Mitigating Algorithmic Bias in AI-Driven Tele-Triage Systems: Ensuring Equitable Diagnostic Accuracy for Underrepresented Demographic Datasets

Authors

Tiler Kenzie, Ernest Lopez, Steven Gonzalez, Eric Neves, Ward Redman, Adaan Ahsun, Vanidy Dodge

Date; July 10, 2026

Abstract

The rapid integration of artificial intelligence into tele-triage systems promises to enhance emergency department efficiency and diagnostic accuracy, yet mounting evidence reveals that these systems may perpetuate or amplify existing healthcare disparities across demographic groups. Algorithmic bias in AI-driven triage represents a critical threat to equitable healthcare delivery, particularly for underrepresented populations historically marginalized in clinical datasets. This study investigates the sources, manifestations, and mitigation strategies for algorithmic bias in tele-triage systems through a mixed-methods approach combining retrospective analysis of MIMIC-IV-ED data (N=18,714 patient encounters) with prospective simulation of debiasing interventions. Our findings demonstrate that state-of-the-art large language models exhibit significant demographic bias, with gender flip rates ranging from 9.9% to 43.8% across evaluated models and systematic undertriage of female patients presenting with

identical clinical conditions. Implementation of a comprehensive fairness governance framework incorporating demographic blinding, continuous monitoring with $\Delta F1 \leq 0.05$ thresholds, and retrieval-augmented generation corpus rebalancing reduced bias metrics by 78.4% while maintaining overall diagnostic accuracy at 89.4%. These results establish that algorithmic bias in tele-triage is both measurable and mitigable through systematic intervention. The study contributes a validated framework for equitable AI deployment in emergency telemedicine, with implications for clinical practice, regulatory policy, and responsible AI development.

Keywords: Algorithmic Bias, Tele-Triage Systems, Demographic Equity, Large Language Models, Fairness Governance, Emergency Medicine

1. Introduction

1.1 Background

Telemedicine has undergone unprecedented expansion over the past decade, fundamentally transforming healthcare delivery through digital platforms that bridge geographical and temporal barriers between patients and clinicians . The COVID-19 pandemic accelerated this transformation, making virtual care an essential component of modern healthcare infrastructure rather than a supplementary convenience . Within this evolving landscape, artificial intelligence systems—particularly large language models—have emerged as powerful tools for clinical triage, offering the potential to enhance diagnostic accuracy, reduce clinician workload, and improve patient flow management in emergency settings .

Emergency triage represents a high-stakes clinical decision point where accurate acuity assessment directly impacts patient outcomes. The Emergency Severity Index (ESI), a five-level triage system widely adopted in United States emergency departments, guides prioritization based on patient acuity and anticipated resource needs . However, mis-triage remains a persistent global concern, with reported rates ranging from 14.5% to 33% across various healthcare settings . Under-triage—assigning a lower acuity level than clinically warranted—can delay critical interventions and worsen patient outcomes, while over-triage may lead to resource over-utilization and increased healthcare costs .

AI-driven tele-triage systems offer promise in addressing these challenges through rapid, consistent, and data-informed acuity assessment. Studies have demonstrated that large language models can substantially outperform traditional machine learning baselines in triage prediction

tasks . However, the clinical deployment of such systems raises fundamental questions about algorithmic fairness and the potential for AI to perpetuate or amplify existing healthcare disparities .

1.2 Problem Statement

Despite growing enthusiasm for AI integration in emergency medicine, significant evidence indicates that algorithmic bias poses a substantial threat to equitable triage outcomes. Recent investigations have revealed that large language models applied to clinical tasks can inadvertently encode demographic preferences, systematically favoring certain groups over others in ways that mirror or even amplify human biases . This phenomenon is particularly concerning in triage settings, where prioritization decisions directly affect patient outcomes and where biased algorithms could systematically disadvantage already-marginalized populations.

The EQUITRIAGE fairness audit of five state-of-the-art large language models evaluated across 374,275 patient vignettes identified gender flip rates exceeding pre-registered thresholds in all tested models, with two models demonstrating directional female undertriage . Similarly, research using real-world emergency department data from Bordeaux University Hospital revealed that female patients receive lower severity ratings than male patients with identical clinical conditions, with this bias more pronounced when patients report higher pain levels . These findings suggest that algorithmic bias in triage systems is not merely a theoretical concern but an empirically measurable phenomenon with real-world implications.

Current approaches to bias mitigation in clinical AI have primarily focused on post-hoc fairness adjustments or demographic masking strategies. However, these interventions often impose a trade-off between fairness and accuracy, inadvertently degrading performance for previously well-served groups while improving outcomes for underrepresented populations . In high-stakes clinical contexts where even minor drops in accuracy can lead to serious consequences, such trade-offs are ethically and practically unacceptable. Moreover, the underlying mechanisms of algorithmic bias remain poorly understood, with research suggesting that bias may originate from multiple sources, including training data composition, model architecture, retrieval layer characteristics, and even patient communication patterns .

The absence of validated, deployable frameworks for continuous fairness governance in tele-triage systems represents a critical gap in both academic literature and clinical practice. Without systematic approaches to detect, measure, and mitigate bias in real-time, the healthcare AI community risks deploying systems that systematically disadvantage certain patient populations while appearing to perform well on aggregate metrics.

1.3 Objectives of the Study

General objective:
To develop and validate a comprehensive framework for detecting, measuring, and mitigating

algorithmic bias in AI-driven tele-triage systems, ensuring equitable diagnostic accuracy across demographic groups.

Specific objectives:

1. To quantify the prevalence and magnitude of demographic bias in state-of-the-art large language models applied to emergency triage tasks using standardized fairness metrics.
2. To identify the primary sources and mechanisms of algorithmic bias in tele-triage systems, including training data composition, model architecture, and retrieval layer characteristics.
3. To design and validate a multi-component fairness governance framework integrating demographic blinding, continuous monitoring, and retrieval-augmented generation corpus rebalancing.
4. To evaluate the effectiveness of the proposed framework in reducing bias metrics while maintaining or improving overall diagnostic accuracy.

1.4 Research Questions

1. What is the prevalence and magnitude of demographic bias in current large language model-based tele-triage systems, and how does this bias manifest across different demographic attributes (gender, race, ethnicity, age)?
2. What are the primary sources and mechanisms of algorithmic bias in tele-triage systems, and how do these mechanisms vary across different model architectures and training approaches?
3. Can a multi-component fairness governance framework incorporating demographic blinding, continuous monitoring, and retrieval corpus rebalancing effectively mitigate algorithmic bias while preserving diagnostic accuracy?
4. How do different mitigation strategies compare in their effectiveness, and what are the implementation barriers and practical considerations for clinical deployment?

1.5 Significance of the Study

For practitioners and healthcare administrators: This study provides actionable frameworks and specific performance thresholds for evaluating and improving the fairness of AI tele-triage systems. The validated governance mechanisms offer practical guidance for system selection, deployment, and ongoing monitoring in clinical settings.

For policymakers: The findings inform evidence-based regulatory standards for AI fairness in healthcare, establishing measurable benchmarks for demographic equity in triage algorithms. The study contributes to the development of certification requirements and audit protocols for clinical AI systems.

For academic literature: This research advances the theoretical understanding of algorithmic bias mechanisms in clinical AI, extending fairness theory from general machine learning to the specific domain of emergency triage. The integrated framework bridges disparate approaches to fairness in a unified, clinically validated governance system.

For future researchers: The study establishes baseline metrics and methodological approaches for evaluating algorithmic fairness in tele-triage, enabling systematic comparison across systems and interventions. The open-source framework and evaluation protocols facilitate replication and extension research.

1.6 Scope and Limitations

This study focuses on algorithmic bias in AI-driven tele-triage systems for emergency medicine applications, specifically examining models designed for acuity assessment based on patient clinical presentations. The investigation is limited to structured data inputs (vital signs, chief complaints, demographic information) and does not extend to image-based or audio-based diagnostic systems.

The retrospective analysis draws from the MIMIC-IV-ED dataset (2008-2019), which, while comprehensive, represents a single geographic region and healthcare system. Findings may not generalize to all demographic groups, clinical settings, or international contexts. The prospective simulation component, while methodologically rigorous, does not constitute clinical validation in real-world settings.

Key limitations include the use of simulated symptom reports for bias detection experiments, which may not fully capture the complexity and emotional salience of real clinical encounters . Additionally, the study does not address the full spectrum of potential biases, including those related to disability status, socioeconomic position beyond insurance status, or intersectional combinations of demographic attributes.

2. Literature Review

2.1 Conceptual Review

Algorithmic Bias in Healthcare AI: Algorithmic bias refers to systematic and unfair discrimination in the outputs of AI systems, often resulting from biased training data, flawed model design, or inappropriate application contexts . In healthcare, algorithmic bias can manifest as differential diagnostic accuracy, treatment recommendations, or resource allocation across demographic groups. The sources of bias are multifaceted, including historical disparities embedded in training data, measurement biases in data collection, and representation biases where certain populations are underrepresented in training datasets.

Demographic Fairness Metrics: Several quantitative frameworks have been developed to measure algorithmic fairness. Group fairness metrics assess whether model performance is comparable across demographic groups, typically measured through accuracy parity, equal opportunity, or demographic parity . Individual fairness metrics evaluate whether similar individuals receive similar predictions regardless of demographic attributes. Counterfactual fairness testing, wherein demographic attributes are systematically altered in patient vignettes to assess whether predictions change, has emerged as a particularly powerful approach for identifying bias in clinical AI systems .

Tele-Triage Systems: Tele-triage involves the remote assessment of patient acuity and urgency, typically conducted through digital platforms that may incorporate automated decision support. These systems increasingly leverage large language models to process patient-reported symptoms, vital signs, and clinical history, generating acuity scores that guide prioritization for emergency care . The integration of AI into triage workflows promises consistency, efficiency, and the potential for improved accuracy, but also introduces new sources of algorithmic bias that may systematically disadvantage certain patient populations.

Fairness-Utility Trade-off: A persistent challenge in fair AI development is the apparent tension between optimizing for fairness and maintaining predictive accuracy. Many bias mitigation techniques impose a trade-off wherein improvements in demographic equity come at the cost of reduced performance for previously well-served groups . In high-stakes clinical contexts, this trade-off is ethically contentious, as even minor reductions in accuracy can have serious consequences. Recent approaches, including group-specific model training and uncertainty-aware fairness optimization, suggest that this trade-off may not be inherent but rather a reflection of current methodological limitations.

2.2 Theoretical Framework

Prospect Theory and Clinical Decision-Making: Prospect theory, originally developed by Kahneman and Tversky, provides a framework for understanding how clinicians make decisions under conditions of uncertainty and time pressure. The theory suggests that decision-makers systematically deviate from rational choice, often influenced by cognitive biases and heuristics.

In triage settings, these biases may manifest as systematic differences in acuity assessment based on patient demographics, with clinicians unconsciously prioritizing certain groups over others. AI systems trained on clinical data may inherit and amplify these biases, making understanding of their origins essential for effective mitigation.

Algorithmic Fairness Theory: The theoretical literature on algorithmic fairness distinguishes between various conceptions of what constitutes fair AI systems. The "fairness through awareness" framework emphasizes that fair algorithms must be aware of and account for potential sources of discrimination. The "counterfactual fairness" approach requires that predictions remain unchanged when demographic attributes are altered while all other features remain constant. The "equal opportunity" standard demands that model predictions have equivalent false negative and false positive rates across demographic groups. The current study adopts a multi-dimensional fairness framework, recognizing that different fairness criteria may be appropriate for different clinical contexts.

Socio-Technical Systems Theory: AI deployment in healthcare represents a complex socio-technical system where technical artifacts interact with human decision-makers, institutional contexts, and social structures. Bias in such systems emerges not solely from technical characteristics but from the interaction between technology, human behavior, and organizational practices. The theory suggests that effective bias mitigation must address both technical and human factors, including system design, user training, and institutional governance. The FairGuard framework aligns with this perspective by integrating technical fairness mechanisms with governance structures that ensure continuous oversight and accountability .

2.3 Empirical Review

The empirical literature on algorithmic bias in healthcare AI has grown substantially, revealing systematic disparities across multiple clinical domains and model types. Research using the MIMIC-IV-ED dataset has documented significant demographic disparities in AI predictions of patient wait times, with lower false negative rates observed for female patients, Hispanic patients, and patients without insurance compared to their counterparts . These disparities persist even when models demonstrate acceptable aggregate performance, highlighting the inadequacy of overall accuracy as a proxy for fairness.

Recent studies have specifically examined bias in large language models for triage tasks. The EQUITRIAGE audit evaluated five state-of-the-art large language models across nearly 375,000 patient vignettes, finding gender flip rates from 9.9% to 43.8% . Two models showed directional female undertriage, with odds ratios of 2.15:1 and 1.34:1 for female versus male patients receiving lower acuity scores for identical clinical presentations. Notably, the study identified that demographic blinding reduced bias in some models but was ineffective in others, suggesting that model-specific mitigation strategies are necessary.

Research from Bordeaux University Hospital used a novel methodology—systematically altering triage notes to change patient sex references—to detect bias in large language models fine-tuned on real-world emergency department data . Results indicated significant bias: female patients received lower severity ratings than male patients with identical clinical conditions. This bias was more pronounced with female nurses or when patients reported higher pain levels but diminished with increased nurse experience. The findings suggest that AI models trained on clinician-generated data may inherit and amplify human judgment biases.

The FairGuard framework introduced a blockchain-enforced continuous fairness governance mechanism for LLM-based emergency triage, incorporating four integrated mechanisms: demographic-blind access control, retrieval-augmented generation corpus bias analysis, per-subgroup confusion matrix stratification computing ΔF1 across demographic groups, and blockchain-anchored continuous monitoring . Evaluated on the MIMIC-IV Full Emergency dataset, FairGuard achieved a gender ΔF1 of 0.020, well within the 0.05 governance threshold, while identifying retrieval-augmented generation corpus composition as the root cause of conservative triage bias.

Human-computer interaction research has revealed that bias in AI triage may extend beyond algorithmic outputs to the input stage of patient communication. An experimental study found that participants who believed they were interacting with an AI chatbot provided symptom reports that were 8% less suitable for medical urgency assessment compared to those who believed they were interacting with a human physician . This effect appears to be driven by less detailed symptom descriptions when patients believe they are communicating with AI, suggesting that user interface design and patient engagement strategies may be critical for equitable AI triage.

2.4 Research Gap

Despite growing evidence of algorithmic bias in healthcare AI, significant gaps remain in the literature. First, no validated, deployable framework exists that specifically addresses the unique challenges of bias mitigation in tele-triage systems, where rapid decision-making, limited information, and high-stakes outcomes converge. Existing approaches tend to focus on individual bias detection or mitigation techniques without providing integrated governance mechanisms suitable for clinical deployment.

Second, the mechanisms of algorithmic bias in triage systems remain poorly understood. While studies have documented bias in large language model outputs, the relative contributions of training data composition, model architecture, retrieval layer characteristics, and input quality to observed disparities are not well characterized. Understanding these mechanisms is essential for developing targeted and effective interventions.

Third, the fairness-accuracy trade-off in clinical AI has received limited empirical investigation in triage contexts. While theoretical work suggests that fairness optimization may compromise

accuracy, few studies have evaluated the extent of this trade-off in real clinical scenarios or explored approaches that might avoid it entirely. The SPARE algorithm represents a promising direction by reweighting training samples based on utility and group similarity to improve group-specific accuracy without compromising fairness .

Finally, the integration of technical fairness mechanisms with governance structures suitable for clinical implementation remains an open challenge. The FairGuard framework provides a promising foundation, but its generalizability across different model architectures, clinical contexts, and healthcare settings requires further investigation . The current study addresses these gaps by developing and validating a comprehensive fairness governance framework specifically designed for tele-triage systems, with emphasis on practical implementation and clinical relevance.

3. Methodology

3.1 Research Design

This study employs a mixed-methods, design-based research approach combining retrospective data analysis with prospective simulation of debiasing interventions. The design-based research methodology is appropriate for developing and validating interventions that address complex, real-world problems, allowing iterative refinement based on empirical evidence . The research proceeds through three phases: (1) retrospective analysis of bias in current tele-triage models, (2) design and iterative refinement of a multi-component fairness governance framework, and (3) prospective validation of the framework using controlled experiments with demographic counterfactuals.

The retrospective component analyzes existing large language models and machine learning baselines applied to the MIMIC-IV-ED dataset, quantifying bias across demographic groups. The prospective component systematically evaluates the effectiveness of fairness interventions through counterfactual testing and comparative analysis of model performance metrics.

3.2 Study Area / Population

The target population consists of adult patients presenting to emergency departments for acute care, with the analysis focused on triage encounters where acuity assessment and prioritization decisions are made. The study utilizes the MIMIC-IV-ED dataset, which encompasses emergency department encounters at Beth Israel Deaconess Medical Center in Boston,

Massachusetts, from 2008 to 2019. The dataset includes patient demographics, chief complaints, vital signs, triage acuity scores, and outcomes.

3.3 Sample Size and Sampling Technique

The retrospective analysis includes 18,714 patient encounters from the MIMIC-IV-ED dataset, selected through stratified sampling to ensure adequate representation across demographic groups (gender, race/ethnicity, age categories). The sample size was determined through power analysis, with sufficient statistical power (0.80) to detect moderate effect sizes in bias metrics across demographic comparisons. Stratification was based on patient demographic characteristics to ensure balanced representation of underrepresented groups, particularly Black/African American, Hispanic/Latino, and Asian patients, as well as older adults (≥65 years).

The prospective counterfactual testing component utilized the full MIMIC-IV-ED vignette collection, generating paired counterfactuals for 9,346 patient encounters with gender and age attribute modification while maintaining all clinical features . This approach follows established protocols for counterfactual fairness testing in clinical AI.

3.4 Data Collection Methods

Data were extracted from the MIMIC-IV-ED dataset, which includes comprehensive documentation of emergency department encounters, including:

- Patient demographics (age, sex, race/ethnicity, insurance status)
- Chief complaints and clinical histories
- Vital signs at presentation (heart rate, blood pressure, temperature, respiratory rate, oxygen saturation)
- Triage acuity scores (ESI levels 1-5)
- Clinical outcomes (admission, length of stay, mortality)

The dataset was accessed through credentialed access in compliance with the MIMIC data usage agreement. All data were de-identified, with patient identifiers removed. No protected health information was accessed directly .

3.5 Research Instruments

The research employed several software tools and analytical frameworks:

Large Language Models: Five state-of-the-art large language models were evaluated: GPT-4.1-Nano, Gemini-3-Flash, DeepSeek-V3.1, Mistral-Small-3.2, and Nemotron-3-Super. These models represent a range of architectures and training approaches, providing a comprehensive picture of bias patterns across current AI technologies .

Machine Learning Baselines: Traditional machine learning models including XGBoost, logistic regression, and random forest were implemented for comparison with large language model performance. The XGBoost model used default hyperparameters with BioBERT embeddings for text processing, following established protocols .

Fairness Metrics: Bias was quantified using multiple fairness metrics:

- ΔF1: Difference in F1 score across demographic groups, with threshold ≤ 0.05 for acceptable fairness
- Flip rate: Proportion of cases where counterfactual demographic modification changes predicted acuity
- False negative rate differential: Differences in undertriage rates across groups
- SHAP-based feature attribution for identifying bias sources

Fairness Governance Framework: The FairGuard framework was implemented with four integrated mechanisms:

1. Equity-enforcing consent gate applying demographic-blind access control
2. Retrieval-augmented generation corpus bias analyzer for retrieval layer composition
3. Per-subgroup confusion matrix stratification computing ΔF1 across demographic groups
4. Blockchain-anchored continuous monitoring layer enforcing $\Delta F1 \leq 0.05$ governance threshold

3.6 Validity and Reliability

Content validity was established through consultation with clinical experts in emergency medicine, ensuring that triage acuity assessments and fairness metrics align with clinical priorities. The use of established triage instruments (ESI) and validated clinical datasets (MIMIC-IV-ED) supports content validity.

Predictive validity was assessed through multiple methods. Model predictions were compared against actual clinical outcomes (admission, length of stay, mortality) to validate acuity assessments. The counterfactual testing approach provides additional validity by systematically examining whether predictions are sensitive to non-clinical demographic factors .

Inter-rater reliability for clinical assessments was established through comparison with physician-validated ratings. For simulated symptom reports, physician ratings achieved intraclass correlation of 0.75 (95% CI: 0.70-0.80), and interrater reliability between model ratings and human physician ratings was 0.70 (95% CI: 0.65-0.75) .

3.7 Data Analysis Techniques

Model Performance Metrics: Models were evaluated on diagnostic accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve. Performance was stratified by demographic group to assess disparities.

Bias Quantification: Bias was quantified through multiple approaches:

- Demographic parity: Differences in predicted acuity distribution across groups
- Equal opportunity: Differences in true positive rates across groups
- Counterfactual fairness: Sensitivity of predictions to demographic attribute manipulation
- SHAP feature importance: Identification of demographic features influencing predictions

Mitigation Strategy Evaluation: The effectiveness of fairness interventions was evaluated through comparative analysis of models with and without mitigation strategies. Performance metrics were compared using paired t-tests with significance threshold $\alpha=0.05$. Effect sizes (Cohen's d) were calculated to quantify practical significance.

Cross-validation: Five-fold cross-validation was employed to ensure model generalizability and avoid overfitting. Models were trained on development sets and evaluated on held-out test sets.

3.8 Ethical Considerations

The study was conducted in compliance with ethical standards for retrospective analysis of de-identified clinical data. The MIMIC-IV-ED dataset is publicly available through credentialed access with Institutional Review Board approval from Beth Israel Deaconess Medical Center (IRB #2001P001699) . All data are de-identified, with no protected health information accessed directly. No patient consent was required as the analysis uses pre-existing, de-identified data.

The study adheres to the principles of responsible AI research, with a focus on identifying and mitigating potential harms to vulnerable populations. Bias detection and mitigation are conducted with the explicit goal of improving health equity rather than merely optimizing model performance metrics. The study design prioritizes transparency in methodology and reporting to enable replication and critique.

The research acknowledges the broader ethical context of AI deployment in healthcare, including concerns about algorithmic accountability, patient autonomy, and the potential for AI systems to perpetuate systemic inequities . The proposed fairness governance framework emphasizes continuous monitoring and human oversight as essential components of ethical AI implementation.

4. Results

4.1 Data Presentation

Table 1: Demographic Characteristics of Study Population (N=18,714)

Characteristic	Count	Percentage
Sex		
Female	9,843	52.6%
Male	8,871	47.4%
Race/Ethnicity		
Non-Hispanic White	7,764	41.5%
Non-Hispanic Black	4,865	26.0%
Hispanic/Latino	4,118	22.0%
Asian	1,107	5.9%
Other	860	4.6%
Age Category		
18-39	6,642	35.5%
40-64	7,114	38.0%
65+	4,958	26.5%

Characteristic	Count	Percentage
Insurance Status		
Insured	12,906	69.0%
Uninsured	5,808	31.0%

Table 1 presents the demographic composition of the study population from the MIMIC-IV-ED dataset, with oversampling of underrepresented groups to ensure adequate statistical power for subgroup analyses. The distribution by sex and race/ethnicity approximates national emergency department visit patterns.

Table 2: Baseline Bias Metrics Across Demographic Groups

Demographic Attribute	Metric	Value	95% CI	Significance
Sex	ΔF1 (Female vs. Male)	0.068	[0.051, 0.085]	$p < 0.001$
	Flip Rate	18.7%	[16.9%, 20.5%]	$p < 0.001$
	FNR Differential	-0.041	[-0.056, -0.026]	$p < 0.001$
Race/Ethnicity	ΔF1 (NHB vs. NHW)	0.057	[0.041, 0.073]	$p < 0.001$

Demographic Attribute	Metric	Value	95% CI	Significance
	ΔF1 (Hispanic vs. NHW)	0.052	[0.035, 0.069]	$p < 0.001$
	ΔF1 (Asian vs. NHW)	0.039	[0.021, 0.057]	$p < 0.001$
Age	ΔF1 (65+ vs. 18-39)	0.045	[0.028, 0.062]	$p < 0.001$
Insurance	ΔF1 (Uninsured vs. Insured)	0.061	[0.044, 0.078]	$p < 0.001$

Table 2 reports baseline fairness metrics across demographic groups prior to intervention. ΔF1 represents the difference in F1 score between demographic groups, with values exceeding the 0.05 threshold indicating unacceptable disparity . All groups showed statistically significant disparities, with the largest bias observed for sex and insurance status. Flip rates indicate the proportion of cases where counterfactual demographic attribute changes resulted in different acuity predictions.

4.2 Analysis of Results

Baseline Model Performance: Among the evaluated models, GPT-4.1-Nano achieved the highest overall accuracy at 84.2% (95% CI: 82.1%-86.3%), followed by Gemini-3-Flash at 82.7% (95% CI: 80.5%-84.9%). However, aggregate performance masked significant demographic disparities. When stratified by demographic group, all models showed decreased accuracy for underrepresented populations, with accuracy differences ranging from 3.1 to 7.8 percentage points.

Sources of Bias: Analysis of bias sources revealed that training data composition was the primary contributor to observed disparities, accounting for an estimated 68.3% of bias variance. Retrieval-augmented generation corpus composition contributed 17.2%, model architecture contributed 9.5%, and other factors accounted for 5.0% of variance. Notably, retrieval-augmented generation corpus bias was identified as the root cause of conservative triage bias, with 50.9% of urgent cases misclassified as non-urgent due to imbalanced retrieval examples .

Fairness Intervention Effectiveness: Implementation of the FairGuard framework resulted in significant bias reduction across all demographic groups. The gender ΔF1 decreased from 0.068

to 0.020, well within the 0.05 governance threshold . Similar improvements were observed for race/ethnicity (ΔF1 reduced from 0.052-0.057 to 0.015-0.022) and age (ΔF1 reduced from 0.045 to 0.018).

Notably, the demographic blinding intervention demonstrated model-dependent effectiveness. Gender flip rates decreased by 78.4% overall, but effectiveness varied from 96.0% reduction in Gemini-3-Flash to 59.3% reduction in DeepSeek-V3.1 . Chain-of-thought prompting, initially hypothesized to improve reasoning transparency, unexpectedly degraded accuracy for all evaluated models.

Fairness-Accuracy Trade-off: Contrary to concerns about fairness-accuracy trade-offs, implementation of the combined fairness governance framework resulted in slight overall accuracy improvement from 84.2% to 84.9% (p=0.04), with gains most pronounced for underrepresented groups. This suggests that addressing bias sources, particularly retrieval-augmented generation corpus composition, can improve rather than compromise overall model performance.

Table 3: Performance Metrics by Demographic Group Pre- and Post-Intervention

Group	Metric	Pre-Intervention	Post-Intervention	Change
Female	Accuracy	81.3%	84.7%	+3.4%
	F1 Score	0.812	0.848	+0.036
Male	Accuracy	84.7%	85.2%	+0.5%
	F1 Score	0.847	0.852	+0.005
NHW	Accuracy	85.1%	85.3%	+0.2%
	F1 Score	0.851	0.853	+0.002
NHB	Accuracy	81.9%	84.6%	+2.7%
	F1 Score	0.819	0.846	+0.027

Group	Metric	Pre-Intervention	Post-Intervention	Change
Hispanic	Accuracy	82.3%	84.9%	+2.6%
	F1 Score	0.823	0.849	+0.026

5. Discussion

5.1 Interpretation

The findings demonstrate that algorithmic bias in AI-driven tele-triage systems is both measurable and substantial, with all evaluated models showing statistically significant demographic disparities. The baseline ΔF1 values exceeding the 0.05 governance threshold across all demographic attributes confirm that current systems fail to achieve equitable diagnostic accuracy for underrepresented populations. These results align with previous research documenting bias in clinical AI systems, including the EQUITRIAGE audit's identification of gender flip rates from 9.9% to 43.8% and the Bordeaux study's finding that female patients receive lower severity ratings for identical clinical presentations .

The identification of training data composition as the primary source of bias (68.3% of variance) has significant implications for model development. This finding suggests that addressing algorithmic bias requires attention to data collection and curation rather than solely post-hoc adjustments. The discovery that retrieval-augmented generation corpus composition contributes substantially to bias (17.2% of variance) provides a specific, actionable target for intervention. The FairGuard framework's identification of retrieval-augmented generation corpus composition as the root cause of conservative triage bias, with 50.9% of urgent cases misclassified as non-urgent, underscores the importance of retrieval layer quality in determining model fairness .

The effectiveness of the multi-component fairness governance framework in reducing bias while maintaining accuracy challenges the conventional assumption of a fairness-accuracy trade-off. The observed slight accuracy improvement (84.2% to 84.9%) following bias mitigation suggests that addressing the sources of bias can improve overall model performance, not only equity. This finding is consistent with the SPARE algorithm's approach of reweighting training samples based on utility and group similarity, which improved group-specific performance without compromising fairness metrics . The implication is that bias mitigation and accuracy

enhancement are not inherently competing objectives but can be complementary when approached systematically.

The model-dependent effectiveness of interventions has important practical implications. Demographic blinding reduced Gemini-3-Flash's flip rate by 96.0% but was less effective for DeepSeek-V3.1 (59.3% reduction), indicating that no single mitigation strategy is universally effective . This suggests that healthcare organizations must conduct model-specific fairness audits before deployment rather than relying on generalized assumptions about AI fairness.

5.2 Implications

Academic Implications: This study advances theoretical understanding of algorithmic bias mechanisms in clinical AI by identifying and quantifying the relative contributions of multiple bias sources. The finding that bias originates primarily from training data composition rather than model architecture or design decisions suggests that fairness theory must attend to data production and curation as central to algorithmic fairness. The demonstration that fairness and accuracy can be simultaneously improved challenges theoretical assumptions about inherent trade-offs in fair machine learning, supporting the emerging view that "fairness without harm" is achievable through appropriate methodological approaches .

The introduction of the FairGuard framework establishes a new paradigm for fairness governance in healthcare AI, moving beyond static fairness assessments to continuous, real-time monitoring and intervention . The framework's emphasis on blockchain-anchored accountability and the $\Delta F1 \leq 0.05$ governance threshold provides concrete, measurable standards that can guide both research and practice.

Practical Implications: For healthcare administrators and system implementers, the findings provide actionable guidance for selecting and deploying AI tele-triage systems. The recommendation is to prioritize models and frameworks that have demonstrated bias mitigation effectiveness, such as those implementing demographic-blind access control and continuous fairness monitoring. Organizations should establish institutional protocols for fairness auditing, including the implementation of $\Delta F1$ monitoring with automated alerts when thresholds are exceeded.

The observation that user interaction patterns can introduce bias at the input stage—with patients providing lower-quality symptom reports when interacting with AI versus human clinicians—suggests that interface design and patient engagement are critical for equitable AI triage . Healthcare organizations should invest in user interface design that explicitly encourages comprehensive symptom reporting, with dynamic prompts that probe for missing information and examples of high-quality reports.

For policymakers, the findings support the establishment of regulatory standards for AI fairness in healthcare, including mandatory fairness audits, continuous monitoring requirements, and performance thresholds similar to the $\Delta F1 \leq 0.05$ governance threshold. The model-dependent

effectiveness of bias mitigation suggests that regulatory frameworks must require device-specific evaluations rather than general model certifications.

5.3 Limitations

1. **Generalizability:** The study draws from a single healthcare system (Beth Israel Deaconess Medical Center) in the United States. Demographic patterns, clinical practices, and healthcare disparities may differ substantially in other settings, limiting the generalizability of findings to international or non-urban contexts.
2. **Demographic Scope:** While the analysis includes gender, race/ethnicity, age, and insurance status, other potentially relevant demographic attributes—including disability status, socioeconomic position beyond insurance, language proficiency, and sexual orientation—were not examined. Intersectional bias (bias at the intersection of multiple demographic attributes) was not systematically analyzed.
3. **Data Limitations:** The MIMIC-IV-ED dataset, while comprehensive, covers the period from 2008-2019 and may not reflect current clinical practices or patient populations. The retrospective nature of the data precludes prospective validation of interventions in real clinical settings. The use of simulated symptom reports for certain analyses, while methodologically rigorous, may not fully capture the complexity of real clinical encounters .
4. **Model Scope:** The study evaluates a limited set of large language models and machine learning baselines. The rapid pace of AI development means that newer models or alternative architectures may exhibit different bias patterns or respond differently to mitigation strategies.
5. **Assumption Stability:** The analysis assumes that historical patterns in triage data remain stable over time. Changes in clinical practice, population demographics, or healthcare systems may alter bias patterns in ways not captured by the current analysis.

5.4 Future Research Directions

1. **Prospective Clinical Validation:** The most critical direction for future research is prospective validation of the fairness governance framework in real-world clinical settings. While the current study demonstrates bias reduction in controlled simulations, real-world implementation may introduce additional complexities related to workflow integration, clinician acceptance, and patient engagement.
2. **Intersectional Bias Analysis:** Future research should examine bias at the intersection of multiple demographic attributes, recognizing that patients may experience compound discrimination based on combinations of gender, race/ethnicity, age, and socioeconomic status. Intersectional fairness requires evaluation methods that can detect bias in small demographic subgroups and interventions tailored to specific intersectional experiences.

3. **International and Cross-Cultural Studies:** Extending bias assessment to international settings with different demographic compositions, triage systems, and cultural contexts is essential for developing globally relevant fairness frameworks. The RemEDy project's planned validation across seven Swiss emergency departments represents one such effort .
4. **Longitudinal Outcomes:** Future research should examine whether bias reduction in triage predictions translates to improved clinical outcomes for underrepresented populations. This requires longitudinal study designs that track patient outcomes from triage through discharge, assessing whether fairer triage predictions lead to more equitable care delivery and health outcomes.

6. Conclusion

This study demonstrates that algorithmic bias in AI-driven tele-triage systems represents a measurable and substantial threat to equitable healthcare delivery, with all evaluated models exhibiting statistically significant demographic disparities exceeding established fairness thresholds. The multi-component fairness governance framework developed in this research—integrating demographic blinding, continuous monitoring with $\Delta F1 \leq 0.05$ governance thresholds, and retrieval-augmented generation corpus rebalancing—effectively reduced bias metrics by 78.4% while improving overall diagnostic accuracy from 84.2% to 84.9%. These findings challenge the conventional assumption of a necessary fairness-accuracy trade-off, suggesting that addressing the sources of bias can enhance both equity and performance.

The study's main contribution is a validated, replicable framework for equitable AI deployment in emergency telemedicine, with specific performance thresholds and governance mechanisms suitable for clinical implementation. For healthcare administrators, the recommendation is to prioritize systems with demonstrated fairness governance capabilities, establish institutional protocols for continuous fairness monitoring, and invest in user interface design that encourages comprehensive patient symptom reporting. As AI continues to transform healthcare delivery, systematic attention to algorithmic fairness is not an optional enhancement but an essential requirement for ensuring that technological innovation serves all patients equitably.

References

1. Shaikh, M., et al. (2026). FairGuard: Blockchain-Enforced Continuous Fairness Governance for Demographically Equitable LLM-Based Emergency Triage Decision Support. *Studies in Health Technology and Informatics*, 338, 403-408.
2. Anonymous. (2026). Reduced symptom reporting quality during human–chatbot versus human–physician interactions. *Nature Health*.
3. UCLA Health. (2025). Researchers develop AI tool to identify undiagnosed Alzheimer's cases while reducing disparities. *UCLA Health News*.
4. Wang, H., Sambamoorthi, N., Hoot, N., Bryant, D., & Sambamoorthi, U. (2025). Evaluating fairness of machine learning prediction of prolonged wait times in Emergency Department with Interpretable eXtreme gradient boosting. *PLOS Digital Health*, 4(3), e0000751.
5. Anonymous. (2026). LLMs for Emergency Triage: Opportunities and Challenges. *arXiv preprint*.
6. Anonymous. (2026). A Review on the Role of Generative AI and Large Language Models in Telemedicine Diagnosis and Clinical Triage. *IEEE Xplore*.
7. Sunny, M. N. M., Sumaiya, U., Akter, M. H., Kabir, F., Munmun, Z. S., Nurani, B., ... & Amin, M. M. (2024). Telemedicine and Remote Healthcare: Bridging the Digital Divide. *South Eastern European Journal of Public Health*, 25, 1500-1510.
8. Nickel, C. H., et al. (2026). Reducing Mis-triage in Emergency Departments (RemEDy): Protocol for Improving Triage Accuracy Through Real-time Evaluation and Artificial Intelligence. *Europe PMC*.
9. Guerra-Adames, A., Avalos-Fernandez, M., Doremus, O., Gil-Jardiné, C., & Lagarde, E. (2024). Uncovering Judgment Biases in Emergency Triage: A Public Health Approach Based on Large Language Models. *Proceedings of Machine Learning Research*, 259, 420-439.
10. Anonymous. (2026). Rethinking fairness in medical imaging: Maximizing group-specific performance with application to skin disease diagnosis. *Medical Image Analysis*, 109, 103950.
11. Wang, H., et al. (2025). Evaluating fairness of machine learning prediction of prolonged wait times in Emergency Department with Interpretable eXtreme gradient boosting. *PMC*, 4(3), e0000751.

12. Anonymous. (2026). EQUITRIAGE: A Fairness Audit of Gender Bias in LLM-Based Emergency Department Triage. *arXiv preprint*.

13. Harvard Medical School. (2025). Researchers Discover Bias in AI Models That Analyze Pathology Samples. *Harvard DBMI News*.

14. Zachariasse, J. M., et al. (2019). Triage systems in emergency departments: A systematic review. *Academic Emergency Medicine*, 26(2), 153-167.

15. Tanabe, P., et al. (2004). The Emergency Severity Index (ESI) 5-level triage system. *Journal of Emergency Nursing*, 30(3), 240-246.