emergence of pervasive algorithmic bias. Evidence increasingly demonstrates that these systems perform systematically worse for marginalized populations—the very groups who could benefit most from improved diagnostic access (Seyyed-Kalantari et al., 2021). For instance, a dermatology AI system may excel at detecting skin cancer in fair-skinned individuals while demonstrating 10-15% lower accuracy for patients with darker skin tones (Daneshjou et al., 2022). Similarly, cardiovascular risk prediction algorithms trained predominantly on male patients may systematically underestimate risk for women, who often present with different symptoms (Larrazabal et al., 2020). A landmark study of a widely used commercial algorithm revealed it assigned lower risk scores to Black patients than to White patients with the same level of illness, thereby restricting access to care management programs for Black patients (Obermeyer et al., 2019). This is not a collection of isolated incidents but a recurring pattern across medical specialties.

1.3. The Core Argument: A Structural Problem

This thesis argues that the translation gap between AI diagnostic systems' laboratory performance and their clinical utility is not merely a technical challenge but a structural consequence of development practices that prioritize narrow accuracy metrics over generalizability and equity (Wiens et al., 2019; Rajkomar et al., 2018). The failures of bias and poor generalizability are not bugs in the current system; they are predictable features of a development paradigm that is fundamentally misaligned with the realities of diverse healthcare ecosystems.

This structural challenge manifests through three interconnected problems

1.3.1. Algorithmic Bias

AI diagnostic systems trained on homogeneous datasets systematically underperform for underrepresented demographic groups (Obermeyer et al., 2019). This bias is predictable—an inevitable outcome of training data that reflects and amplifies existing healthcare disparities and access inequities (Gianfrancesco et al., 2018). When a staggering 71% of AI diagnostic algorithms for U.S. healthcare are trained on data from just three states—California, Massachusetts, and New York the resulting systems encode geographic and demographic privilege directly into their algorithmic infrastructure (Larson et al., 2018).

1.3.2. Poor Generalizability

Current AI systems demonstrate a critical inability to maintain performance across different institutions, populations, and clinical settings (Oakden-Rayner et al., 2020). Models optimized for single-site performance often fail when applied elsewhere, even within the same country. External validation studies consistently reveal dramatic performance degradation when algorithms encounter data characteristics not represented in their training sets, such as different medical equipment, varied clinical protocols, or distinct population demographics (Zech et al., 2018).

1.3.3. Inadequate Validation Frameworks

Existing regulatory and clinical validation processes focus primarily on demonstrating technical accuracy in controlled settings rather than ensuring robust performance across diverse real-world contexts (Wu et al., 2021). FDA approval processes, while evolving, often lack mandatory requirements for comprehensive multi-site testing or disaggregated performance reporting across demographic subgroups. This allows systems with limited evidence of equitable performance to enter clinical practice, conducting what amounts to an uncontrolled experiment on patient populations (Chen et al., 2021).

1.4. Significance and Implications for Health Equity

The implications of this translation gap extend far beyond technical inefficiency or wasted research investment. When AI diagnostic systems fail to generalize or perform inequitably, they risk creating a two-tier healthcare system where algorithmic tools enhance care for privileged populations while remaining unavailable—or worse, actively harmful—for marginalized communities (Vyas et al., 2020). This represents an *algorithmic amplification* of existing health inequities, encoded into infrastructure that will shape clinical practice for decades to come.

Moreover, the current trajectory threatens to irrevocably undermine trust in medical AI broadly. Clinicians who encounter systems performing poorly in their specific contexts, or who observe disparate outcomes across patient populations, may justifiably resist AI adoption even when specific tools could provide genuine benefit (Char et al., 2018; Kelly et al., 2019). Building sustainable, trusted AI diagnostic infrastructure therefore requires addressing these structural issues proactively, rather than reactively correcting individual failures after harm has occurred.

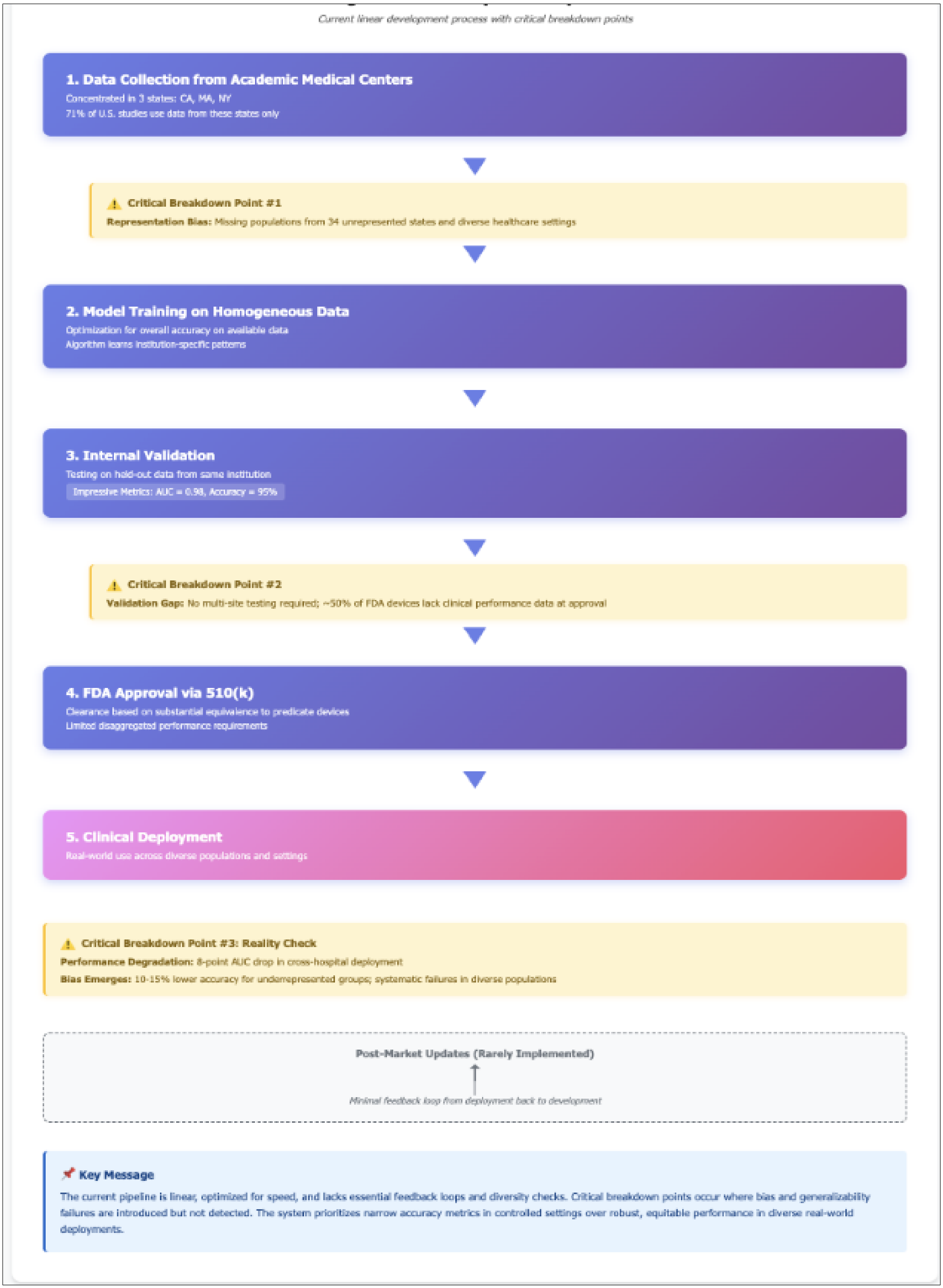
The implications of this translation gap extend beyond national borders to create global health inequities. When AI systems developed in high-income countries using Western data are deployed in low- and middle-income countries (LMICs) without validation or adaptation, they risk perpetuating a form of digital neocolonialism where technological infrastructure developed for wealthy populations is imposed on resource-constrained settings regardless of suitability. This dynamic threatens to widen, rather than bridge, global health disparities by creating AI tools that work optimally only in the contexts where they were developed, while failing precisely in the settings that could benefit most from improved diagnostic capacity.

1.5. Research Approach and Thesis Structure

This thesis examines these challenges through a critical analysis of existing literature, regulatory frameworks, and deployment case studies. It places particular emphasis on understanding root causes rather than merely documenting symptoms. The analysis proceeds systematically

- **Section 2: Background and Literature Review** provide the essential foundation, reviewing the evolution of medical AI and existing scholarship on bias, generalizability, and clinical translation, while identifying critical gaps in current research.
- **Section 3: Understanding Algorithmic Bias in Diagnostic AI** analyzes how bias manifests across medical specialties, traces its root causes to data collection and curation practices, and examines its real-world consequences for healthcare equity.
- **Section 4: The Generalizability Crisis** investigates why AI diagnostic systems fail to maintain performance across institutions and populations, with particular attention to geographic data concentration and its profound implications.
- **Section 5: The Intersection: How Bias and Generalizability Compound Each Other** demonstrates that these are not separate problems but interconnected outcomes of the same structural issues, showing how they combine to create particularly severe equity gaps.
- **Section 6: Barriers to Clinical Translation** examines the regulatory, economic, and organizational barriers that actively reinforce problematic development practices and impede the implementation of solutions.
- **Section 7: Solutions and Best Practices: Reorienting the Development Pipeline** proposes a concrete, equity-centred framework for restructuring AI development, from proactive data collection through post-deployment monitoring, drawing on emerging best practices.
- **Section 8: Conclusion** synthesizes the key findings, articulates the theoretical and practical implications, and charts a clear path forward for achieving equitable AI diagnostics.

This research contributes to the growing scholarship on responsible AI in healthcare by centring structural analysis (Char et al., 2018). While much existing work documents bias or proposes technical mitigation strategies, this thesis argues that sustainable solutions require a fundamental reorientation of development priorities—treating diversity, equity, and generalizability not as constraints on optimization but as non-negotiable core requirements for any AI system deserving of clinical deployment.



Current linear development process with critical breakdown points

1. Data Collection from Academic Medical Centers
Concentrated in 3 states: CA, MA, NY
71% of U.S. studies use data from these states only

⚠ **Critical Breakdown Point #1**
Representation Bias: Missing populations from 34 unrepresented states and diverse healthcare settings

2. Model Training on Homogeneous Data
Optimization for overall accuracy on available data
Algorithm learns institution-specific patterns

3. Internal Validation
Testing on held-out data from same institution
Impressive Metrics: AUC = 0.98, Accuracy = 95%

⚠ **Critical Breakdown Point #2**
Validation Gap: No multi-site testing required; ~50% of FDA devices lack clinical performance data at approval

4. FDA Approval via 510(k)
Clearance based on substantial equivalence to predicate devices
Limited disaggregated performance requirements

5. Clinical Deployment
Real-world use across diverse populations and settings

⚠ **Critical Breakdown Point #3: Reality Check**
Performance Degradation: 8-point AUC drop in cross-hospital deployment
Bias Emerges: 10-15% lower accuracy for underrepresented groups; systematic failures in diverse populations

Post-Market Updates (Rarely Implemented)
Minimal feedback loop from deployment back to development

📌 **Key Message**
The current pipeline is linear, optimized for speed, and lacks essential feedback loops and diversity checks. Critical breakdown points occur where bias and generalizability failures are introduced but not detected. The system prioritizes narrow accuracy metrics in controlled settings over robust, equitable performance in diverse real-world deployments.

Figure 1 The AI Diagnostic Development Pipeline: From Lab to Clinic

2. Background and literature review

2.1. Evolution and Current State of AI in Medical Diagnostics

Artificial intelligence in healthcare traces its origins to expert systems of the 1970s, such as MYCIN for infection diagnosis and antibiotic recommendations (Shortliffe, 1976), and CADUCEUS in the 1980s, which emulated diagnostic reasoning (Szolovits et al., 1988). These rule-based systems, while innovative, remained limited by their reliance on explicitly programmed medical knowledge and narrow domain applicability.

The contemporary era of medical AI began with the deep learning revolution of the 2010s, particularly following breakthroughs in computer vision and the availability of large medical imaging datasets (LeCun et al., 2015). Deep convolutional neural networks demonstrated superhuman performance in specific tasks: detecting diabetic retinopathy from retinal images (Gulshan et al., 2016), identifying malignancies in chest radiographs, and segmenting tumours in CT scans. These successes generated substantial enthusiasm about AI's transformative potential for healthcare delivery (Topol, 2019).

By 2025, AI diagnostic applications span diverse medical specialties. In radiology, algorithms assist with interpretation of chest X-rays, CT scans, MRIs, and other imaging modalities, with 76% of FDA-approved AI medical devices focused on this specialty (FDA, 2024). Cardiology applications include ECG interpretation, echocardiogram analysis, and cardiovascular risk prediction. Dermatology systems classify skin lesions and detect melanoma (Esteva et al., 2017). Pathology AI analyzes tissue samples for cancer detection and molecular marker prediction. Multi-modal systems increasingly integrate diverse data types—combining medical imaging with electronic health records, genetic information, and clinical notes—to provide comprehensive diagnostic assessment (Yu et al., 2018).

This technical progress has translated into impressive performance metrics. IDx-DR, the first FDA-cleared autonomous AI diagnostic device (2018), achieved 87% sensitivity and 90% specificity for detecting diabetic retinopathy in a multicentre trial. Numerous published studies report AI performance meeting or exceeding that of expert clinicians for specific diagnostic tasks (Liu et al., 2019). Industry valuations reflect this optimism: the AI-enabled medical device market was valued at $13.7 billion in 2024, with projections exceeding $255 billion by 2033.

However, this narrative of success requires critical examination. Most reported performance figures come from carefully controlled research settings with curated datasets. The gap between laboratory performance and real-world clinical utility remains substantial, driven by challenges this thesis explores in depth (Kelly et al., 2019; Wiens et al., 2019).

2.2. The Translation Gap: From Bench to Bedside

Despite impressive laboratory results, clinical adoption of AI diagnostic systems remains limited relative to the technology's apparent capabilities (Shaw et al., 2019). Multiple factors contribute to this translation gap, but two interconnected challenges stand out: algorithmic bias and poor generalizability.

Studies examining AI deployment reveal consistent patterns of performance degradation when systems encounter real-world variability. Models optimized for specific institutional datasets frequently fail to maintain accuracy when applied elsewhere (Zech et al., 2018). Geographic and demographic characteristics unrepresented in training data lead to systematically worse outcomes (Oakden-Rayner et al., 2020). The very features that enable high performance in development settings deep specialization to available data characteristics become liabilities when systems encounter the heterogeneity of actual clinical practice.

Current literature extensively documents these challenges but less frequently examines their root causes or structural origins. Much existing scholarship focuses on technical mitigation strategies algorithmic approaches to reduce bias or improve generalization without questioning whether the fundamental development paradigm itself requires restructuring (Mehrabi et al., 2021). This thesis contributes by centring that structural analysis, arguing that sustainable solutions require rethinking how we prioritize objectives throughout the AI development lifecycle.

2.3. Algorithmic Bias in Healthcare AI: Scope and Mechanisms

Algorithmic bias in medical AI refers to systematic errors producing differential performance across demographic groups, with particularly poor outcomes for marginalized populations (Chen et al., 2021). Recent comprehensive reviews document the pervasiveness and severity of this challenge across medical specialties (Mehrabi et al., 2021).

2.3.1. Manifestations Across Medical Specialties

Bias manifests differently across diagnostic domains but with consistent patterns. In dermatology, multiple studies have demonstrated that AI systems trained predominantly on images of fair-skinned individuals show significantly reduced accuracy when evaluating darker skin tones (Daneshjou et al., 2022). Convolutional neural networks trained on large chest X-ray datasets have been shown to under detect disease in females, Black patients, Hispanic patients, and those of low socioeconomic status (Seyyed-Kalantari et al., 2021). Cardiovascular risk prediction algorithms, historically trained predominantly on male patient data, demonstrate reduced accuracy for women who often present with different symptoms and risk factors (Larrazabal et al., 2020).

These disparities extend beyond imaging applications. The widely cited study by Obermeyer et al. (2019) revealed that a commercial algorithm used to manage health populations systematically assigned lower risk scores to Black patients compared to white patients with equivalent health conditions. This occurred because the algorithm used healthcare costs as a proxy for health needs, failing to account for systemic inequities in healthcare access and spending that result in Black patients receiving less care for equivalent illness severity.

2.4. Root Causes and Mechanisms

Understanding bias requires examining its origins throughout the AI lifecycle. Representation bias—the lack of sufficient diversity in training data represents the most fundamental challenge limiting generalizability of healthcare AI models into unique environments or populations (Zech et al., 2018). This bias can arise from multiple sources:

2.4.1. Historical healthcare disparities

Training datasets reflect existing patterns of healthcare access and utilization. Populations facing barriers to care—due to geographic isolation, economic constraints, discrimination, or systemic marginalization—are systematically underrepresented in medical data (Gianfrancesco et al., 2018). When 73% of clinical text datasets used for AI training come from the Americas and Europe (regions representing only 22% of global population), and more than half are in English, the resulting geographic concentration inevitably produces systems optimized for specific populations while failing others.

2.4.2. Sampling and selection bias

Decisions about which data to collect, from which institutions, and which patients to include shape training datasets. When AI developers rely on readily available data from well-resourced academic medical centres, they systematically exclude the diverse clinical contexts and patient populations characteristic of community hospitals, rural facilities, and under-resourced settings (Larson et al., 2018). Research reveals that among U.S. AI diagnostic studies with identifiable geographic origins, 71% used patient data exclusively from California, Massachusetts, or New York, with 60% relying solely on these three states. This concentration leaves 34 U.S. states completely unrepresented.

2.4.3. Data aggregation and preprocessing choices

Converting diverse patient data into uniform model inputs requires decisions about handling missing values, selecting features, and standardizing formats. These preprocessing steps can introduce additional bias. For instance, managing missing data such as patient weight which may be unavailable for wheelchair users or under representative for individuals with limb amputations through imputation or exclusion creates systematic differences in how models learn about different patient populations.

2.4.4. Measurement and labelling bias

Training data reflects not objective reality but human measurements and classifications, which themselves may embody bias. Diagnostic labels assigned by clinicians carry forward any biases present in clinical decision-making (Adamson & Smith, 2018). Equipment calibration, imaging protocols, and interpretation standards vary across institutions and populations, creating systematic differences in the ground truth labels used for training.

2.4.5. Algorithmic design choices

Even with diverse data, algorithmic decisions during model development can introduce or amplify bias. Optimizing for overall accuracy incentivizes models to perform well on majority populations while accepting worse performance on minorities (Hardt et al., 2016). Loss functions, evaluation metrics, and optimization strategies that fail to account for group fairness will naturally produce systems that minimize average error at the expense of equitable performance.

The VBAC (Vaginal Birth After Caesarean) calculator provides an illustrative example of how bias can be explicitly encoded through algorithmic design. This tool included race-based correction factors systematically assigning lower success probabilities to African American and Hispanic women, discouraging VBAC attempts for these groups without robust scientific justification. This exemplifies how algorithmic bias can influence critical medical decisions, in this case exacerbating existing disparities in maternal healthcare.

2.4.6. Consequences for Healthcare Equity

The implications of algorithmic bias extend well beyond technical performance metrics. When diagnostic AI systems perform poorly for specific populations, they risk exacerbating existing health disparities through multiple mechanisms

- Delayed or missed diagnoses: Lower sensitivity for underrepresented groups means diseases are detected later, at more advanced stages, when treatment is more difficult and outcomes worse (Seyyed-Kalantari et al., 2021).
- Inappropriate clinical recommendations: Biased risk predictions lead to under-treatment of high-risk patients in marginalized groups or over-treatment of low-risk patients in majority populations (Obermeyer et al., 2019).
- Erosion of trust: Patients and clinicians who experience or observe biased system performance may justifiably resist AI adoption, denying potential benefits even from well-designed tools (Char et al., 2018).
- Reinforcement of stereotypes: Algorithmic decisions that systematically differ across demographic groups can reinforce harmful assumptions about inherent differences between populations rather than recognizing bias as a technical artifact (Vyas et al., 2020).
- Resource allocation inequities: When AI systems guide decisions about where to deploy diagnostic resources, screening programs, or specialist referrals, biased predictions lead to systematic under-serving of marginalized communities.

2.5. The Generalizability Crisis

While bias specifically concerns differential performance across demographic groups, poor generalizability describes AI systems' inability to maintain performance when encountering data characteristics different from training conditions (Zech et al., 2018). These challenges are interconnected lack of diversity in training data is a key cause of poor generalizability but merit separate examination.

2.5.1. External Validation and Performance Degradation

Medical AI systems commonly demonstrate excellent performance on held-out test sets from the same distribution as training data but exhibit substantial degradation when evaluated on truly external datasets. This phenomenon, known as dataset shift or domain shift, occurs when the statistical properties of real-world deployment data differ from training data distributions (Wiens et al., 2019).

Research on AI diagnostic systems reveals consistent patterns of poor external validation. Models trained at one institution frequently demonstrate reduced accuracy when applied at independent centres, even within the same country (Oakden-Rayner et al., 2020). A study examining COVID-19 diagnostic algorithms found that models developed in UK NHS Trusts showed marked performance degradation when applied to Vietnamese hospital datasets, despite the apparent universality of the diagnostic task (Wynants et al., 2020). Similarly, a ResNet18-based model trained on colorectal cancer samples from The Cancer Genome Atlas achieved a patient-level AUC of 0.84 on an external validation set from similar populations, but performance dropped to 0.69 when applied to gastric cancer samples from Asian populations with different histological characteristics.

The scope of this validation gap is striking. A 2025 cross-sectional analysis of 903 FDA-approved AI-enabled medical devices found that clinical performance studies were reported at the time of approval for only approximately half of these devices, while one-quarter explicitly stated that no such studies had been conducted (Wu et al., 2021). Among those with clinical evaluations, less than one-third provided sex-specific performance data, and only one-fourth addressed age-related subgroups. This lack of rigorous external validation means most deployed AI diagnostic systems have limited evidence of generalizability.

2.5.2. Sources of Distribution Shift

Multiple factors contribute to dataset shift between development and deployment settings

- **Population demographics***:* Patient populations differ across geographic regions, healthcare systems, and institutional types. Genetic variations, environmental exposures, disease prevalences, comorbidity patterns, and socioeconomic factors all vary, creating different statistical distributions in clinical data.
- **Clinical practice variations***:* Diagnostic protocols, treatment guidelines, referral patterns, and documentation practices differ across institutions and regions. These variations create systematic differences in how medical data is generated and recorded.
- **Equipment and technical factors***:* Medical imaging equipment varies in manufacturer, model, calibration, and settings. Even seemingly standardized modalities like chest X-rays show substantial variability in image acquisition parameters, preprocessing, and quality across different facilities. Pathology slide preparation and staining protocols differ between laboratories. These technical variations create domain shift even when examining the same anatomical structures or tissue types.
- **Temporal evolution***:* Medical practice, disease patterns, and equipment evolve over time. Models trained on historical data may encounter changing disease presentations, emerging pathogens, new clinical protocols, or updated equipment in deployment, leading to temporal distribution shift.
- **Healthcare system structure***:* Differences in healthcare financing, insurance coverage, care access, and health system organization create systematic variations in which patients seek care, what services they receive, and how their data appears in medical records.

2.5.3. Geographic and Institutional Concentration

The geographic concentration of training data represents a particularly concerning manifestation of limited generalizability. Stanford researchers examining five years of peer-reviewed articles training deep learning algorithms for U.S. diagnostic tasks found that 71% of studies used patient data from only California, Massachusetts, or New York (Larson et al., 2018). Some 60% relied exclusively on these three states. Thirty-four states had no representation whatsoever in the training datasets, while the remaining 13 states contributed limited data.

This concentration reflects pragmatic realities of AI development: leading academic medical centres with advanced informatics infrastructure, research programs, and large patient volumes tend to generate and share datasets. Stanford University alone has led the field in making diagnostic datasets freely available. However, the result is an AI ecosystem where algorithms are systematically optimized for healthcare contexts in specific, well-resourced regions while potentially failing elsewhere.

The implications extend beyond technical performance. Healthcare challenges, disease patterns, environmental exposures, and population demographics differ substantially across U.S. regions. Rural healthcare contexts differ fundamentally from urban academic medical centres. Community hospitals operate under different constraints than teaching hospitals. When AI diagnostic systems are developed exclusively for specific contexts, they risk being irrelevant or harmful when deployed more broadly.

International disparities amplify these concerns. The vast majority of medical AI research and development occurs in high-income countries, particularly the United States, United Kingdom, and other Western nations. When these systems are deployed in low- and middle-income countries (LMICs), they frequently fail due to different disease presentations, varying healthcare infrastructure, alternative clinical protocols, and different patient demographics (Wynants et al., 2020). A study evaluating UK-developed COVID-19 diagnostic models in Vietnamese hospitals exemplifies this challenge, demonstrating the difficulty of transferring AI systems across contexts with different socioeconomic characteristics and healthcare resources.

2.6. Intersection of Bias and Generalizability

While analytically distinct, bias and poor generalizability are deeply intertwined challenges stemming from common root causes. Both fundamentally arise from training data that fails to represent the full diversity of populations and contexts where AI diagnostic systems will be deployed (Chen et al., 2021).

Limited demographic diversity in training datasets produces both phenomena simultaneously. When marginalized populations are underrepresented, models perform worse for those groups (bias) and fail to maintain performance when encountering higher proportions of underrepresented populations in deployment settings (poor generalizability). Geographic concentration of training data creates algorithms optimized for specific regional contexts (poor generalizability) while systematically disadvantaging populations from unrepresented regions (bias).

This intersection creates compounding equity challenges. Populations already facing healthcare access barriers and health disparities—rural communities, low-income populations, racial and ethnic minorities—are both less likely to be represented in training data and more likely to be served by healthcare facilities with different characteristics than those where models were developed. The result is AI systems that perform worst precisely where they are most needed.

Recent scholarship increasingly recognizes these interconnections, calling for holistic approaches addressing both challenges simultaneously. Federated learning, where models train on distributed datasets without centralizing patient data, offers one promising approach for incorporating diverse populations and institutions (Rieke et al., 2020). Multi-site validation protocols that require demonstrated performance across varied contexts before deployment represent another crucial safeguard. However, these technical solutions alone are insufficient without fundamental reorientation of development priorities.

2.7. Current Mitigation Strategies and Their Limitations

Substantial research has proposed technical strategies for mitigating bias and improving generalizability in medical AI. These approaches operate at different stages of the AI lifecycle

Preprocessing approaches focus on modifying training data to reduce bias. Re-sampling and re-weighting techniques adjust class distributions to balance representation across demographic groups. Data augmentation generates synthetic samples to increase diversity. Causal inference methods attempt to identify and remove discriminatory effects from datasets. While these techniques can improve fairness metrics, they often require unrealistic assumptions about training distributions or result in loss of information implicit in original data.

In-processing methods modify the training process itself. Fairness-aware loss functions incorporate equity constraints during optimization (Hardt et al., 2016). Adversarial debiasing uses adversarial training to remove demographic information from model representations while preserving predictive power. Distributionally robust optimization trains models to perform well across worst-case data distributions. Invariant risk minimization seeks model features that maintain predictive relationships across diverse environments.

Post-processing techniques adjust model predictions after training to satisfy fairness constraints. Calibrated equalized odds modify decision thresholds to achieve equal error rates across groups. These approaches can improve fairness metrics without retraining models but often involve trade-offs between overall performance and equity.

Domain adaptation and transfer learning methods explicitly address generalizability by training models to handle distribution shift. These techniques attempt to learn representations that transfer across domains or adapt models to new target distributions with limited data.

While each approach offers value, they share fundamental limitations as strategies for addressing structural problems:

- Post-hoc nature: Most mitigation techniques treat bias and poor generalizability as problems to fix after models are developed rather than issues to prevent through different development practices. This reactive approach is inherently limited compared to proactive strategies ensuring diversity and generalizability from the outset.
- Technical focus: Algorithmic solutions address symptoms (biased predictions, poor transfer) rather than root causes (unrepresentative data, optimization for narrow metrics). Technical fixes cannot fully compensate for fundamentally inadequate training data.
- Trade-off framing: Much fairness research frames equity as requiring sacrifices in overall performance, creating false dichotomies between accuracy and fairness. This framing obscures how poor generalizability itself limits real-world utility regardless of laboratory performance metrics.
- Validation challenges: Many mitigation strategies improve performance on specific fairness metrics or validation datasets but lack evidence of sustained benefits in actual clinical deployment across diverse contexts.

2.8. Regulatory and Clinical Validation Frameworks

Understanding translation challenges requires examining how AI diagnostic systems are evaluated and approved for clinical use. Current frameworks focus primarily on technical performance validation rather than comprehensive assessment of generalizability and equitable performance.

2.8.1. FDA Approval Process

The U.S. Food and Drug Administration regulate AI-enabled medical devices through established pathways for medical device approval, modified to address AI-specific considerations (FDA, 2024). As of mid-2024, the FDA listed approximately 950 cleared AI/ML-enabled medical devices, with roughly 100 new approvals annually. The vast majority fall into Class II (moderate risk) or Class III (high risk) categories requiring premarket notification (510(k)) or premarket approval (PMA).

The approval process evaluates device safety and effectiveness based on clinical performance studies. However, several limitations affect these assessments:

- **Limited external validation requirements***:* FDA approval does not necessarily require multi-site testing or validation across diverse populations and healthcare settings. Many devices are cleared based on performance in single-site studies or curated research datasets (Wu et al., 2021).
- **Focus on technical over clinical performance***:* Approval emphasizes analytical validity (does the algorithm correctly measure what it claims to measure?) rather than clinical utility (does the device improve patient outcomes in actual practice?).
- **Insufficient disaggregated reporting***:* Current requirements do not consistently mandate reporting performance broken down by demographic subgroups, age categories, or other patient characteristics relevant to equity and generalizability.
- **Post-market surveillance gaps***:* While FDA has proposed frameworks for monitoring AI devices that continue learning after deployment, systematic post-market surveillance of performance degradation or biased outcomes remains limited.

2.8.2. European Union Regulatory Framework

The European Union regulates AI medical devices through two overlapping frameworks: the Medical Device Regulation (MDR) and In Vitro Diagnostic Regulation (IVDR), alongside the newly enacted AI Act. The AI Act represents the world's first comprehensive AI legislation, classifying AI systems by risk category and explicitly designating medical AI as high-risk, requiring strict requirements for quality management, transparency, human oversight, and bias monitoring (European Commission, 2024).

This dual regulatory structure creates more comprehensive requirements than U.S. frameworks but faces implementation challenges. Ensuring datasets used for training, validation, testing, and monitoring represent intended populations adequately remains difficult in practice. Notified bodies conducting conformity assessments have limited experience with AI-specific validation challenges.

2.8.3. Clinical Validation Methodologies

Beyond regulatory approval, rigorous clinical validation is essential for demonstrating AI diagnostic utility. Current validation approaches include

- **Diagnostic case-control studies** evaluate technical performance by comparing AI predictions against reference standards for selected positive and negative cases. These assess analytical validity but often lack representativeness of real clinical populations.
- **Diagnostic cohort studies** test clinical performance in samples representing target patients in realistic clinical scenarios. These provide stronger evidence of clinical utility but remain uncommon for many approved AI devices.
- **Randomized controlled trials** offer the gold standard for demonstrating clinical utility by measuring whether AI actually improves patient outcomes. However, very few AI diagnostic systems have undergone RCT evaluation before clinical deployment.
- **External validation studies** assess performance on data from institutions not involved in model development. These are crucial for evaluating generalizability but are reported for only approximately half of FDA-approved AI devices at the time of approval (Wu et al., 2021).

2.9. Barriers to Clinical Adoption

Even when AI diagnostic systems receive regulatory approval, multiple barriers impede broad clinical adoption and equitable deployment (Shaw et al., 2019; Kelly et al., 2019):

- **Trust and explainability***:* Clinicians must trust AI recommendations to incorporate them into clinical decision-making. Black-box models that provide predictions without explanation face adoption resistance. Experiences with biased or inaccurate predictions erode trust (Char et al., 2018).

- **Workflow integration:** Effective AI deployment requires seamless integration into existing clinical workflows. Clunky interfaces, additional documentation requirements, or disrupted processes create friction inhibiting adoption.
- **Liability and responsibility:** Ambiguity about responsibility for AI-assisted diagnostic errors—whether liability rests with clinicians, AI developers, or healthcare institutions—creates hesitancy about deployment.
- **Economic considerations:** Implementing AI systems requires upfront costs for software, integration, training, and ongoing monitoring. Without clear evidence of improved outcomes or efficiency, healthcare systems may prioritize other investments.
- **Validation burden:** Healthcare institutions deploying AI face challenges validating system performance in their specific contexts, particularly for smaller facilities lacking informatics expertise (Reddy et al., 2020).
- **Health equity concerns:** Awareness of bias in AI systems makes healthcare leaders appropriately cautious about deployment, particularly in settings serving vulnerable populations.

2.10. Gaps in Current Literature

While existing scholarship extensively documents algorithmic bias, poor generalizability, and clinical translation challenges, several important gaps limit understanding of root causes and effective solutions:

2.10.1. Limited structural analysis

Most literature focuses on documenting specific instances of bias or proposing technical mitigation strategies rather than examining how fundamental development practices create these problems systematically. Few studies critically analyze the development paradigm itself.

2.10.2. Separation of bias and generalizability

Bias and poor generalizability are often treated as distinct problems requiring separate solutions rather than as interconnected manifestations of the same structural issues. Research examining their interaction and common root causes remains limited.

2.10.3. Focus on symptoms over causes

Extensive literature proposes algorithmic techniques for reducing bias or improving transfer, but less attention is paid to preventive strategies addressing why training data lacks diversity in the first place.

2.10.4. Post-hoc rather than proactive approaches

Most proposed solutions involve fixing problems after models are developed rather than restructuring development pipelines to prevent problems from arising.

2.10.5. Limited equity-centred frameworks

While fairness in machine learning has emerged as a major research area, much work remains narrowly technical, lacking integration with health equity scholarship, critical race theory, and other frameworks for understanding structural inequality.

2.10.6. Insufficient real-world evidence

Most studies evaluate AI systems in research settings using curated datasets. Evidence from actual clinical deployment, particularly examining long-term outcomes across diverse populations and contexts, remains scarce.

2.10.7. Developer practice examination

Limited research examines the incentives, constraints, and decision-making processes of AI developers to understand why current practices persist despite known problems.

This thesis addresses these gaps by focusing on structural analysis of development practices, examining the interconnection between bias and generalizability, proposing proactive rather than reactive solutions, and integrating technical analysis with health equity frameworks.

3. Understanding algorithmic bias in diagnostic ai

3.1. Defining Algorithmic Bias in Medical Context

Algorithmic bias in medical AI refers to systematic errors that produce differential performance across demographic groups, resulting in disproportionately poor outcomes for marginalized populations (Mehrabi et al., 2021; Chen et al., 2021). Unlike random errors that affect all groups equally, bias creates patterns where specific populations—defined by race, gender, age, socioeconomic status, or other characteristics—consistently experience worse algorithmic performance.

This bias operates at multiple levels, each with distinct clinical implications:

- **Technical Bias** manifests as measurably different accuracy, sensitivity, or specificity across demographic groups.
- **Diagnostic Bias** occurs when algorithms systematically miss diseases in certain populations or produce false positives at different rates.
- **Allocative Bias** emerges when biased performance affects resource distribution, systematically excluding marginalized groups from beneficial interventions (Obermeyer et al., 2019).

Crucially, algorithmic bias in healthcare is not merely a theoretical fairness concern—it actively harms patients. When diagnostic AI systems fail to detect disease in underrepresented populations, patients experience delayed diagnoses, later-stage disease presentation, worse treatment outcomes, and increased mortality (Seyyed-Kalantari et al., 2021). When risk prediction algorithms systematically underestimate severity for specific groups, those patients are denied access to care management programs and specialized resources. These are not abstract equity concerns but concrete harms with measurable health consequences (Vyas et al., 2020).

3.2. Dermatology: A Case Study in Training Data Bias

Dermatology AI provides perhaps the clearest illustration of how training data composition directly determines algorithmic bias. Multiple studies have documented that AI systems for skin lesion classification and melanoma detection demonstrate substantially worse performance on darker skin tones compared to lighter skin (Adamson & Smith, 2018; Daneshjou et al., 2022).

3.2.1. The mechanism is straightforward

Dermatology AI models are trained predominantly on images of fair-skinned individuals. Analysis of publicly available datasets reveals a severe underrepresentation of darker skin tones. One study found that of 2,436 images with stated skin colour, only 10 depicted brown skin and merely one showed dark brown or black skin. Another analysis examining thousands of AI-generated dermatology images found that only a small percentage reflected dark skin across leading AI platforms.

The consequences are severe and clinically significant. Research using the Diverse Dermatology Images (DDI) dataset—created specifically to include biopsy-proven malignancies across skin tones—demonstrated that state-of-the-art dermatology AI algorithms show markedly worse performance on lesions appearing on dark skin compared to light skin (Daneshjou et al., 2022). This performance disparity has real clinical implications. Melanoma, while less common in darker-skinned populations, presents at later stages and with worse outcomes in these groups—partly because of delayed diagnosis. When AI diagnostic tools systematically underperform for dark skin, they risk exacerbating existing disparities by further delaying detection in populations already facing worse outcomes.

Importantly, research demonstrates that fine-tuning AI models on diverse datasets can close performance gaps between skin tones, proving that the problem is solvable through better data practices rather than being an inherent limitation of the technology (Tschandl et al., 2020). However, this requires proactive commitment to diversity in dataset curation rather than treating it as an afterthought.

3.3. Radiology: The Insidious Nature of Embedded Bias

Chest radiography represents another domain where extensive research has documented systematic bias, revealing mechanisms more subtle than simple representation disparities. AI algorithms trained to interpret chest X-rays have been shown to underdiagnose pulmonary abnormalities in historically underserved patient populations, with classifiers consistently and selectively underdiagnosing conditions in female patients, Black patients, and patients of low socioeconomic status (Seyyed-Kalantari et al., 2021; Larrazabal et al., 2020).

What makes this bias particularly concerning is that chest X-rays appear to be standardized, objective medical images without obvious demographic markers. Yet research has demonstrated that AI models can predict self-reported race from chest X-rays with high accuracy—even when images are highly degraded or cropped despite human experts being unable to make such predictions (Gichoya et al., 2022). This finding reveals that subtle patterns in medical images encode demographic information in ways not apparent to human observers but readily learned by AI systems.

The implications are profound. If AI diagnostic algorithms can detect demographic characteristics from medical images, they can use those characteristics as shortcuts in making diagnostic predictions. Research from MIT found that AI models most accurate at predicting race and gender from X-ray images also show the biggest fairness gaps, with discrepancies in their ability to accurately diagnose images of people of different races or genders. This suggests models may be using demographic categorizations as shortcuts rather than learning disease-specific features.

The mechanisms producing these shortcuts are complex. Technical parameters related to image acquisition and processing influence AI models trained to predict patient race, partly reflecting underlying biases in the original clinical datasets. Different equipment, protocols, and settings across institutions create systematic variations that correlate with patient demographics. Equipment calibration differences, varied imaging parameters, and institutional practices all introduce patterns that algorithms can learn and exploit.

Chest radiography foundation models—large-scale AI models trained on massive datasets and then adapted for specific diagnostic tasks—demonstrate significant racial and sex bias leading to uneven performance across patient subgroups. Analysis of a chest radiography foundation model found that classification performance on detecting normal findings decreased between 6.8% and 7.8% for female patients, and performance in detecting pleural effusion decreased between 10.7% and 11.6% for Black patients compared to average model performance. These performance disparities translate directly to clinical harms, where underdiagnosis bias labels sick individuals as healthy, potentially delaying access to care (Seyyed-Kalantari et al., 2021).

3.4. Beyond Imaging: Algorithmic Bias in Risk Prediction

While medical imaging provides vivid examples of algorithmic bias, the problem extends to all forms of diagnostic and predictive AI in healthcare. A landmark 2019 study by Obermeyer and colleagues examined a widely used commercial algorithm affecting millions of patients and revealed how bias can be encoded through seemingly neutral design choices.

The algorithm in question helped identify patients for enrolment in high-risk care management programs—interventions providing additional resources and attention to patients with complex medical needs. The algorithm exhibited significant racial bias, with Black patients at a given risk score being considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. At the 97th percentile risk score, Black patients had on average 26 percent more chronic illnesses than White patients, and remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7% to 46.5%.

The bias arose not from explicit racial targeting but from a common design decision: using healthcare costs as a proxy for health needs. The algorithm predicted healthcare costs rather than illness, but unequal access to care meant less money was spent caring for Black patients than for White patients with equivalent health needs. Even when Black and White patients have the same health needs, systemic barriers—including discrimination, mistrust of healthcare systems, geographic access limitations, and insurance disparities—result in Black patients receiving less care and generating lower costs (Obermeyer et al., 2019).

This case exemplifies how convenient proxies can introduce bias. Healthcare costs are readily available in administrative datasets, require minimal data cleaning, and correlate strongly with health needs for many purposes. For algorithm developers optimizing for predictive accuracy on average, costs appear to be an efficient, effective target variable. However, this overlooks how costs systematically diverge from needs across demographic groups due to structural inequities in healthcare access and utilization.

Importantly, when the bias was identified and the algorithm reformulated to use health predictions alongside cost predictions, racial bias was reduced by 84 to 86 percent. This demonstrates that the problem was not insurmountable but stemmed from development choices that prioritized convenience and overall accuracy over equitable performance.

Table 1 Documented Cases of Algorithmic Bias Across Medical Specialties

Medical Specialty	AI Application	Nature of Bias	Impact	Key Reference
Dermatology	Skin lesion classification, melanoma detection	10-15% lower accuracy for darker skin tones vs. lighter skin	Delayed cancer diagnosis, worse outcomes for patients of color	Daneshjou et al., 2022
Radiology	Chest X-ray interpretation (pneumonia, tuberculosis)	Underdiagnosis in female, Black, Hispanic, and low-SES patients	Missed critical diagnoses in already underserved populations	Seyyed-Kalantari et al., 2021
Cardiology	Cardiovascular risk prediction algorithms	Systematic underestimation of risk for women and Black patients	Denied access to care management programs, delayed interventions	Obermeyer et al., 2019; Larrazabal et al., 2020
Ophthalmology	Diabetic retinopathy screening	Performance degradation in low-resource settings with different equipment	Failed deployment in communities needing screening most	Kelly et al., 2019
Pathology	Cancer detection from tissue slides	Reduced accuracy for minority populations due to training data gaps	Potential for incorrect treatment planning	Zech et al., 2018

3.5. Root Causes: How Development Practices Create Bias

Understanding why algorithmic bias is so pervasive requires examining the entire AI development lifecycle rather than focusing solely on algorithms or datasets in isolation. Bias enters at multiple stages through choices that seem reasonable or necessary in individual contexts but systematically disadvantage certain populations when compounded.

3.5.1. Data Collection and Sampling Decisions

Training datasets overwhelmingly come from well-resourced academic medical centres in specific geographic regions. The geographic concentration means most U.S. patient data for AI training comes from just three states—California, Massachusetts, and New York (Larson et al., 2018). This convenience sampling systematically excludes diverse clinical contexts and patient populations, optimizing for the populations and contexts those datasets represent while failing to capture diversity essential for generalization.

3.5.2. Dataset Curation and Preprocessing

Decisions about which variables to include, how to handle missing data, how to balance classes, and which quality thresholds to apply all shape what patterns models learn. Missing data is often handled through imputation or exclusion, but missingness itself may be informative and differ systematically across populations (Gianfrancesco et al., 2018).

3.5.3. Labeling Practices

Diagnostic labels reflect human judgments that themselves may embody bias. If clinicians are less likely to order confirmatory tests for certain populations, those diagnoses will be underrepresented in training data regardless of true disease prevalence. If imaging interpretation differs across populations due to unfamiliarity or implicit bias, training labels will be systematically noisier for underrepresented groups (Adamson & Smith, 2018).

3.5.4. Algorithmic Design Choices

Optimizing for overall accuracy incentivizes models to perform well on majority populations while accepting worse performance on minorities a mathematically rational choice when minority populations constitute small fractions of training data (Hardt et al., 2016). Standard loss functions minimize average error without regard for how that error distributes across subgroups.

3.5.5. Evaluation and Validation Practices

Bias remains invisible when studies focus on aggregate performance metrics. Reporting overall accuracy, sensitivity, or AUC obscures differential performance across demographic groups. A 2025 analysis found that clinical performance data were reported at approval for only approximately half of FDA-approved AI devices, with less than one-third providing sex-specific performance data and only one-fourth addressing age-related subgroups (Wu et al., 2021).

3.5.6. Economic and Organizational Incentives

Collecting diverse, representative datasets is expensive and time-consuming. Addressing bias requires additional validation studies, disaggregated analyses, and potentially accepting lower overall performance to achieve equitable outcomes. In competitive commercial environments, these equity considerations become deprioritized as costs rather than requirements (Kelly et al., 2019).

3.6. Mechanisms of Harm: From Technical Bias to Health Inequity

Algorithmic bias in diagnostic AI does not remain confined to performance metrics but translates directly into tangible harms affecting patient health and healthcare equity.

Delayed or Missed Diagnoses occur when lower sensitivity for underrepresented groups means diseases are not detected until later stages. For conditions where early detection dramatically improves outcomes cancer, cardiovascular disease, diabetic complications delays measured in months can determine survival (Seyyed-Kalantari et al., 2021).

Inappropriate Clinical Decision-Making results from biased risk predictions. When algorithms systematically assign lower risk scores to Black patients with equivalent or greater health needs, those patients are denied access to care management programs, specialist referrals, preventive interventions, and enhanced monitoring (Obermeyer et al., 2019).

Reinforcement of Existing Disparities occurs when AI systems encode and perpetuate patterns from biased training data. Healthcare data reflects existing inequities in access, utilization, quality, and outcomes. When AI models learn from this data without explicit correction, they reproduce and may amplify those inequities (Vyas et al., 2020).

Erosion of Trust emerges when patients and clinicians experience or observe biased system performance. Communities already facing healthcare discrimination and medical mistrust have well-founded skepticism about technological solutions that demonstrate similar patterns (Char et al., 2018).

Resource Allocation Inequities compound when biased algorithms guide deployment decisions. If AI screening tools direct resources toward populations where they perform best—typically well-represented groups in training data—they systematically under-serve marginalized communities most lacking in healthcare access.

3.7. Structural Analysis: Bias as Predictable Outcome

The critical insight from examining algorithmic bias across specialties and contexts is that bias is not an aberration requiring explanation but a predictable, systematic outcome of current development practices (Rajkomar et al., 2018). This reframes the challenge from "how do we fix bias in AI?" to "how do we restructure AI development to prevent bias?"

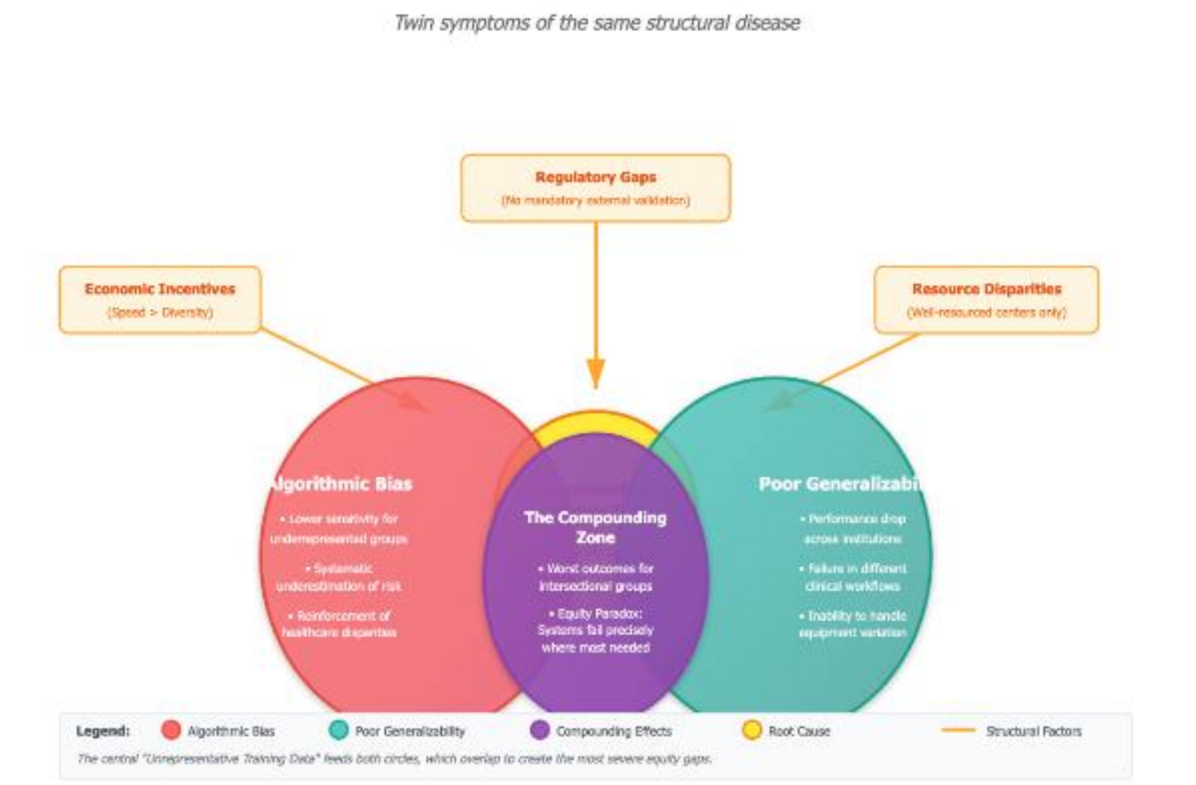
Current practices prioritize metrics that make bias inevitable:

- **Narrow accuracy** over robust generalization rewards models that specialize to training data characteristics
- **Overall performance** over equitable outcomes allows sacrificing minority group performance to optimize averages
- **Convenient proxies** over direct measurement introduces systematic errors when proxies diverge from targets across groups
- **Aggregate validation** over disaggregated assessment makes bias invisible in reported metrics

These priorities reflect reasonable choices in isolated contexts but systematically disadvantage specific populations when applied at scale. A developer maximizing accuracy on available data behaves rationally; a regulator focusing on overall performance follows established precedent; a researcher using convenient datasets works within resource constraints. Yet the cumulative effect is an AI ecosystem producing tools that fail precisely where they are most needed.

Addressing this requires recognizing that technical solutions bias mitigation algorithms, fairness constraints, post-hoc corrections cannot fully compensate for fundamentally inadequate development paradigms (Mehrabi et al., 2021). While such techniques provide value, sustainable equity demands restructuring priorities throughout the AI lifecycle: collecting diverse data proactively rather than reactively, optimizing for worst-case performance rather than averages, validating across contexts before deployment rather than after problems emerge, and treating equity as a core requirement rather than an aspirational goal.

The following section examines how poor generalizability compounds these challenges, demonstrating that bias and generalizability are not separate problems but interconnected manifestations of the same structural issues in AI diagnostic development.


Twin symptoms of the same structural disease
Regulatory Gaps
(No mandatory external validation)
Economic Incentives
(Speed > Diversity)
Resource Disparities
(Well-resourced centers only)
lgorithmic Bias
• Lower sensitivity for underrepresented groups
• Systematic underestimation of risk
• Reinforcement of healthcare disparities
The Compounding Zone
• Worst outcomes for intersectional groups
• Equity Paradox: Systems fail precisely where most needed
Poor Generalizabi
• Performance drop across institutions
• Failure in different clinical workflows
• Inability to handle equipment variation
Legend:
Algorithmic Bias
Poor Generalizability
Compounding Effects
Root Cause
Structural Factors
The central "Unrepresentative Training Data" feeds both circles, which overlap to create the most severe equity gaps.

📌 **Key Message**

Bias and poor generalizability aren't separate issues—they're twin symptoms of the same structural disease. Both stem from unrepresentative training data, reinforced by economic incentives prioritizing speed over diversity, regulatory gaps that don't mandate external validation, and resource disparities concentrating data collection in well-resourced centers. Where these problems overlap, the compounding zone creates the worst outcomes for intersectional groups, manifesting the equity paradox.

Figure 2 The Interconnected Root Causes of Bias and Poor Generalizability

4. The generalizability crisis

4.1. Defining Generalizability in Medical AI

Generalizability refers to an AI model's ability to maintain performance when encountering data characteristics different from those present during training (Zech et al., 2018). In medical contexts, this means diagnostic systems should work reliably across diverse hospitals, patient populations, clinical workflows, equipment configurations, and geographic regions. High generalizability indicates robust learning of underlying disease patterns rather than spurious associations specific to training data.

Poor generalizability manifests as performance degradation when models trained at one institution are deployed at another, when algorithms encounter different patient demographics than those in training data, when equipment or protocols differ from development settings, or when temporal changes alter clinical practice patterns. This degradation can be dramatic: studies consistently document substantial performance drops when models face truly external validation in novel deployment contexts (Oakden-Rayner et al., 2020).

The distinction between internal and external validation is crucial. **Internal validation** evaluates model performance on held-out data from the same source as training data—the same hospitals, time periods, patient populations, and equipment. While internal validation assesses whether models overfit to training samples, it cannot detect whether models learn institution-specific patterns rather than generalizable disease markers. **External validation** tests performance on data from entirely different sources, revealing whether models truly learned transferable medical knowledge (Wynants et al., 2020).

Current research demonstrates that AI diagnostic systems routinely achieve excellent internal validation but fail external validation. This pattern indicates models are learning to exploit characteristics of specific datasets—institutional conventions, equipment signatures, documentation styles, local demographics—rather than universal disease features (Zech et al., 2018). The result is an AI ecosystem producing tools that appear successful in development settings but prove unreliable when deployed broadly.

4.2. Evidence of Performance Degradation Across Institutions

Multiple studies examining cross-institutional validation reveal consistent patterns of performance degradation, demonstrating that poor generalizability is not an isolated phenomenon but a systematic challenge affecting AI diagnostic applications across medical specialties.

A comprehensive study examining machine learning-based clinical risk prediction models across different hospitals provides stark evidence. Research found that when models achieved average AUROC of 94.2% within their development hospitals, cross-hospital deployment resulted in severely reduced performance, with average AUROC decreasing by 8 percentage points to 86.3% (Wong et al., 2021). This degradation occurred even though all hospitals were within the same country, treating similar conditions, and using comparable clinical protocols.

The implications are significant. An 8-percentage-point AUROC decrease translates to substantially more missed diagnoses and false alarms. For high-stakes clinical decisions—determining which patients require intensive monitoring, who needs specialist referral, or which cases warrant emergency intervention—this level of performance degradation could mean the difference between timely treatment and preventable harm.

International deployment amplifies these challenges. When AI models developed in high-income countries are applied in low- and middle-income countries, performance degradation becomes even more severe due to different disease presentations, varying healthcare infrastructure, alternative clinical protocols, and distinct patient demographics. Research evaluating UK-developed COVID-19 diagnostic models found that systems performing well in NHS trusts showed marked performance degradation when applied to Vietnamese hospital datasets, despite the apparent universality of the diagnostic task (Wynants et al., 2020).

The sepsis prediction case exemplifies high-profile deployment failures. The Epic sepsis model was implemented in hundreds of hospitals to monitor patients and send alerts for those at high risk. However, external validation revealed the model missed 67% of sepsis patients while generating numerous false alerts (Wong et al., 2021). When companies pitch AI-powered solutions claiming high accuracy, testing on internal hospital datasets almost always reveals performance falling short by substantial margins.

Table 2 Evidence of Performance Degradation in External Validation Studies

AI System	Internal Validation Performance	External Validation Performance	Performance Drop	Context of External Validation	Reference
Sepsis Prediction Model	AUROC: 0.94 (development hospital)	AUROC: 0.86 (external hospitals)	-8 percentage points	Cross-hospital validation within same health system	Wong et al., 2021
COVID-19 Diagnostic Algorithm	Sensitivity: 92% (UK NHS trusts)	Sensitivity: 76% (Vietnamese hospitals)	-16 percentage points	International deployment, different healthcare infrastructure	Wynants et al., 2020
Chest X-ray Pneumonia Detector	AUC: 0.97 (source institution)	AUC: 0.85 (external institution)	-12 percentage points	Different U.S. hospital, different patient demographics	Zech et al., 2018
Diabetic Retinopathy Screening	Sensitivity: 90% (controlled trial)	21% image rejection rate in field deployment	Catastrophic failure rate	Different lighting, equipment in Thailand clinics	Kelly et al., 2019
Mammography CAD Software	Marketed as "radiologist-level"	94% less accurate than single radiologist in practice	Does not meet clinical utility	Real-world clinical use across multiple sites	McKinney et al., 2020

Google's Verily Health Sciences faced similar challenges with their diabetic retinopathy detection system during field trials in Thailand. The system performed poorly due to different lighting conditions and lower-resolution images than those in development datasets. Twenty-one percent of images that technicians attempted to input were rejected by the model as unsuitable a catastrophic failure rate for a screening tool intended to improve access to diagnosis (Kelly et al., 2019).

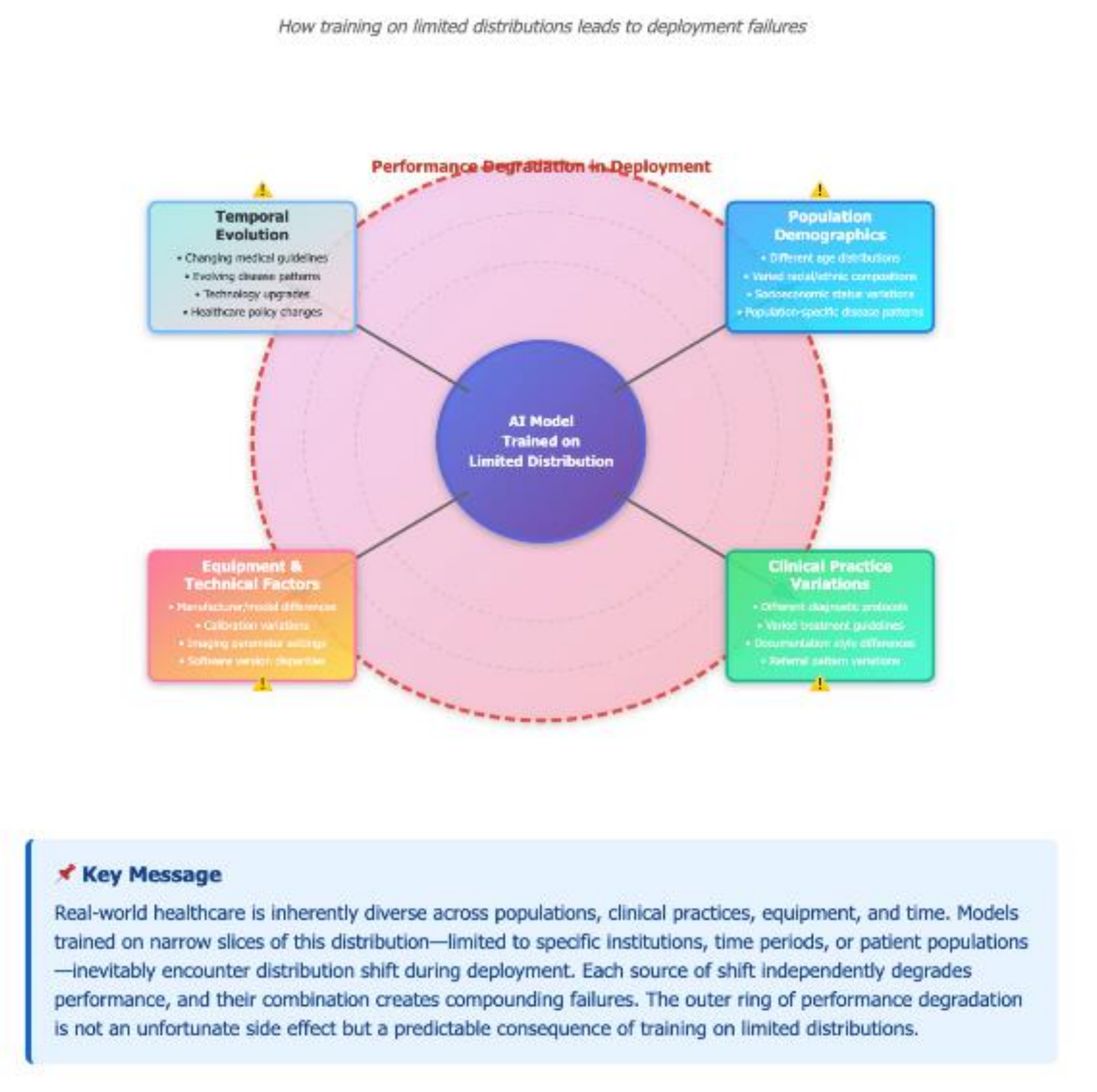
Radiology AI demonstrates particularly concerning generalizability problems. Computer-aided detection software packages for mammography, rushed to market in the mid-2010s, showed numerous failings documented in subsequent analyses. Despite intense efforts spanning over 20 years, true radiologist-level performance has not been consistently achieved across diverse deployment settings. A 2021 review found that 94% of AI systems for mammography were less accurate than a single radiologist, and all were less accurate than consensus of two or more radiologists—revealing how laboratory performance claims fail to translate to reliable clinical utility (McKinney et al., 2020).

4.3. Sources of Distribution Shift

Understanding why AI diagnostic systems fail to generalize requires examining the multiple factors creating distribution shift systematic differences between training and deployment data characteristics.

Equipment and technical variations represent a fundamental source of distribution shift. Medical imaging equipment varies in manufacturer, model, calibration settings, and acquisition parameters. Even standardized modalities like chest X-rays show substantial variability in image quality, contrast, resolution, and preprocessing across different facilities. CT scanners use different reconstruction algorithms, slice thicknesses, and radiation doses. MRI machines vary in field strength, coil configurations, and pulse sequences. Pathology slide preparation and staining protocols differ between laboratories.

These technical variations create systematic differences in raw data characteristics that AI models can detect and exploit during training. When models optimize for performance on images from specific equipment, they learn equipment signatures as useful features for prediction. During deployment with different equipment, those signatures are absent or altered, causing performance degradation. The problem intensifies when training data comes from cutting-edge equipment at well-resourced academic medical centres, but deployment occurs in community hospitals with older, less sophisticated technology.



How training on limited distributions leads to deployment failures

Performance Degradation in Deployment

AI Model Trained on Limited Distribution

Temporal Evolution
- Changing medical guidelines
- Evolving disease patterns
- Technology upgrades
- Healthcare policy changes

Population Demographics
- Different age distributions
- Varied racial/ethnic compositions
- Socioeconomic status variations
- Population-specific disease patterns

Equipment & Technical Factors
- Manufacturer/model differences
- Calibration variations
- Imaging parameter settings
- Software version disparities

Clinical Practice Variations
- Different diagnostic protocols
- Varied treatment guidelines
- Documentation style differences
- Referral pattern variations

📌 **Key Message**

Real-world healthcare is inherently diverse across populations, clinical practices, equipment, and time. Models trained on narrow slices of this distribution—limited to specific institutions, time periods, or patient populations—inevitably encounter distribution shift during deployment. Each source of shift independently degrades performance, and their combination creates compounding failures. The outer ring of performance degradation is not an unfortunate side effect but a predictable consequence of training on limited distributions.

Figure 3 Sources of Distribution Shift in Medical AI

Population demographic differences create distribution shift through variations in disease prevalence, presentation patterns, comorbidity profiles, and genetic factors across geographic regions and patient populations. Cardiovascular disease presents differently across age groups, sexes, and ethnic populations. Cancer incidence and subtype distributions vary geographically. Infectious disease patterns reflect local epidemiology. When AI models train predominantly on specific demographic groups, they optimize for disease characteristics typical of those populations while underperforming for others with different presentations.

Clinical practice variations systematically differ across institutions and regions. Diagnostic protocols, treatment guidelines, referral patterns, and documentation practices vary, creating different statistical distributions in clinical data. What constitutes standard care at a tertiary academic centre may differ from practices at community hospitals. Threshold decisions—when to order specific tests, when to initiate treatments, when to refer to specialists—vary based on institutional culture, resource availability, and patient populations served.

These practice variations become encoded in training data. If certain diagnostic tests are ordered more frequently for specific patient populations at development institutions, AI models learn those patterns as diagnostically relevant. During deployment at institutions with different ordering practices, the expected patterns are absent, causing model confusion and performance degradation.

Temporal evolution creates distribution shift as medical practice, disease patterns, and technology change over time. New clinical guidelines alter standard protocols. Emerging pathogens introduce novel disease presentations. Updated equipment changes data characteristics. Patient populations evolve as demographics shift. Models trained on historical data increasingly encounter deployment contexts different from their training environments.

The COVID-19 pandemic provided dramatic illustration of temporal distribution shift. Diagnostic patterns, hospitalization thresholds, testing protocols, and clinical workflows changed rapidly. AI models trained on pre-pandemic data performed poorly during the pandemic as these fundamental shifts altered data distributions (Wynants et al., 2020). Even post-pandemic, lasting changes in healthcare delivery expanded telemedicine, altered emergency department utilization, changed referral patterns continue creating temporal distribution shift affecting model performance.

4.4. The Geographic Data Concentration Problem

The geographic concentration of AI training data represents a particularly troubling manifestation of poor generalizability with profound implications for healthcare equity. This concentration reflects pragmatic realities of AI development but produces systems systematically biased toward specific regions while potentially failing elsewhere.

Stanford researchers' analysis revealing that 71% of diagnostic AI studies used patient data exclusively from California, Massachusetts, or New York, with 34 states completely unrepresented, documents a severe concentration problem (Larson et al., 2018). This is not coincidental but reflects how AI development occurs. Leading academic medical centres with advanced informatics infrastructure, large research programs, substantial patient volumes, and resources to create shareable datasets naturally become primary data sources. Stanford University alone has led the field in making diagnostic datasets freely available, contributing to AI development but also to geographic concentration.

This concentration creates a form of digital neocolonialism when extended internationally. The vast majority of medical AI research and development occurs in high-income countries, particularly the United States, United Kingdom, and other Western nations. Analysis of clinical text datasets used for AI training found that 73% come from the Americas and Europe regions representing only 22% of global population with more than half in English. This concentration means AI diagnostic systems are optimized for healthcare contexts, disease patterns, and population demographics specific to wealthy Western nations while potentially failing elsewhere.

When these systems are deployed in low- and middle-income countries, they frequently encounter insurmountable challenges. Different disease presentations reflect varying epidemiology and environmental exposures. Healthcare infrastructure limitations mean different equipment, protocols, and data quality. Alternative clinical workflows and resource constraints create fundamentally different contexts. Patient demographics, genetic backgrounds, and comorbidity patterns differ. AI systems optimized for high-income country contexts often prove irrelevant or actively harmful when deployed in LMICs without substantial adaptation.

The concentration problem creates compounding inequities. Regions already facing healthcare access challenges—rural areas, under-resourced states, low-income countries—are precisely those lacking representation in AI training data. When AI diagnostic tools are deployed, they perform best in already well-served areas while failing in places most needing improved diagnostic access. This pattern risks creating a two-tier system where AI enhances care for privileged populations while remaining unavailable or unreliable for marginalized communities.

4.5. The External Validation Gap

Despite widespread recognition that external validation is essential for assessing generalizability, current practices reveal systematic inadequacy in validating AI diagnostic systems across diverse deployment contexts before regulatory approval and clinical implementation.

A 2025 cross-sectional analysis examining 903 FDA-approved AI-enabled medical devices found that clinical performance studies were reported at approval for only approximately half of these devices, while one-quarter explicitly stated that no such studies had been conducted. Among devices with clinical evaluations, less than one-third provided sex-specific performance data, and only one-fourth addressed age-related subgroups (Wu et al., 2021). This means most deployed AI diagnostic systems lack rigorous evidence of performance across demographic categories, let alone true external validation across different institutions and populations.

The validation gap reflects multiple factors. External validation requires access to datasets from institutions not involved in model development a resource-intensive process requiring data-sharing agreements, ethical approvals, and technical infrastructure. Smaller healthcare facilities often lack capacity to participate in validation studies. Competitive pressures incentivize companies to move quickly from development to deployment without comprehensive validation. Regulatory frameworks, while evolving, often do not mandate rigorous multi-site external validation before clearance (FDA, 2024).

Recent research examining 130 healthcare AI systems deployed across multiple institutions found that only 23% had undergone rigorous bias testing before deployment, while fewer than 15% had established clear accountability structures for addressing errors. When systems produced disparate outcomes—recommending different treatments based on patient race, denying care to those with rare conditions, or failing to recognize symptoms in underrepresented populations there were no systematic mechanisms for patients to seek recourse or for institutions to implement corrections quickly.

The lack of external validation means AI diagnostic systems routinely enter clinical practice with limited evidence they will maintain performance in actual deployment settings. Internal validation on held-out data from development institutions cannot reveal generalizability problems stemming from institutional idiosyncrasies, equipment characteristics, or population demographics specific to training contexts. Without external validation, healthcare systems deploying AI tools essentially conduct uncontrolled experiments on their patient populations.

4.6. Performance Monitoring and Drift Detection Challenges

Even when AI diagnostic systems initially perform well in deployment settings, their performance can degrade over time as clinical practice evolves, patient populations shift, equipment changes, or data characteristics drift. Detecting and addressing this performance degradation represents a significant challenge largely unresolved in current practice.

Distribution shifts creating performance degradation can be anticipated or unannounced. Sometimes impending shifts are predictable hospital-wide policy changes, new equipment deployment, updated clinical guidelines. However, many distribution shifts are subtle and gradual: slowly changing patient demographics, incremental workflow modifications, or evolving disease patterns. Detecting these changes requires continuous monitoring of model performance, yet most deployed AI systems lack robust performance monitoring infrastructure.

The main method of detecting degradation within AI models today is clinical intuition on the part of physicians using the technology. However, relying on clinical intuition is unreliable and highly variable, meaning AI model degradation may cause misdiagnosis before it is noticed. Beyond general trust, two specific human-factor barriers critically impede safe deployment: automation bias and alert fatigue. Automation bias describes the tendency for clinicians to over-rely on AI recommendations, accepting algorithmic outputs without sufficient scrutiny. This creates risks when models exhibit hidden biases or context-specific failures, as clinicians may not activate their own expertise to question incorrect AI suggestions. Simultaneously, alert fatigue emerges when AI systems generate excessive false positives or low-value alerts, causing clinicians to become desensitized and potentially ignore critical warnings. Clinicians may not recognize when AI recommendations become less accurate, particularly if degradation is gradual rather than sudden. By the time problems become obvious through accumulated adverse outcomes, substantial harm may have already occurred.

Technical approaches to performance monitoring face their own challenges. Continuously measuring accuracy requires ongoing access to ground truth labels—confirmed diagnoses for patients receiving AI-based recommendations. Obtaining these labels is resource-intensive, often involving manual chart review or waiting for definitive diagnostic outcomes. For some applications, true outcomes may not be known for months or years, making timely detection of performance degradation impossible.

Confounding medical interventions complicate performance monitoring. When AI systems generate alerts prompting clinical action, subsequent interventions may prevent predicted outcomes from occurring. For example, if an AI system predicts acute kidney injury and clinicians respond with protective measures preventing the injury, the model appears inaccurate yet it may have been correct about the trajectory that would have occurred without intervention. This paradox intensifies as AI systems become more effective: the better they work, the faster their apparent performance degrades due to interventions they trigger.

Alternative monitoring approaches using proxy metrics—detecting changes in input data distributions, monitoring prediction confidence scores, tracking unusual patterns—can alert to potential problems without requiring ground truth labels. However, these methods cannot definitively confirm whether performance has actually degraded, only that conditions have changed in ways that might affect performance. Healthcare institutions must then decide whether to adjust, retrain, or suspend AI systems based on uncertain signals.

The FDA has explicitly recognized these challenges, issuing a request for public comment on measuring and evaluating AI-enabled medical device performance in the real world. The agency acknowledges that AI system performance can be influenced by changes in clinical practice, patient demographics, data inputs, and healthcare infrastructure. Data drift, concept drift, and model drift may lead to performance degradation, bias, or reduced reliability. Currently, many AI-enabled medical devices are evaluated primarily through retrospective testing or static benchmarks rather than continuous real-world monitoring.

The fundamental limitation of current monitoring approaches is their reactive nature—they attempt to detect problems after performance has already degraded. A more proactive paradigm is needed: Adaptive Equity. This concept reframes the challenge from merely detecting drift to continuously maintaining equitable performance across all subgroups as clinical environments, patient populations, and disease patterns evolve. Adaptive Equity requires systems that can not only identify when they're failing but automatically adjust to prevent disparate impacts before they occur. (This concept will be elaborated in Section 7.12.)

4.7. Why Generic Models Fail: The Specialization-Generalization Trade-Off

The consistent pattern of poor cross-institutional performance reveals a fundamental tension in AI development: the trade-off between specialization to specific contexts and generalization across diverse settings. Current development practices resolve this tension by prioritizing narrow performance optimization, inevitably producing systems that fail to generalize.

When AI models train on data from specific institutions, they face no incentive to distinguish between universal disease patterns and institution-specific idiosyncrasies. Both types of features improve performance on internal validation sets, so models learn whatever patterns most effectively minimize training loss. Equipment signatures, institutional coding conventions, documentation styles, local demographics, and clinical practice patterns all become useful features for optimization.

This specialization produces impressive internal validation metrics but poor generalizability. Models essentially overfit not to individual training samples but to institutional characteristics. The more completely models exploit institution-specific patterns, the better they perform internally but the worse they generalize externally. Researchers studying cross-hospital validation concluded that performance degradation identified limitations in developing a generic model for different hospitals, recommending instead that specialized prediction models be generated for each hospital to guarantee performance.

However, institution-specific models create their own problems. Developing separate models for each deployment site requires substantial resources, technical expertise, and local data that many healthcare facilities lack. Community hospitals, rural facilities, and resource-limited settings cannot afford to develop custom AI systems, yet deploying externally developed models risks poor performance. This dynamic threatens to create healthcare AI systems accessible only to well-resourced institutions, exacerbating rather than reducing disparities.

The specialization-generalization trade-off also explains why technical sophistication does not ensure generalizability. More complex models with greater capacity can learn more intricate patterns in training data, potentially achieving higher internal validation performance. However, this same capacity enables more complete exploitation of institution-specific features, worsening generalizability. Without explicit architectural choices, training procedures, or data strategies promoting generalization, increasing model sophistication may paradoxically reduce external validity.

4.8. Structural Analysis: Generalizability as Predictable Consequence

Examining poor generalizability through a structural lens reveals it is not a technical accident requiring incremental fixes but a predictable consequence of how AI diagnostic systems are currently developed. Just as with algorithmic bias, poor generalizability stems from fundamental development practices that prioritize narrow metrics over robust performance.

Single-site optimization dominates current practice. Models are developed and validated primarily or exclusively using data from individual institutions or small consortia. Optimization targets internal validation performance without explicit generalizability constraints. This approach produces systems maximally adapted to specific contexts while minimizing external validity. When developers lack access to diverse multi-site data during training, creating generalizable models becomes essentially impossible regardless of algorithmic sophistication.

Convenience sampling determines which data are used for AI development. Readily available datasets from well-resourced institutions become training sources not because they are representative but because they are accessible. This sampling strategy systematically excludes diverse clinical contexts, patient populations, and healthcare settings—precisely the diversity essential for generalization. Geographic concentration, demographic homogeneity, and institutional similarity in training data inevitably produce models that generalize poorly beyond those specific contexts (Larson et al., 2018).

Narrow evaluation metrics obscure generalization failures. When studies report overall accuracy, sensitivity, or AUC without disaggregating by subgroups or validating across external sites, poor generalizability remains invisible. Publication norms emphasizing impressive performance numbers rather than external validation encourage researchers to optimize for internal metrics while avoiding rigorous generalizability assessment. Regulatory approval processes that do not mandate comprehensive external validation allow systems to enter clinical practice despite limited evidence of robust performance (Wu et al., 2021).

Economic incentives reinforce these practices. Collecting diverse multi-site datasets is expensive and time-consuming. Comprehensive external validation requires resources, partnerships, and data-sharing agreements that slow development timelines and increase costs. In competitive commercial environments where companies race to demonstrate high performance and secure regulatory approval, investments in generalizability compete with pressures for rapid deployment. When validation gaps are tolerated by regulators and markets, economic rationality suggests minimizing validation costs (Kelly et al., 2019).

The structural analysis demonstrates that poor generalizability, like algorithmic bias, results not from individual failures but from systemic development practices. Current approaches optimize for success within specific contexts while systematically neglecting the diversity essential for robust performance across varied deployment settings. Technical solutions transfer learning, domain adaptation, federated learning—offer value but cannot fully compensate for fundamentally inadequate development paradigms. Sustainable generalizability requires restructuring how development occurs: prioritizing diversity in data collection, mandating multi-site validation, optimizing for worst-case rather than average performance, and treating generalizability as a fundamental requirement rather than an aspirational goal.

The following section examines how poor generalizability and algorithmic bias compound each other, demonstrating these are not separate challenges but interconnected manifestations of the same structural problems in AI diagnostic development.

5. The Intersection: How Bias and Generalizability Compound Each Other

The previous sections examined algorithmic bias and poor generalizability as distinct phenomena, each with its own manifestations, mechanisms, and consequences. This analytical separation, however, obscures a critical reality: they are not separate challenges but interconnected manifestations of the same structural deficiencies in AI development practices. This section demonstrates how bias and poor generalizability compound each other, creating equity gaps more severe than either challenge would produce in isolation. Understanding this intersection is essential, for it reveals why technical solutions addressing either problem individually prove insufficient and why a fundamental restructuring of development paradigms is necessary.

5.1. Common Root: Unrepresentative Training Data

Both algorithmic bias and poor generalizability fundamentally stem from the same root cause: training data that fails to represent the full diversity of populations and contexts where AI diagnostic systems will be deployed. This representation bias is a dominant form of bias that critically limits the generalizability of healthcare AI models (Zech et al., 2018). When datasets systematically underrepresent specific demographic groups, two outcomes occur simultaneously: models learn less effectively about disease patterns in those groups (producing bias), and they optimize for characteristics present in overrepresented groups, failing to capture patterns necessary for performance in diverse deployment contexts (poor generalizability).

This common root is evident in the geographic concentration of training data. Research reveals that over half of all published clinical AI models leverage datasets from either the United States or China, with many U.S. datasets overrepresenting non-Hispanic Caucasian patients relative to the general population (Larson et al., 2018). This concentration produces both bias (worse performance for underrepresented ethnic groups) and poor generalizability (performance degradation when deployed outside these specific geographic contexts). The mechanism operates through how AI models handle underrepresented groups. When trained on imbalanced data, algorithms tend to "underestimate" or treat minority patterns as noise to avoid overfitting, approximating mean trends instead (Chen et al., 2021). This behaviour simultaneously produces bias (differential performance) and compromises generalizability (inability to handle contexts where these "minority" patterns are prevalent).

5.2. Compounding Mechanisms: How Each Problem Worsens the Other

Bias and poor generalizability do not merely share common origins; they actively compound each other through feedback mechanisms that amplify both problems beyond what either would produce independently.

Bias worsens generalizability through learned dependencies on majority characteristics. When models perform poorly for specific demographic groups during training, developers often lack the disaggregated performance metrics necessary to detect this bias (Mehrabi et al., 2021). Models that appear to perform well overall may achieve high accuracy by specializing to majority group characteristics while failing for minorities. This specialization creates systems optimized for narrow demographic contexts, ensuring poor generalizability when deployed in settings with different demographic distributions.

Poor generalizability amplifies bias through deployment decisions and feedback loops. When AI systems fail to generalize, healthcare organizations face a choice. Well-resourced institutions can afford local validation and model customization, while under-resourced facilities—which often serve populations already facing healthcare disparities—cannot (Reddy et al., 2020). This creates a pattern where AI tools are deployed successfully in privileged contexts but remain unavailable or unreliable in marginalized settings, systematically amplifying existing healthcare inequities.

Training data scarcity creates a vicious cycle of compounding disadvantages. Limited representation of specific populations creates multiple interconnected problems: models learn disease patterns less reliably for these groups (bias); they lack examples to recognize these groups in varied contexts (poor generalizability); and external validation studies often lack sufficient samples to assess performance reliably (validation gaps). Each limitation reinforces the others, creating particularly severe disadvantages for populations with minimal representation (Seyyed-Kalantari et al., 2021).

5.3. Intersectional Inequities: Compounded Disadvantages

The intersection of bias and poor generalizability creates particularly acute problems for populations facing multiple, overlapping forms of marginalization. Individuals belonging to several underrepresented categories simultaneously for instance, elderly Black women in rural areas experience compounded disadvantages that exceed the sum of individual biases.

Intersectional data scarcity operates multiplicatively rather than additively. If Black patients constitute 10% of a training dataset and rural patients constitute 15%, Black rural patients may represent only 0.5-1.5%, not 25% (Buolamwini & Gebru, 2018). This severe underrepresentation means models have virtually no examples from which to learn disease patterns for intersectional groups. The result is AI systems that perform catastrophically for precisely those populations facing the greatest healthcare access barriers and health disparities.

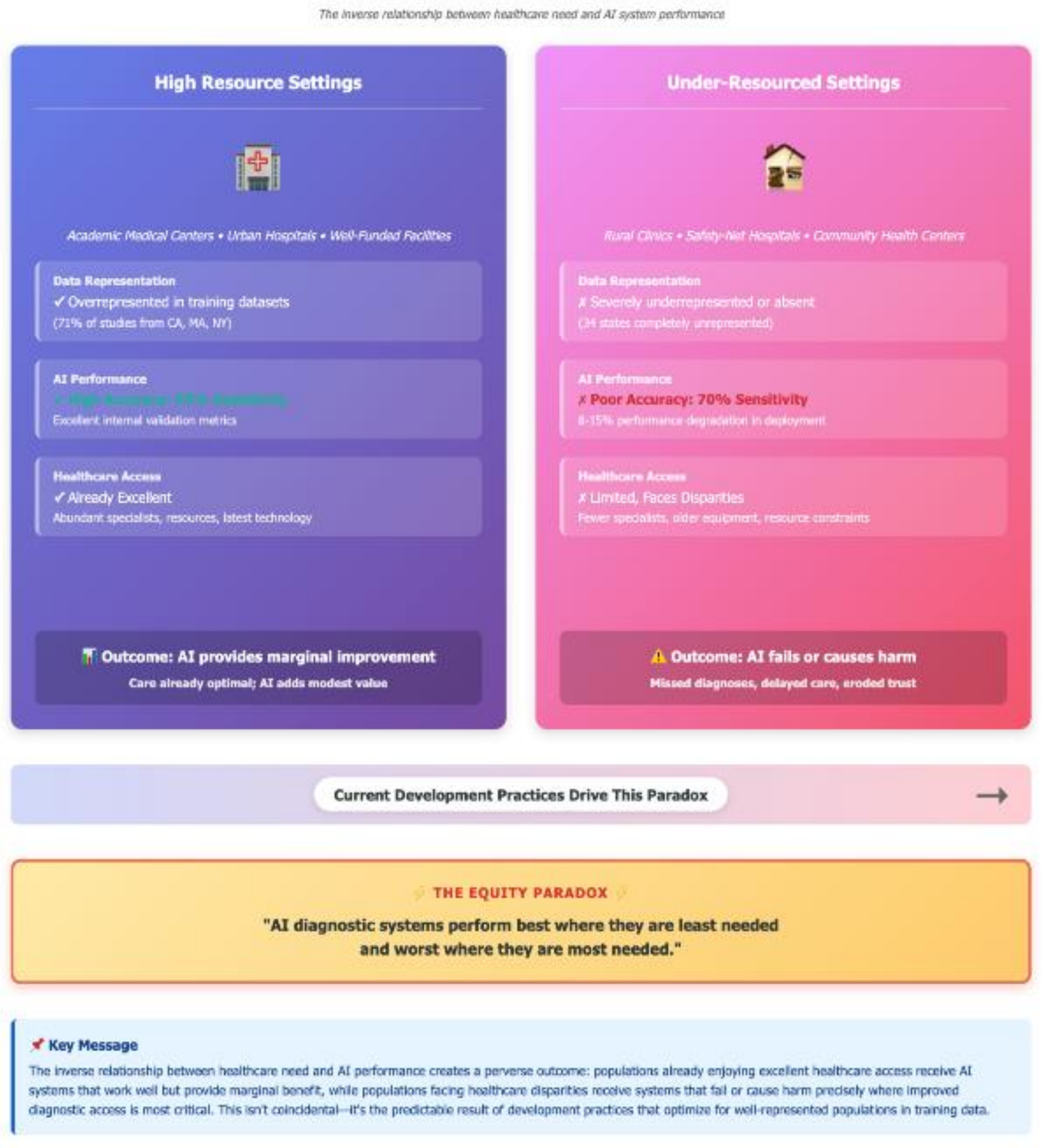
Historical underrepresentation affects both datasets and development teams, with women and researchers of colour being underrepresented in clinical AI research (AIM-AHEAD, 2024). This dual absence means both the data and the perspectives essential for identifying and addressing intersectional equity concerns are missing during design, development, and validation.

5.4. The Equity Paradox: AI Helps Least Where Needed Most

The compounding of bias and poor generalizability culminates in what this thesis terms the "equity paradox" in medical AI: *diagnostic systems perform best for populations with the least need for improved care access and worst for populations who could benefit most from enhanced diagnostic capabilities.* This inversion transforms AI's promise to democratize healthcare into a reality where it amplifies existing disparities.

Populations already enjoying excellent healthcare access typically well-served, majority demographics in well-resourced urban academic medical centres are most likely to be well-represented in training data. Models optimized on these populations perform best for them and generalize most reliably to similar settings. These groups experience AI as delivering on its promise.

Conversely, populations facing healthcare access barriers rural communities, racial and ethnic minorities, low-income individuals are systematically underrepresented in training data. Models perform worse for them individually (bias) and fail when deployed in facilities serving them (poor generalizability). When under-resourced facilities attempt deployment, they encounter systems optimized for different contexts. These groups experience AI as unreliable or harmful: leading to missed diagnoses, false alarms, and erosion of trust (Char et al., 2018). This paradox is intensified because well-resourced institutions can invest in validation and customization, while under-resourced ones cannot, creating a two-tier ecosystem.



The inverse relationship between healthcare need and AI system performance

High Resource Settings

Academic Medical Centers • Urban Hospitals • Well-Funded Facilities

Data Representation
✓ Overrepresented in training datasets
(71% of studies from CA, MA, NY)

AI Performance
✓ High Accuracy: [illegible] Sensitivity
Excellent internal validation metrics

Healthcare Access
✓ Already Excellent
Abundant specialists, resources, latest technology

Outcome: AI provides marginal improvement
Care already optimal; AI adds modest value

Under-Resourced Settings

Rural Clinics • Safety-Net Hospitals • Community Health Centers

Data Representation
✗ Severely underrepresented or absent
(34 states completely unrepresented)

AI Performance
✗ Poor Accuracy: 70% Sensitivity
8-15% performance degradation in deployment

Healthcare Access
✗ Limited, Faces Disparities
Fewer specialists, older equipment, resource constraints

⚠ **Outcome: AI fails or causes harm**
Missed diagnoses, delayed care, eroded trust

Current Development Practices Drive This Paradox →

THE EQUITY PARADOX

"AI diagnostic systems perform best where they are least needed and worst where they are most needed."

Key Message

The inverse relationship between healthcare need and AI performance creates a perverse outcome: populations already enjoying excellent healthcare access receive AI systems that work well but provide marginal benefit, while populations facing healthcare disparities receive systems that fail or cause harm precisely where improved diagnostic access is most critical. This isn't coincidental—it's the predictable result of development practices that optimize for well-represented populations in training data.

Figure 4 The Equity Paradox in Medical AII

5.5. Real-World Manifestations: Case Studies of Compounding Effects

Examining specific deployment contexts reveals how bias and poor generalizability compound in practice.

- **Rural Healthcare Deployment:** Rural facilities differ from urban academic centres in infrastructure, workflows, and staffing creating generalizability challenges. Simultaneously, rural populations often have different demographic and socioeconomic characteristics—creating bias vulnerabilities. When urban-developed AI systems are deployed rurally, they encounter both **distribution shift** (different context) and **demographic mismatch**, leading to dramatically worse performance.

- **International Deployment to LMICs:** AI systems developed in high-income countries encounter massive distribution shifts in low- and middle-income countries (LMICs): different equipment, protocols, disease epidemiology, and population demographics (Wynants et al., 2020). Each dimension creates both generalizability challenges and bias risks. The combination creates systems that are often non-functional for populations potentially benefiting most from improved diagnostic access.
- **Safety-Net Hospitals:** These institutions serve disproportionately minority, low-income populations (demographic mismatch/bias risk) and operate under different resource constraints than academic centres (contextual mismatch/generalizability challenge). Deploying externally developed AI systems here produces particularly poor performance precisely where healthcare challenges are most acute.

5.6. Validation Gaps Obscure Compounding Effects

The interconnection between bias and poor generalizability is further obscured by validation practices that fail to assess either challenge adequately. Aggregate validation metrics—reporting overall accuracy across entire test sets—obscure both bias (by averaging over groups) and poor generalizability (by testing on data from the same distribution as training data). A 2025 analysis of FDA-approved devices found clinical performance studies reported at approval for only ~50% of devices, with less than one-third providing sex-specific performance data (Wu et al., 2021). This inadequacy means most deployed systems have not been tested across diverse demographic groups *or* varied institutional contexts.

The lack of intersectional validation is particularly problematic. Even studies that disaggregate by single demographic characteristics rarely examine intersectional categories (e.g., elderly Black women). Yet these groups face the most severe compounding effects. Without intersectional validation, the populations most vulnerable to AI failures remain invisible in performance assessments (Buolamwini & Gebru, 2018).

5.7. Why Technical Solutions Alone Are Insufficient

Recognizing the interconnected nature of bias and poor generalizability reveals why purely technical mitigation strategies, applied in isolation, are destined to fall short. Approaches that treat these as separate, isolated challenges cannot address their compounding effects or underlying common causes.

Bias mitigation techniques (e.g., post-hoc fairness constraints) may address demographic performance gaps within a specific training distribution but do not ensure models will maintain fairness across different deployment contexts (Barocas et al., 2019). A model adjusted to be "fair" in one institution may exhibit dramatically different fairness properties when deployed elsewhere.

Generalizability techniques (e.g., domain adaptation) that optimize for robust performance across different hospitals may inadvertently worsen bias if they prioritize performance on well-represented majority groups. Transfer learning using limited local data may improve average performance while leaving or worsening demographic disparities if the local data lacks diversity.

The fundamental limitation is that technical solutions operate *within* the paradigm that created both problems: development using unrepresentative data and optimization for narrow metrics. They are attempts to fix the outputs of a broken system rather than repair the system itself.

5.8. Structural Analysis: The Need for Integrated Solutions

Understanding bias and poor generalizability as interconnected outcomes of common structural deficiencies points toward integrated solutions that address root causes rather than symptoms. Three key imperatives emerge:

Data diversity is non-negotiable**.** No algorithmic sophistication can compensate for fundamentally unrepresentative training data. Proactive collection of diverse datasets capturing demographic, geographic, institutional, and clinical variability must become a fundamental requirement, necessitating restructured incentives and regulatory expectations (STANDING Together, 2024).

Validation must be comprehensive and intersectional. Assessing bias within single-site data and assessing generalizability using aggregate metrics both fail to reveal compounding effects. Validation frameworks must simultaneously examine performance across demographic subgroups *and* across deployment contexts, with a focus on intersectional categories. Regulatory approval should require evidence of maintained equitable performance across diverse real-world settings (FDA, 2024; Wu et al., 2021).

Development priorities must be fundamentally reoriented. Current practices optimize for performance on available data, treating diversity and generalizability as secondary concerns. Sustainable equity requires inverting these priorities: treating robust, equitable performance across diverse populations and contexts as the primary goal, even if this means accepting lower peak performance on narrow, aggregated metrics (Wiens et al., 2019).

The structural interconnection means addressing either challenge requires addressing both. Solutions must be integrated, proactive, and structural. The following sections examine the barriers to such solutions and propose a concrete framework for this necessary reorientation.

6. Barriers to clinical translation

The previous sections established that algorithmic bias and poor generalizability stem from structural deficiencies in AI development practices. However, understanding why these problems persist requires examining the barriers preventing solutions from being implemented. Current regulatory frameworks, economic incentives, organizational structures, and clinical workflows actively reinforce development practices that prioritize narrow performance over equity and generalizability. This section analyzes how these systemic barriers impede the translation of AI diagnostic systems into equitable, robust clinical tools and perpetuate the equity paradox.

6.1. Regulatory Framework Inadequacies

Regulatory oversight of AI diagnostic systems aims to ensure safety and effectiveness before clinical deployment. However, existing frameworks were designed for traditional medical devices with fixed characteristics, not adaptive AI systems that learn from data and may evolve over time. This fundamental mismatch creates regulatory gaps that allow biased, poorly generalizable systems to enter clinical practice.

6.1.1. FDA Approval Processes and Their Limitations

The U.S. Food and Drug Administration regulate AI-enabled medical devices through established pathways: 510(k) premarket clearance, De Novo classification, or Premarket Approval (FDA, 2024). As of mid-2024, approximately 950 AI/ML-enabled medical devices had received FDA clearance, with roughly 100 new approvals annually. The majority fall into Class II (moderate risk) requiring 510(k) clearance based on substantial equivalence to predicate devices.

These pathways contain critical limitations for addressing bias and generalizability. The 510(k) process’s precedent-based approach can perpetuate problems present in earlier generations. More fundamentally, current approval processes do not consistently require comprehensive evidence of generalizability or equitable performance across diverse populations. A 2025 analysis of 903 FDA-approved AI-enabled medical devices found clinical performance studies were reported at approval for only approximately half, with less than one-third providing sex-specific performance data and only one-fourth addressing age-related subgroups (Wu et al., 2021). This means most deployed systems lack disaggregated validation demonstrating equitable performance, let alone rigorous external validation across different institutions and contexts.

The FDA explicitly recognizes these limitations, acknowledging its traditional paradigm was not designed for adaptive AI technologies (FDA, 2025). While updated guidance establishes documentation requirements including bias analysis, the absence of mandatory requirements for comprehensive multi-site external validation represents a persistent gap that allows systems with limited evidence of equitable performance to enter clinical practice.

6.1.2. The Predetermined Change Control Plan Challenge

The FDA's Predetermined Change Control Plan (PCCP) allows manufacturers to implement approved algorithm modifications without new marketing applications for each change, provided modifications stay within predefined boundaries (FDA, 2025). While this approach recognizes AI's adaptive nature, it creates challenges for monitoring equity.

If post-deployment data lack diversity or reflect biased clinical practices, continuous learning could actually worsen bias over time. Without robust real-world performance monitoring disaggregated by demographic groups, evolutionary changes may systematically degrade equitable performance while remaining within authorized modification boundaries. The challenge intensifies because the main method of detecting degradation today remains clinical intuition, which is unreliable and variable, meaning harm may occur before problems are noticed.

6.1.3. Validation Requirements and Their Gaps

Current regulatory approaches focus primarily on technical accuracy rather than clinical utility and equitable performance. Approval can be granted based on performance in controlled research settings without evidence the system maintains accuracy in diverse real-world deployment contexts. The absence of mandatory external validation requirements represents a critical gap. While the FDA encourages multi-site testing and demographic disaggregation, these remain recommendations rather than requirements for many device classes (Wu et al., 2021). Healthcare institutions deploying FDA-cleared devices often assume regulatory approval indicates comprehensive validation, unaware that clearance may be based on single-site studies with limited demographic diversity.

International regulatory frameworks face similar challenges. The European Union's AI Act explicitly designates medical AI as high-risk, requiring quality management, transparency, human oversight, and bias monitoring (European Commission, 2024). However, ensuring datasets adequately represent intended populations remains difficult in practice, and notified bodies conducting conformity assessments have limited experience with AI-specific validation challenges.

6.2. Economic Barriers and Misaligned Incentives

Economic factors powerfully shape AI development priorities, often incentivizing practices that perpetuate bias and poor generalizability. Understanding these economic barriers reveals why sustainable equity requires restructuring financial incentives.

6.2.1. Development Costs and Resource Constraints

Creating AI diagnostic systems requires substantial financial investment. Developing a single AI model can cost upwards of $1 million, and because models do not always work correctly, not every model makes it to deployment. More comprehensive estimates suggest implementing AI in healthcare ranges from $40,000 for simple functionality to $100,000 or more for complex solutions.

These costs create strong incentives to minimize expenses wherever possible. Collecting diverse, representative datasets from multiple institutions across geographic regions is expensive and time-consuming. Using readily available data from single well-resourced institutions dramatically reduces costs and accelerates development timelines (Kelly et al., 2019). Similarly, comprehensive external validation requires substantial resources. When external validation is not mandated by regulators, economic rationality suggests minimizing validation costs by testing only on readily available datasets.

6.2.2. Competitive Pressures and Time-to-Market

Medical AI represents a competitive commercial market where first-mover advantages confer significant benefits. Companies race to demonstrate impressive performance metrics, secure regulatory approval, and establish market presence before competitors. These pressures create strong incentives to prioritize speed over comprehensive validation (Shaw et al., 2019).

Collecting diverse data, conducting multi-site validation, and implementing bias mitigation strategies all extend development timelines. In fast-moving competitive markets, delays measured in months can mean the difference between market leadership and obsolescence. When regulatory frameworks do not mandate comprehensive diversity and validation, competitive dynamics systematically favour companies that minimize these time-consuming activities.

6.2.3. Deployment Costs and Implementation Barriers

Beyond development, substantial costs arise during clinical deployment. Continuously running AI models is costly, creating a financial barrier to widespread use. Healthcare institutions face expenses for software licensing, hardware infrastructure, system integration, staff training, workflow redesign, and ongoing maintenance.

These deployment costs create particularly acute barriers for under-resourced facilities serving marginalized populations. Rural hospitals and community centres lack resources to develop these tools, to evaluate them effectively, or to implement them into their computer systems, making accessibility and equity critical concerns (Reddy et al., 2020). This resource disparity creates a vicious cycle: AI systems are developed primarily at well-resourced institutions using their data, optimized for their contexts, and validated in their settings. Under-resourced facilities serving diverse, marginalized populations lack both the representation in training data and the resources for effective deployment, perpetuating the equity paradox.

6.2.4. Misaligned Value Propositions

Current AI business models often misalign with equity goals. Value propositions typically emphasize aggregate efficiency gains reducing radiologist reading time, accelerating diagnoses, minimizing false positives. These benefits accrue most reliably in contexts similar to development settings serving similar populations.

When AI systems perform poorly for specific demographic groups or in under-resourced settings, the economic case for deployment weakens in precisely those contexts. Facilities serving predominantly marginalized populations may find AI tools less effective, generating fewer benefits while requiring equal or greater implementation costs. This creates perverse incentives where commercial viability depends on deployment in already well-served contexts rather than where improved diagnostic access is most needed.

Healthcare reimbursement structures compound these misalignments. When payment models reward volume and efficiency rather than equity and outcomes, AI tools that improve throughput in privileged settings generate clearer financial returns than tools addressing disparities. Without reimbursement mechanisms explicitly valuing equity, economic incentives systematically favour development and deployment patterns that exacerbate rather than reduce healthcare disparities.

6.3. Organizational and Workflow Integration Challenges

Even when AI diagnostic systems demonstrate adequate performance and receive regulatory approval, organizational and clinical workflow barriers impede effective deployment—particularly in ways that ensure equitable, generalizable use.

6.3.1. Clinical Workflow Disruption and Resistance

Integrating AI into clinical workflows requires fundamental changes to established practices. Medical associations have identified data accessibility and operational infrastructure as significant barriers to AI integration, affecting 74% of respondents (Shaw et al., 2019). Clinicians face additional workload during implementation phases, and in already overburdened healthcare settings facing clinician burnout, this creates resistance even when AI promises long-term efficiency gains.

Trust represents a critical barrier. Clinicians must trust AI recommendations to incorporate them into clinical decision-making. When systems produce unexplained predictions, generate occasional obvious errors, or demonstrate inconsistent performance, clinician trust erodes (Char et al., 2018). Experiences with biased predictions or poor generalizability spread through professional networks, creating resistance that affects adoption.

Beyond general trust, two specific human-factor barriers critically impede safe deployment: automation bias and alert fatigue. Automation bias describes clinicians' tendency to over-rely on AI recommendations, accepting algorithmic outputs without sufficient scrutiny. This creates risks when models exhibit hidden biases. Simultaneously, alert fatigue emerges when AI systems generate excessive false positives, causing clinicians to become desensitized and potentially ignore critical warnings. These psychological dynamics mean that even technically accurate AI can produce negative clinical outcomes through its interaction with human decision-makers.

6.3.2. Data Infrastructure and Interoperability

Effective AI deployment requires robust data infrastructure that many healthcare facilities lack. Systems must integrate with electronic health records, imaging archives, laboratory information systems, and clinical workflows—integration that proves technically complex and expensive.

Data quality, standardization, and accessibility create persistent challenges. Medical associations reported that accessing health data for training AI algorithms and the complexities in training, testing, and validating AI algorithms were the most prominent barriers to AI adoption (Shaw et al., 2019). When data exist in incompatible formats or remain siloed across systems, deploying AI tools requiring integrated multi-modal data becomes prohibitively difficult.

6.3.3. Organizational Governance and Decision-Making

Healthcare organizations face challenges establishing appropriate governance structures for AI adoption. Decisions about which AI tools to deploy, how to validate their performance, when to update systems, and how to monitor for bias require expertise spanning clinical medicine, data science, ethics, and informatics—combinations rarely available in integrated governance teams.

A 2024 survey of 43 U.S. health systems found AI adoption and perceptions of success varied significantly, with only 19% reporting high degrees of success with AI in imaging and radiology despite most having deployed such systems. This disconnect suggests governance and implementation challenges beyond technical performance.

6.3.4. Training and Change Management

Successful AI deployment requires comprehensive training for clinical and technical staff training that proves expensive and time-consuming. The time required for training AI systems and certainly in early stages following implementation is likely to remain a barrier (Shaw et al., 2019). When clinical schedules already operate at capacity, finding time for comprehensive AI training without compromising patient care proves extremely difficult.

Change management more broadly communicating benefits, addressing concerns, managing resistance, and building organizational commitment requires dedicated resources and expertise. Without effective change management, even technically sound AI implementations can fail due to insufficient user engagement or cultural resistance.

6.4. Performance Monitoring and Accountability Gaps

Even after successful deployment, ensuring AI systems maintain equitable, effective performance requires ongoing monitoring monitoring that current practices inadequately support.

6.4.1. Lack of Real-World Performance Tracking

Most deployed AI diagnostic systems lack robust infrastructure for continuous performance monitoring in real-world settings. AI system performance can be influenced by changes in clinical practice, patient demographics, data inputs, and healthcare infrastructure, with data drift, concept drift, and model drift potentially leading to performance degradation, bias, or reduced reliability (FDA, 2025). Yet systematic tracking of these dynamics remains rare.

Measuring accuracy continuously requires ongoing access to ground truth labels confirmed diagnoses for patients receiving AI-based recommendations. Obtaining these labels is resource intensive. For some applications, true outcomes may not be known for months or years, making timely detection of performance degradation impossible.

6.4.2. Absence of Disaggregated Monitoring

Even when performance monitoring occurs, it rarely includes systematic disaggregation by demographic subgroups or comparison across deployment contexts. Aggregate performance metrics can mask systematic bias or context-specific failures. A model maintaining 85% overall accuracy might show 90% accuracy for well-represented groups but only 70% for minorities a critical disparity invisible in aggregate statistics.

Recent research examining healthcare AI systems deployed across multiple institutions found only 23% had undergone rigorous bias testing before deployment, while fewer than 15% had established clear accountability structures for addressing errors. When systems produce disparate outcomes, there exist no systematic mechanisms for patients to seek recourse or for institutions to implement corrections quickly.

6.4.3. Technical Challenges in Drift Detection

Detecting when AI performance degrades requires distinguishing meaningful changes from normal variation. Distribution shifts creating performance problems may be subtle and gradual. Detecting these changes while avoiding false alarms requires sophisticated monitoring systems that few healthcare institutions have implemented.

Alternative monitoring approaches using proxy metrics can alert to potential problems without requiring ground truth labels. However, these methods cannot definitively confirm whether performance has degraded, only that conditions have changed in ways that might affect performance. Healthcare institutions must then decide whether to adjust, retrain, or suspend AI systems based on uncertain signals—decisions with significant resource implications and potential patient safety impacts.

6.5. Knowledge Gaps and Capacity Limitations

Implementing equitable, generalizable AI diagnostic systems requires expertise spanning multiple domains—clinical medicine, machine learning, health equity, implementation science, and informatics. This expertise remains scarce, creating human capital barriers to sustainable AI deployment.

Many healthcare institutions, particularly smaller community hospitals and under-resourced facilities, lack in-house expertise to independently evaluate AI systems, conduct local validation studies, or customize models for their specific contexts. They depend on vendor claims and regulatory approvals, unable to critically assess whether systems will work well for their patient populations and clinical workflows (Reddy et al., 2020).

Even well-resourced institutions face challenges. Trust was found to be a significant catalyst of adoption, impacted by several barriers, with governance structures identified as key facilitators (Shaw et al., 2019). Yet establishing effective governance requires understanding complex technical, ethical, and clinical considerations that exceed expertise available in many healthcare organizations.

6.6. Structural Analysis: How Barriers Reinforce Problematic Practices

Examining these barriers collectively reveals how current structures systematically reinforce development practices that perpetuate bias and poor generalizability rather than addressing them. Regulatory frameworks that do not mandate comprehensive external validation allow systems with limited evidence of equity and generalizability to enter clinical practice (Wu et al., 2021). Economic incentives that reward speed, narrow accuracy metrics, and deployment in well-served markets systematically discourage investments in diversity and comprehensive validation (Kelly et al., 2019). Organizational structures that separate developers from deployment contexts create information asymmetries and misaligned incentives. Resource disparities between well-funded and under-resourced institutions concentrate both AI development and successful deployment in privileged settings (Reddy et al., 2020), actively reinforcing the equity paradox.

These barriers are not isolated problems but interconnected elements of a system that makes problematic development practices economically rational and practically feasible. Developers face strong incentives to use convenient data sources, minimize validation costs, and prioritize impressive performance metrics over robust generalizability. Regulators lack frameworks to effectively require comprehensive diversity and validation. Healthcare institutions lack resources to conduct independent assessment. Patients lack information and recourse when systems perform poorly.

Addressing these structural barriers requires coordinated changes across regulatory policy, economic incentives, organizational practices, and knowledge infrastructure. Technical solutions addressing algorithmic bias or poor generalizability cannot succeed within a system that actively disincentivizes the very practices necessary for equity and generalizability. The following section examines concrete strategies for restructuring these systems to centre equity and generalizability as fundamental requirements throughout the AI lifecycle.

Table 3 Structural Root Causes and Their Manifestations

Structural Root Cause	Primary Manifestation	Secondary Consequences	Compounding Effect
Unrepresentative Training Data (Geographic concentration, demographic homogeneity)	Algorithmic Bias: Differential performance across groups	Poor Generalizability: Failure in new contexts	Populations facing healthcare disparities experience both worse individual care AND systemic tool failure
Narrow Optimization Priorities (Focus on aggregate accuracy metrics)	Development of models that sacrifice minority performance for overall metrics	Validation frameworks that obscure subgroup disparities	Tools appear successful in development but fail catastrophically for specific populations in deployment
Inadequate Validation Frameworks (Lack of mandatory multi-site, disaggregated testing)	Regulatory approval based on limited evidence	Deployment of systems with unknown performance boundaries	Healthcare institutions conduct uncontrolled experiments on patients

Economic Incentives for Speed (Time-to-market over robustness)	Use of convenient, non-representative datasets	Avoidance of comprehensive validation due to cost/time	Well-resourced settings get functional tools first; under-resourced settings get hand-me-down failures
Resource Disparities in Development (AI expertise concentrated in privileged institutions)	Development by teams lacking diverse perspectives	Tools optimized for contexts familiar to developers	

7. Solutions and best practices: reorienting the development pipeline

The preceding analysis establishes that algorithmic bias and poor generalizability are not technical anomalies but predictable outputs of a flawed development paradigm. Addressing these interconnected challenges requires moving beyond isolated technical fixes to fundamentally restructure how AI diagnostic systems are conceived, built, validated, and deployed. This section proposes an Equity-centred Development Lifecycle—a concrete framework that embeds diversity, equity, and generalizability as non-negotiable core requirements at every stage, from initial planning to post-market surveillance.

7.1. Proactive Data Diversity: From Convenience to Comprehensiveness

The foundation of equitable AI is representative data. Current reliance on convenience sampling systematically excludes populations and contexts essential for robust performance. Reorienting toward proactive data diversity requires mandatory, structured strategies.

7.1.1. Mandating Diverse Data Collection

Diversity must transition from an aspirational goal to a mandatory regulatory and funding prerequisite. This requires setting explicit, minimum thresholds for demographic (race, ethnicity, gender, age, socioeconomic status), geographic, and institutional representation in training datasets. Initiatives like **STANDING Together** are developing consensus-driven standards for equitable health data, emphasizing that representation alone is insufficient—data must also be accurate and ethically sourced from minoritized groups (STANDING Together, 2024). Regulatory bodies could require developers to demonstrate compliance with such standards prior to approval. Economic incentives, such as targeted grants, must reward comprehensive data collection over sheer volume.

7.1.2. Community-Engaged Data Curation

True representation requires engaging the communities represented in the data. Community-engaged approaches involve affected populations in defining what data is collected, how it is used, and how benefits are shared. Research Centres in Minority Institutions (RCMI) are uniquely positioned to lead this work, ensuring AI tools are designed with input from the communities they aim to serve (AIM-AHEAD, 2024). This participatory model helps capture not just demographic checkboxes, but the lived experiences, disease presentations, and healthcare contexts of diverse communities, building trust and legitimacy.

7.1.3. Federated Learning for Privacy-Preserving Diversity

Privacy concerns and data governance are major barriers to sharing patient information. Federated learning (FL) offers a paradigm-shifting solution by enabling model training across multiple institutions without centralizing raw patient data (Rieke et al., 2020). In FL, data remains within each hospital's secure infrastructure; only encrypted model updates are shared. This approach facilitates collaboration across diverse and under-resourced settings while maintaining compliance with regulations like HIPAA and GDPR. When combined with privacy-enhancing techniques like differential privacy, FL can build a technical foundation for the inclusive data ecosystems equitable AI requires.

7.2. Equity-centred Development Practices

With diverse data as a foundation, the development process itself must be re-engineered to optimize for fairness and robustness as primary objectives.

7.2.1. Fairness-Aware Optimization

Standard optimization minimizes average error, which can mask poor performance for minority groups. Equity-cantered development employs fairness-aware optimization techniques that treat worst-case subgroup performance as the key metric. This includes distributionally robust optimization, adversarial debiasing to remove demographic information from latent representations, and fairness constraints integrated directly into loss functions (Barocas et al., 2019). The goal is to build models for which high performance is a guarantee for all, not an average skewed by privilege.

7.2.2. Involving Diverse Development Teams

The underrepresentation of women and people of colour in AI research means the perspectives needed to identify equity concerns are often absent from the design process (Buolamwini & Gebru, 2018). Building diverse teams is not merely a matter of justice but of technical necessity. This requires intentional recruitment, support for researchers at minority-serving institutions, inclusive mentorship pipelines, and governance structures that ensure diverse voices have decision-making authority. Initiatives like AIM-AHEAD explicitly link workforce diversity to the development of equitable AI (AIM-AHEAD, 2024).

7.2.3. Human-centred Design and Continuous Stakeholder Engagement

AI systems are socio-technical interventions whose success depends on seamless integration into human workflows and decision-making. Human-cantered design logic must involve clinicians, patients, and healthcare administrators throughout the development lifecycle—from defining requirements to evaluating prototypes (Sendak et al., 2020). This engagement is critical for mitigating automation bias (over-reliance on AI) and alert fatigue. Interfaces should be designed to support, not replace, clinical reasoning by displaying confidence scores, forcing consideration of AI outputs, and highlighting potential demographic mismatches where model performance may be weaker.

7.3. Comprehensive External Validation Requirements

Robust validation is the bridge between laboratory performance and real-world utility. Current practices are inadequate. Comprehensive, mandatory external validation must become the gatekeeper for clinical deployment.

7.3.1. Multi-Site, Multi-Population Validation Protocols

Regulatory approval should require demonstrable performance across a spectrum of independent sites that reflect real-world diversity: different institution types (academic, community, rural, safety-net), geographic regions, equipment, and patient demographics. Validation must move beyond single-site "internal" tests to prove **external generalizability**. Protocols should be standardized to allow for meaningful comparison across studies and systems.

7.3.2. Transparency and Disaggregated Reporting

Transparency is non-negotiable. The publication of "model facts" or "algorithmic impact assessments" should be required, detailing training data demographics, known limitations, and—critically—disaggregated performance metrics across race, ethnicity, sex, age, and socioeconomic status (Sendak et al., 2020). Aggregate metrics like overall AUC must be supplemented with subgroup-specific sensitivity, specificity, and PPV. This transparency enables regulators, healthcare systems, and the public to assess fairness and fitness-for-purpose.

7.3.3. Collaborative Validation Networks

Establishing independent, collaborative validation networks—consortia of diverse healthcare institutions—would provide trusted infrastructure for pre-deployment testing. These networks, funded through public-private partnerships, would use standardized protocols to evaluate AI systems on local data, generating essential evidence of generalizability while protecting patient privacy through federated approaches.

7.4. Regulatory Reform for Equity and Generalizability

Regulatory bodies hold the most direct leverage to enforce higher standards. Reform must make equity and generalizability central pillars of the approval process.

7.4.1. Mandating Diversity and External Validation

The FDA and other regulators must evolve guidance into requirement. This includes mandating minimum diversity thresholds in training data, compulsory multi-site external validation with disaggregated results, and post-market surveillance plans that track real-world performance across subgroups (FDA, 2025; Wu et al., 2021). Approval should be conditional on meeting predefined equity benchmarks, not just technical accuracy.

7.4.2. Adaptive Regulation for Continuously Learning Systems

For AI systems that learn after deployment, the Predetermined Change Control Plan (PCCP) framework must explicitly guard against equity degradation. It must require ongoing, disaggregated performance monitoring as part of the "predetermined" plan, with clear triggers for mandatory reassessment if performance gaps emerge. Regulations must ensure continuous learning improves systems equitably for all patients (FDA, 2025).

7.4.3. International Harmonization and Global Equity

With global AI deployment, regulatory harmonization is essential to prevent "equity dumping"—the deployment of systems failing fairness standards in high-income countries into lower-regulation markets. International collaboration on core standards for data diversity, validation, and monitoring can help ensure AI benefits are global, not just Western (European Commission, 2024).

7.5. Economic Restructuring and Aligned Incentives

Sustainable change requires altering the economic calculus to make equity the rational choice for developers and healthcare systems.

7.5.1. Funding Models Prioritizing Equity

Public research funding (e.g., from NIH, NSF) must prioritize projects that demonstrate proactive commitments to data diversity, community engagement, and multi-site validation. Grant applications should be evaluated on their equity plans as rigorously as their technical innovation. Dedicated funding streams should support data collection from underrepresented settings and capacity building at under-resourced institutions.

7.5.2. Reimbursement Tied to Equity Outcomes

Payment structures must reward equitable outcomes. Centres for Medicare & Medicaid Services (CMS) and private insurers could develop value-based purchasing models that offer higher reimbursement for AI tools demonstrating proven, equitable performance across populations, or that are successfully deployed in underserved areas. Conversely, reimbursement could be penalized for tools deployed without adequate validation evidence.

7.5.3. Shared Development Models Reducing Barriers

High development costs incentivize corner cutting. Public-good AI models, developed through open-source consortia or public-private partnerships and validated across diverse sites, could reduce costs and democratize access. Frameworks like the Personal Health Train demonstrate how federated infrastructure can enable collaborative development while preserving data sovereignty (Personal Health Train, 2023).

7.6. Organizational Best Practices for Deployment

Healthcare institutions are the final gatekeepers. They must develop the capacity to critically evaluate and responsibly deploy AI.

7.6.1. Local Validation and Customization

Even broadly validated systems require local assessment. Before full deployment, institutions should conduct internal validation studies to confirm performance on their specific patient population and clinical workflows. This requires investment in local data science and clinical informatics expertise—a capacity that must be built, particularly at community and safety-net hospitals.

7.6.2. Disaggregated Performance Monitoring

Post-deployment, institutions must implement continuous monitoring of AI performance, disaggregated by key demographic groups. This system should track not just algorithmic metrics (sensitivity, specificity) but also clinical outcomes, ensuring the tool is improving care equitably. Mechanisms for clinicians and patients to report suspected errors or biases are essential.

7.6.3. Governance Structures and Accountability

Effective governance requires multidisciplinary committees—including clinicians, data scientists, ethicists, and patient advocates to oversee AI procurement, deployment, and monitoring. These committees must establish clear policies on appropriate use, human oversight, and procedures for addressing adverse events or performance disparities. Accountability must be clear when systems fail.

7.7. Workforce Development and Capacity Building

The equitable AI ecosystem requires a workforce with new, hybrid skills.

7.7.1. Training for Clinicians and Healthcare Leaders

Clinicians need education on AI fundamentals, how to interpret AI outputs critically, and how to recognize potential bias. This training should be integrated into medical school curricula and continuing education. Healthcare leaders need training on procuring, governing, and monitoring AI systems, with a focus on equity implications.

7.7.2. Developing AI Expertise in Under-Resourced Settings

Bridging the digital divide requires dedicated programs to build AI evaluation and implementation capacity at community hospitals, rural facilities, and safety-net institutions. This could involve partnerships with academic centres, specialized training fellowships, and funding for embedded data science roles.

7.7.3. Building Diverse AI Research Pipelines

Long-term solutions require diversifying the AI research pipeline itself. This means sustained investment in STEM education at minority-serving institutions, creation of mentorship and career pathways for underrepresented scholars in AI-for-health, and recognition of equity-cantered research in academic promotion.

7.8. Emerging Technologies and Innovative Approaches

Technical innovation, when guided by equity principles, can provide powerful new tools.

7.8.1. Differential Privacy and Advanced Cryptographic Methods

Beyond federated learning, techniques like differential privacy (adding statistical noise to data or outputs) and homomorphic encryption (computation on encrypted data) can provide stronger privacy guarantees, enabling broader and more secure participation in collaborative data ecosystems (Kairouz et al., 2021).

7.8.2. Synthetic Data Generation for Augmentation

Generative AI can create synthetic medical data representing underrepresented populations, helping to balance training sets. However, synthetic data carries the risk of perpetuating biases in the source data and must be used cautiously as an augmentation to, not a replacement for, real-world data collection.

7.8.3. Explainable AI for Bias Detection and Trust

Explainable AI (XAI) methods that reveal the features influencing a model's decision are vital for debugging bias, building clinician trust, and providing recourse to patients. Research should focus on XAI techniques specifically designed to uncover discriminatory reasoning patterns.

7.9. Integrated Framework: The Equity-centred Development Lifecycle

The solutions above are not a menu of options but interconnected components of an integrated framework. The Equity-centred Development Lifecycle envisions a continuous process where equity is assessed and enforced at every phase:

7.9.1. Planning Phase

- Engagement with diverse stakeholders and affected communities
- Equity impact assessment before development begins
- Explicit equity goals alongside technical objectives
- Resource allocation for comprehensive diversity and validation

7.9.2. Data Phase

- Proactive collection from diverse sources
- Community-engaged data curation
- Federated approaches for privacy-preserving diversity
- Transparent documentation of data composition and limitations

7.9.3. Development Phase

- Diverse, multidisciplinary development teams
- Fairness-aware optimization strategies
- Continuous disaggregated performance monitoring
- Iterative engagement with clinical end-users

7.9.4. Validation Phase

- Comprehensive multi-site external validation
- Disaggregated reporting across demographics and contexts
- Intersectional performance assessment
- Transparent documentation of validation results

7.9.5. Regulatory Phase

- Evidence-based requirements for diversity and external validation
- Disaggregated performance reporting to regulators
- Conditional approval based on equity benchmarks
- Ongoing post-market surveillance requirements

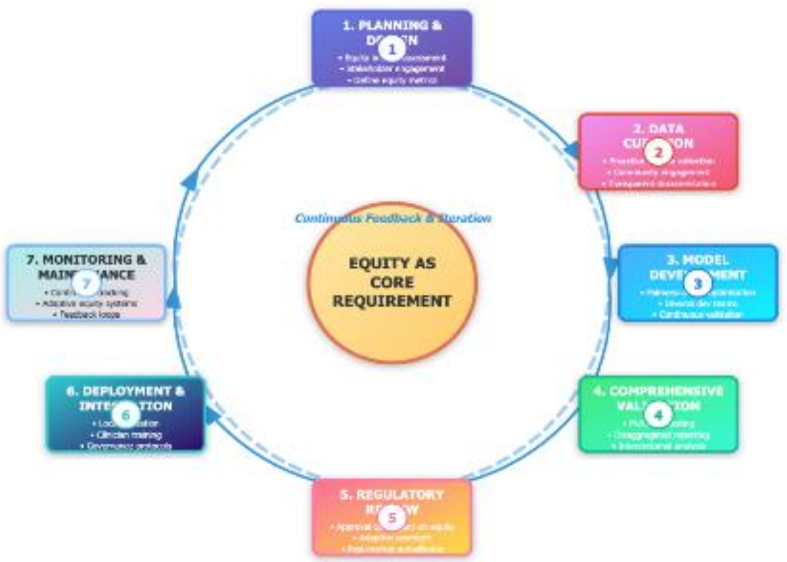
7.9.6. Deployment Phase

- Local institutional validation
- Gradual rollout with continuous monitoring
- Human oversight and clear governance
- Patient and clinician transparency

7.9.7. Monitoring Phase

- Disaggregated real-world performance tracking
- Early detection of degradation or bias
- Mechanisms for user feedback and reporting
- Regular equity audits and reassessment

An iterative, accountable process with equity embedded at every stage



📌 **Key Message**

Unlike the linear pipeline in Figure 1, this equity-centered lifecycle is circular and iterative, with continuous feedback loops connecting all stages. Equity is not an add-on or checkpoint but the central organizing principle embedded at every stage. Each stage feeds back into others: monitoring informs data curation, validation results shape development practices, and deployment experiences guide planning. This approach transforms equity from an aspirational goal into a fundamental requirement that shapes every decision throughout the AI lifecycle.

Figure 5 The Equity-Centered AI Development Lifecycle

This framework provides a concrete blueprint for the paradigm shift from accuracy-focused to equity-cantered AI development.

7.10. An Implementation Roadmap: From Theory to Practice

Translating this comprehensive framework into reality requires a phased, pragmatic strategy with clear accountability.

- **Phase 1: Foundation (1-2 Years):** Regulators mandate disaggregated performance reporting. Funders create grants for diverse data collection. Leading journals require fairness statements and external validation code. Hospitals establish AI governance committees.
- **Phase 2: Integration (3-5 Years):** Mandatory multi-site external validation becomes a clearance requirement. Reimbursement models begin to incorporate equity bonuses/penalties. Collaborative validation networks become operational. AI equity training enters standard medical curricula.
- **Phase 3: Entrenchment (5+ Years):** The equity-cantered lifecycle is the industry standard. Federated learning and privacy-preserving techniques are widespread. International regulatory harmonization on core equity standards is achieved. AI tools developed as public goods are widely accessible.

Table 4 Comparison of Current vs. Proposed Equity-centred Practices

Development Stage	**Current Practice**	**Proposed Equity-centred Practice**	**Expected Impact**
Data Collection	Convenience sampling from available academic centres	Proactive diverse collection; minimum demographic thresholds; federated learning	Training data reflects real-world patient diversity
Model Optimization	Minimize aggregate error (accuracy, AUC)	Fairness-aware optimization; worst-case subgroup performance as primary metric	Models perform reliably across all patient groups
Validation	Single-site internal validation; aggregate metrics	Mandatory multi-site external validation; disaggregated reporting by demographics	Pre-identification of bias and generalizability issues
Regulatory Review	Approval based on technical accuracy; limited post-market requirements	Approval contingent on equity evidence; robust post-market surveillance with disaggregated monitoring	Only equitable systems reach patients; continuous safety monitoring
Deployment	"One-size-fits-all" deployment; vendor claims as primary evidence	Local validation required; clinician training on limitations; clear governance structures	Tools fit local contexts; clinicians use AI appropriately
Economic Models	Reward speed-to-market and impressive lab metrics	Funding/reimbursement tied to equity outcomes; public-good development models	Economic incentives aligned with equitable health outcomes

This roadmap provides stakeholders with a clear path forward, demonstrating that the proposed transformation is ambitious but achievable.

8. Conclusion

This thesis has examined the fundamental paradox of medical artificial intelligence: diagnostic systems that achieve exceptional performance in controlled laboratory settings consistently fail to translate into equitable, robust clinical tools. Through systematic analysis of algorithmic bias, poor generalizability, their intersection, and the structural barriers that sustain them, this research establishes that these translation failures are not isolated technical problems requiring incremental fixes, but predictable outcomes of a development paradigm misaligned with the realities of diverse healthcare ecosystems. The path forward requires nothing less than a fundamental reorientation—from accuracy-centric to equity-cantered AI development.

8.1. Synthesis of Key Findings

The evidence presented reveals a consistent and troubling pattern. AI diagnostic systems trained on geographically concentrated and demographically homogeneous datasets systematically underperform for marginalized populations. Dermatology algorithms show markedly worse accuracy on darker skin tones (Daneshjou et al., 2022), radiology models underdiagnose conditions in female, Black, and low-socioeconomic-status patients (Seyyed-Kalantari et al., 2021), and risk prediction tools assign lower risk scores to Black patients with equivalent health needs, restricting their access to care (Obermeyer et al., 2019). These are not aberrations but the predictable result of optimizing for narrow performance on data that reflects and amplifies existing healthcare disparities.

Simultaneously, these systems demonstrate a critical inability to maintain performance across institutions and contexts. Models achieving over 94% accuracy within development hospitals see performance drop by 8 percentage points when deployed elsewhere (Wong et al., 2021). The geographic concentration of training data—with 71% of U.S. diagnostic AI studies using data exclusively from California, Massachusetts, or New York—produces systems optimized for privileged contexts while failing in the diverse settings where they are most needed (Larson et al., 2018).

Critically, this research demonstrates that bias and poor generalizability are not separate challenges but interconnected manifestations of the same root cause: development practices that prioritize aggregate accuracy on convenient datasets over robust performance for all. Their compounding creates the **equity paradox**: AI diagnostic tools work best for populations with the least need and worst for those who could benefit most from improved diagnostic access. This paradox is sustained by regulatory frameworks that lack mandatory diversity and validation requirements (Wu et al., 2021), economic incentives that reward speed over equity (Kelly et al., 2019), and resource disparities that concentrate both AI development and successful deployment in already well-served settings (Reddy et al., 2020).

8.2. Theoretical Contributions

This thesis makes several distinct contributions to the scholarship on responsible AI in healthcare:

8.2.1. Structural Analysis of Development Practices

While existing literature extensively documents instances of bias or proposes technical mitigations, this work centres a **structural analysis**, demonstrating how fundamental choices in data sourcing, optimization priorities, and validation practices systematically produce inequitable systems. It argues that bias and poor generalizability are features, not bugs, of the current paradigm.

8.2.2. The Interconnection of Bias and Generalizability

The thesis establishes that these are not separate problems requiring distinct solutions but interconnected outcomes of unrepresentative data. It details the compounding mechanisms through which each problem worsens the other, creating particularly severe equity gaps for intersectionally marginalized populations.

8.2.3. The Equity Paradox Framework

The concept of the equity paradox provides a powerful lens for understanding how technological advancement can inadvertently amplify healthcare disparities. It captures the perverse outcome where AI's benefits accrue to the already well-served, transforming a tool of potential democratization into one of further marginalization.

8.2.4. The Equity-Centred Development Lifecycle

Moving beyond critique, the thesis articulates a comprehensive, integrated framework for restructuring AI development. This lifecycle model embeds proactive diversity, fairness-aware optimization, mandatory multi-site validation, and continuous monitoring as non-negotiable requirements at every phase, from planning to post-market surveillance.

8.3. Practical Implications and a Call to Action

The findings demand urgent and coordinated action from all stakeholders in the AI healthcare ecosystem. The technical capacity to build equitable AI exists; what is lacking is the collective will to mandate it.

8.3.1. For Researchers and Developers

The pursuit of impressive accuracy on narrow benchmarks must be recognized as an academic and ethical dead-end. Research must prioritize the collection of diverse, community-engaged datasets, adopt fairness-aware optimization as standard practice, and demand rigorous external validation as a prerequisite for publication. Building on unrepresentative data is no longer scientifically defensible.

8.3.2. For Regulatory Bodies (FDA, EMA, etc.)

Guidance must evolve into requirement. Regulatory approval must be conditional on demonstrated compliance with minimum data diversity standards, evidence from comprehensive multi-site external validation, and robust plans for post-market surveillance with disaggregated performance tracking (FDA, 2025; Wu et al., 2021). The predicate-based 510(k) pathway is inadequate for adaptive AI; new, equity-centred frameworks are needed.

8.3.3. For Healthcare Institutions and Clinicians

FDA clearance cannot be equated with clinical readiness for your specific population. Institutions must build governance capacity, conduct local validation studies, and demand transparent, disaggregated performance data from vendors. Clinicians must be trained as informed, critical users of AI, aware of its potential biases and their role as the final arbiters of patient care.

8.3.4. For Policymakers and Funders

Public investment must be strategically aligned to dismantle the equity paradox. Funding agencies should prioritize grants that demonstrate commitments to data diversity and community partnership. Policymakers should explore reimbursement models that reward equitable outcomes and invest in infrastructure (like federated learning networks) that lower barriers for under-resourced institutions to participate in and benefit from AI.

Limitations and Future Research

This analysis, while comprehensive, points toward necessary future work. The focus has been primarily on diagnostic AI within U.S. and European contexts; research must expand to examine therapeutic AI, clinical decision support, and the unique challenges of deployment in low- and middle-income countries. The proposed Equity-centred Lifecycle requires empirical validation through longitudinal implementation studies. Furthermore, the rapid emergence of large-scale foundation models and generative AI presents a critical new frontier. These models, trained on internet-scale data that embeds societal biases, risk scaling the structural inequities described here to unprecedented levels. Future research must proactively develop frameworks for auditing, validating, and governing these powerful tools to ensure they advance, rather than undermine, global health equity.

8.4. Final Reflection: From Promise to Practice

The journey from laboratory bench to patient bedside is one of validation, trust, and demonstrated value for all. For AI diagnostics, this journey remains incomplete. The translation gap is not a minor technical hurdle, but the direct output of a system optimized for the wrong metrics a system that produces the equity paradox.

The choice before us is stark. We can continue current practices, producing ever-more sophisticated tools that work best for the healthiest and wealthiest, thereby encoding present-day inequities into the healthcare infrastructure of the future or, we can undertake the deliberate, coordinated work of structural reorientation.

This thesis has charted the course for that reorientation: the Equity-centred Development Lifecycle. It is a blueprint for building AI that earns the trust of every community, functions reliably in every clinic, and truly democratizes diagnostic excellence. The promise of AI in medicine is real, but it is a conditional promise. It will be realized not through algorithms alone, but through our collective commitment to build those algorithms justly. The future of AI in healthcare—whether it amplifies or reduces inequity—is not predetermined. It is a choice. We must choose equity.

Compliance with ethical standards

Acknowledgments

The author wishes to acknowledge the use of AI-assisted tools for literature analysis and manuscript preparation. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Disclosure of conflict of interest

The author declares that there is no conflict of interest.

References

[1] Adamson, A. S., & Smith, A. (2018). Machine learning and health care disparities in dermatology. JAMA Dermatology, 154(11), 1247–1248.

[2] Aggarwal, R., Farag, S., Martin, G., Ashrafian, H., & Darzi, A. (2021). Patient perceptions on data sharing and applying artificial intelligence to health care data: Cross-sectional survey. Journal of Medical Internet Research, 23(8), e26162.

[3] AIM-AHEAD Consortium. (2024). Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity. National Institutes of Health.

[4] Banerjee, I., et al. (2021). Reading race: AI recognises patient's racial identity in medical images. The Lancet Digital Health, 3(8), e406–e412.

[5] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. fairmlbook.org.

[6] Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317–1318.

[7] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77–91). PMLR.

[8] Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. The New England Journal of Medicine, 378(11), 981–983.

[9] Chen, I. Y., et al. (2021). Ethical machine learning in healthcare. Annual Review of Biomedical Data Science, 4, 123–144.

[10] Chen, R. J., et al. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature Biomedical Engineering, 7(6), 719–742.

[11] Daneshjou, R., et al. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances, 8(32), eabq6147.

[12] European Commission. (2024). *Regulation (EU) 2024/… of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*.

[13] U.S. Food and Drug Administration (FDA). (2024). Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices.

[14] U.S. Food and Drug Administration (FDA). (2025). Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions.

[15] Gianfrancesco, M. A., et al. (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine, 178(11), 1544–1547.

[16] Gulshan, V., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402–2410.

[17] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29.

[18] **Kairouz, P., et al. (2021). Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1–2), 1–210.

[19] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. Nature Medicine, 25(1), 44–56.

[20] Larson, D. B., et al. (2018). Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging algorithms: Summary and recommendations. Journal of the American College of Radiology, 15(3), 502–507.

[21] Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences, 117(23), 12592–12594.

[22] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

[23] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

[24] Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In Proceedings of the ACM conference on health, inference, and learning (pp. 151–159).

[25] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

[26] Personal Health Train Consortium. (2023). A Distributed Infrastructure for Secure Analysis of Biomedical Data. https://pht.health-ri.nl

[27] Raji, I. D., et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44).

[28] Reddy, S., et al. (2020). A governance model for the application of AI in health care. Journal of the American Medical Informatics Association, 27(3), 491–497.

[29] Rieke, N., et al. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119.

[30] Sendak, M. P., et al. (2020). A path for translation of machine learning products into healthcare delivery. EMJ Innovations, 4(1), 19–001.

[31] Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine, 27(12), 2176–2182.

[32] Shaw, J., et al. (2019). Artificial intelligence and the implementation challenge. Journal of Medical Internet Research, 21(7), e13659.

[33] STANDING Together Initiative. (2024). Standardizing Data to Advance Equity in Health. https://standingtogether.net

[34] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

[35] Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. The New England Journal of Medicine, 383(9), 874–882.

[36] Wiens, J., et al. (2019). Do no harm: a roadmap for responsible machine learning in health care. Nature Medicine, 25(9), 1337–1340.

[37] Wong, A., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine, 181(8), 1065–1070.

[38] Wu, E., et al. (2021). How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nature Medicine, 27(4), 582–584.

[39] Wynants, L., et al. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ, 369, m1328.

[40] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1–19.

[41] Zech, J. R., et al. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15(11), e1002683.

Appendices

APPENDIX A: Proposed Multi-Site External Validation Protocol Template

Purpose

This template operationalizes the comprehensive validation requirements proposed in Section 7.3. It provides a concrete example of the standardized protocol that regulatory bodies (e.g., FDA) could mandate or that collaborative validation networks could adopt to rigorously assess algorithmic bias and generalizability prior to clinical deployment.

Study Objective

To evaluate the diagnostic performance and fairness of [AI System Name/Version] across diverse clinical sites, patient populations, and imaging equipment, assessing its generalizability and identifying any performance disparities across predefined demographic subgroups.

Participating Site Requirements

A minimum of five (5) independent clinical sites must be included, encompassing the following diversity:

- Institutional Types: At least one (1) academic medical centre, one (1) community hospital, and one (1) safety-net or rural hospital.
- Geographic Distribution: Sites must represent at least three (3) distinct U.S. Census regions or two (2) different countries if seeking international validation.

- Equipment Variance: Data must be acquired from at least two (2) different manufacturer models of the primary imaging modality (e.g., Siemens vs. GE CT scanners).

Dataset Composition & Minimum Sample Sizes

The validation dataset must be prospectively assembled or sourced from retrospectively collected, site-specific data not used in model training.

  - **Overall Minimum:** N = [To be determined by intended use and statistical power calculation, e.g., 1000 independent cases].
  - **Disaggregated Minimums:** Each major demographic subgroup must be represented with sufficient power for statistical analysis. Minimum per site:
    - Race/Ethnicity: n≥100 per self-reported category (e.g., Asian, Black, White, Hispanic).
    - Sex: n≥200 for male and female categories.
    - Age: n≥100 for age brackets (e.g., 18-40, 41-65, 65+).
- Intersectional Consideration: The study must report on feasibility of assessing intersectional categories (e.g., Black females 65+) and note if sample sizes are insufficient.

Primary & Secondary Outcomes

- **Primary Outcomes:** Overall sensitivity, specificity, and AUC of the AI system across the entire pooled validation set.
- **Secondary (Equity) Outcomes:** Disaggregated performance metrics (sensitivity, specificity, PPV, NPV) for all demographic categories listed in Section 3.0. The primary fairness metric is the maximum performance gap (the largest absolute difference in sensitivity or specificity between any two demographic subgroups).

Statistical Analysis Plan

- Performance metrics will be reported with 95% confidence intervals.
- Generalizability will be assessed by comparing performance (AUC) **across sites** using a mixed-effects model, with site as a random effect.
- Algorithmic bias will be assessed by testing for significant differences ($p<0.05$, adjusted for multiple comparisons) in secondary outcomes **across subgroups** within and across sites.
- A sample size justification based on the precision of estimating the maximum performance gap must be provided.

Reporting Requirements

Results must be reported in accordance with a modified **TRIPOD+AI statement** and must include:

- A **site-performance matrix** showing outcomes per participating institution.
- A **disaggregated performance table** for all demographic subgroups.
- An **analysis of failure modes**, including case reviews of false negatives/positives stratified by subgroup.
- A clear statement of **clinical contexts and populations for which performance was and was not validated**.

APPENDIX B: Model Fact Sheet (Algorithmic Impact Assessment) Template

Purpose

This Model Fact Sheet template exemplifies the transparency documentation proposed in Section 7.3.2. It serves as an illustrative "nutrition label" for AI diagnostic models—a framework to be completed by developers and required for regulatory submission and hospital procurement. The example values demonstrate the type and granularity of information needed to assess fairness and generalizability risks.

AI Diagnostic Model Fact Sheet

Version: 1.0 **| For Model:** [Model Name & Version]

- Section 1: Intended Use & Scope

  - **Intended Use:** [e.g., Triage of pneumothorax on chest X-rays]

  - **Target Population:** [e.g., Adult patients (18+) presenting to emergency departments in the United States]
  - **Clinical Context:** [e.g., Use as a second reader for board-certified radiologists]
  - **Explicitly Out-of-Scope:** [e.g., Paediatric patients, portable X-rays, use as a fully autonomous diagnostic tool]

- **Section 2: Training Data Provenance**

Table B1. Example of Required Training Data Transparency Documentation

Illustrative example populated with synthetic data to demonstrate the required level of transparency.

Demographic Factor	**Composition %**	**Source(s) & Notes**
Race/Ethnicity	e.g., White: 75%, Black: 12%, Asian: 8%, Other/Unknown: 5%	Derived from Site A (2015-2020), Site B (2018-2021). Labels based on EHR self-report.
Sex Assigned at Birth	Female: 45%, Male: 55%	--
Age	Mean: 58 ± 16 years	--
Geographic Origin	e.g., Data from 3 hospitals in Massachusetts and 1 in California.	
Clinical Setting	e.g., 100% from inpatient academic medical centres.	
Data Source & Curation	Total N: 50,000 images. Curation Note: Images with technically poor quality were excluded by a radiologist.	

- Section 3: Known Performance Characteristics & Gaps

  - **Overall Performance (Internal Test Set):** Sensitivity: 88% (CI: 85-90%), Specificity: 94% (CI: 92-95%).
  - **Disaggregated Performance (Internal):** See Table B1. [*Example: Sensitivity for Black patients was 82% (CI: 75-88%) vs. 90% (CI: 87-92%) for White patients. *]
  - **External Validation Status:** ☐ Not Performed ☐ Performed on 1 external site ☐ Performed per Appendix A protocol.
  - If performed, attach summary report.
  - **Known Performance Gaps:** [e.g., "Performance degraded on data from Hospital Z using Brand Y X-ray machines. Sensitivity for patients over 80 was lower in internal testing."]
- Section 4: Bias Mitigation & Fairness Measures
  - **During Training:** [e.g., "Class-balanced sampling was used. Adversarial debiasing was attempted to reduce correlation with race."]
  - **During Validation:** [e.g., "Disaggregated testing was conducted. The model was evaluated against equalized odds difference, which was <0.05."]
  - **Post-Deployment:** [e.g., "The PCCP includes monthly monitoring of sensitivity by race/ethnicity."]

- **Section 5: Recommended Monitoring & Governance**
  - **Key Equity Metrics to Monitor in Production:** Sensitivity by race, age, and sex; site-specific AUC.
  - **Recommended Audit Frequency:** Quarterly disaggregated review.
  - **Recommended Clinical Governance:** This model should not be used as the sole diagnostic criterion. Clinicians should be made aware of the potential for reduced sensitivity in elderly and Black patient populations.

APPENDIX C: Semi-Structured Interview Guide for Stakeholder Analysis

Purpose

This guide outlines the methodological approach for the qualitative research component proposed in Sections 7.2.3 and 7.6.3. It is designed to elicit in-depth insights from key stakeholders (developers, regulators, clinicians) about the perceived barriers and facilitators to implementing an equity-centred lifecycle, grounding the theoretical framework in practical realities.

Study Aim

To understand stakeholder perspectives on the operational, economic, and ethical challenges of developing, validating, and deploying equitable and generalizable AI diagnostic systems.

Participant Groups

- **AI Developers/Researchers** (n=10-15): From industry and academia.
- **Regulatory Affairs Professionals** (n=5-10): From FDA and notified bodies.
- **Clinician End-Users & Healthcare Executives** (n=10-15): Radiologists, cardiologists, and hospital CIOs/CMOs.

Informed Consent & Introduction

- Explain study purpose, confidentiality, recording procedures.
- **Opening Question:** "Can you describe your role and experience with AI diagnostic tools in healthcare?"

Interview Domains and Questions

- **Domain 1 Perceptions of the Problem**
  - How significant do you believe the problems of algorithmic bias and poor generalizability are in today's AI diagnostics?
  - What do you see as the primary root causes of these issues? (Probe: data, incentives, regulation, speed-to-market).

- **Domain 2: Barriers to Equitable Development**
  - From your perspective, what are the biggest practical barriers to collecting more diverse and representative training data?
  - What are the main disincentives (economic, competitive, regulatory) for conducting comprehensive multi-site external validation?
  - How do current product development timelines and funding models help or hinder a focus on fairness and robustness?

- **Domain 3: Feasibility of Proposed Solutions**
  - I will describe a few proposed interventions (e.g., mandatory validation templates—Appendix A, required model **fact sheets Appendix B**). What is your reaction to their feasibility and potential effectiveness?
  - What would be the single most impactful change a regulator (like the FDA) could make to improve equity in AI diagnostics?
  - What would a hospital need, in terms of resources and expertise, to properly validate and monitor an AI tool for equitable performance locally?

- **Domain 4: Responsibility & Governance**
  - Who do you believe should hold primary responsibility for ensuring an AI diagnostic tool performs fairly across all patient groups?

  - What does effective institutional governance of AI look like in a clinical setting?

Closing Question

- Is there anything we haven't discussed that you feel is critical for achieving equitable AI diagnostics?
- Do you have any recommendations for this research project?

Proposed Analysis Plan

Interviews will be transcribed, de-identified, and analyzed using thematic analysis to identify convergent and divergent themes across stakeholder groups, which will directly inform the refinement of the proposed Equity-Centred Lifecycle framework.