

ORIGINAL RESEARCH

Open access

Explainable Artificial Intelligence Model for Real-Time Monitoring of Algorithmic Bias in Clinical Triage Systems Using Demographic Drift, Outcome Disparities, Prediction Confidence, and Referral Decision Patterns

Sophie Laurent^{1*}, Pierre Dubois¹, Marc Lefevre², Claire Moreau¹

Abstract

Clinical triage algorithms increasingly influence access to emergency care, specialty referral, admission, and follow-up. As patient populations and clinical practice patterns change, these systems can silently drift toward biased performance. Current fairness assessments are often retrospective, episodic, and disconnected from operational triage workflows. They may identify inequity after harm has already accumulated rather than detecting emerging bias as it develops. This article proposes an explainable AI model for continuous monitoring of algorithmic bias in clinical triage systems. The model focuses on demographic drift, outcome disparities, prediction confidence, and referral decision patterns as complementary bias signals. The proposed model uses operational triage logs, demographic distributions, prediction outputs, outcome indicators, and referral decisions to generate a conceptual fairness risk signal. SHAP-based explanation modules decompose the signal into interpretable contributors for clinical governance teams. Conceptually, the model would detect divergence in referral patterns across demographic groups, identify whether the divergence coincides with demographic drift, flag subgroup-specific prediction confidence concerns, and explain the likely drivers of the alert. The output would support timely review rather than automated punitive action. The model could transform algorithmic fairness from a periodic retrospective report into a continuous, transparent, and operationally actionable surveillance system. It is designed as a governance-oriented framework rather than an experimental performance claim.

Keywords Health equity, Explainable artificial intelligence, Algorithmic bias, Clinical triage, Demographic drift, Fairness monitoring

*Correspondence:

Sophie Laurent

sophie.laurent@gmail.com

¹ Department of Digital Health Systems and Informatics, Faculty of Medicine, Sorbonne University, Paris, France

² Department of Intelligent Healthcare Engineering, Faculty of Engineering, École Polytechnique, Paris, France

Introduction

Artificial intelligence is increasingly embedded in clinical triage, risk prediction, and decision support, where it can influence who receives urgent evaluation, specialty referral, hospital admission, or additional diagnostic workup. These systems may inherit inequities from electronic health record data, clinical documentation practices, historical access patterns, and proxy variables that encode structural disadvantage [1, 2]. Concerns about biased algorithms are especially important in triage because decisions made early in the care pathway can shape downstream resource allocation and

clinical outcomes. A fairness-aware monitoring system is therefore needed not only to inspect models before deployment but also to observe how they behave in real clinical operations [3, 4].

One-time fairness audits are insufficient because algorithmic behavior can change after deployment as patient populations, disease prevalence, care pathways, and coding practices evolve. Dataset shift and clinical workflow drift can degrade model validity even when the original model appeared acceptable during development or validation [5, 6]. In health care, this degradation may not be evenly distributed across demographic groups, creating new disparities in calibration, sensitivity, or referral decisions. Continuous monitoring is therefore essential for identifying emerging inequity before it becomes normalized within institutional practice [7, 8].

Explainable artificial intelligence provides a mechanism for moving beyond disparity detection toward disparity interpretation. Methods such as SHAP can attribute model outputs or fairness risk signals to specific features, patient segments, temporal shifts, or operational patterns, making fairness alerts more actionable for clinicians and governance teams [9, 10]. Explainability is also important because health systems must be able to distinguish between a technical failure, a data-quality issue, a workflow change, and a potentially inequitable decision pathway. When used carefully, XAI can help translate bias surveillance from a statistical report into a clinically interpretable governance process [11, 12].

This article proposes an explainable AI model for real-time monitoring of algorithmic bias in clinical triage systems using four complementary signal streams: demographic drift, outcome disparities, prediction confidence, and referral decision patterns. The model is conceptual and governance-oriented, designed to monitor deployed triage algorithms rather than replace clinical judgment or claim experimental performance. It is motivated by evidence that health care algorithms can reproduce racial and ethnic disparities, that clinical AI can degrade over time, and that post-deployment accountability requires structured oversight. By combining fairness metrics with SHAP-based explanations, the model would be expected to support earlier detection, clearer interpretation, and more consistent response to emerging bias.

Background

Algorithmic bias in clinical triage systems

Algorithmic bias in clinical triage can arise when prediction models learn from historical patterns of unequal access, incomplete measurement, differential documentation, or biased clinical decisions. The well-known example of a population health management algorithm that underestimated the needs of Black patients illustrates how seemingly neutral proxy outcomes can embed racial inequity [1]. Similar risks extend to emergency triage, symptom assessment, deterioration prediction, and admission support because these tools often rely on routinely collected data that reflect prior clinical and social processes [2, 13]. In triage settings, biased outputs may influence urgency assignment, referral routing, diagnostic escalation, or admission decisions, making fairness monitoring a direct patient-safety concern [14, 15].

Demographic drift and concept drift

Demographic drift occurs when the population served by a triage system changes relative to the population represented during model development, while concept drift occurs when relationships among predictors, outcomes, and clinical decisions evolve over time. In clinical AI, both forms of drift can produce subgroup-specific degradation because changes in age mix, insurance coverage, race and ethnicity composition, language needs, disease prevalence, or care pathways may affect groups differently [5, 6]. A model that appeared equitable during validation could therefore become unfair if the institution begins serving a different patient population or if practice patterns change. Continuous drift monitoring is a necessary complement to pre-deployment validation because fairness is not a static property of a model [7, 8].

Outcome disparities and prediction confidence as bias signals

Outcome disparities provide direct evidence that a triage model may be functioning differently across patient groups, particularly when error patterns, calibration, or downstream clinical outcomes diverge in ways that cannot be clinically justified. Prediction confidence can add an additional fairness signal because a model may be overly certain for some groups and uncertain for others, reflecting uneven data representation or subgroup-specific miscalibration [16, 17]. In clinical imaging and risk prediction, studies have shown that models may encode demographic information or perform differently across under-served populations, reinforcing the need to examine both predictions and confidence patterns by group [18, 19]. A monitoring model should therefore evaluate not only whether decisions differ, but also whether the model expresses confidence in ways that could amplify inequitable triage decisions [20, 21].

Referral patterns as manifestation of triage bias

Referral and admission decisions are important manifestations of triage bias because they translate algorithmic scores into access to services, specialty evaluation, and clinical escalation. Even when a model does not directly determine a referral, its risk estimate, acuity recommendation, or decision-support prompt can influence clinician behavior and shape downstream care pathways [4, 14]. Differences in referral patterns by race, ethnicity, insurance, language, age, or disability status may reflect clinical need, but they may also reveal inequitable decision pathways when similar patients receive different levels of escalation. A bias-monitoring system should therefore treat referral decisions as operational outcomes that require fairness analysis alongside prediction errors and calibration signals [22, 23].

Explainable AI and fairness monitoring tools

Explainable AI methods can help fairness monitoring systems identify why a disparity signal has emerged and which features, subgroups, or workflow changes are contributing to it. SHAP-based explanation frameworks have been used to connect local predictions with global model behavior, making them suitable for translating complex bias signals into interpretable summaries for clinical stakeholders [9, 10]. Broader responsible AI frameworks emphasize that transparency, monitoring, documentation, and governance are required for safe clinical deployment rather than optional additions after model development [7, 11]. However, many existing approaches remain retrospective, motivating a real-time model that connects fairness metrics, drift detection, confidence analysis, referral monitoring, and explanation in a single operational architecture [24, 25].

Model Development Overview

High-level monitoring pipeline

The proposed monitoring pipeline would receive streaming inputs from triage predictions, operational logs, demographic records, clinical outcomes, and referral decisions. It would compute fairness-relevant summaries over recent operational periods, compare them with reference behavior, and pass the resulting signals to an explainable bias-risk model. This architecture separates the monitored triage algorithm from the monitoring layer, allowing the framework to observe any deployed triage system without requiring direct access to its internal training process [6, 7]. Alerts would be routed to a governance dashboard where fairness metrics, drift indicators, confidence patterns, referral distributions, and SHAP explanations are presented together for human review [9, 22].

Figure 1 illustrates the proposed explainable AI architecture for converting real-time triage data streams into fairness risk alerts, SHAP-based explanations, governance review, and documented mitigation actions.

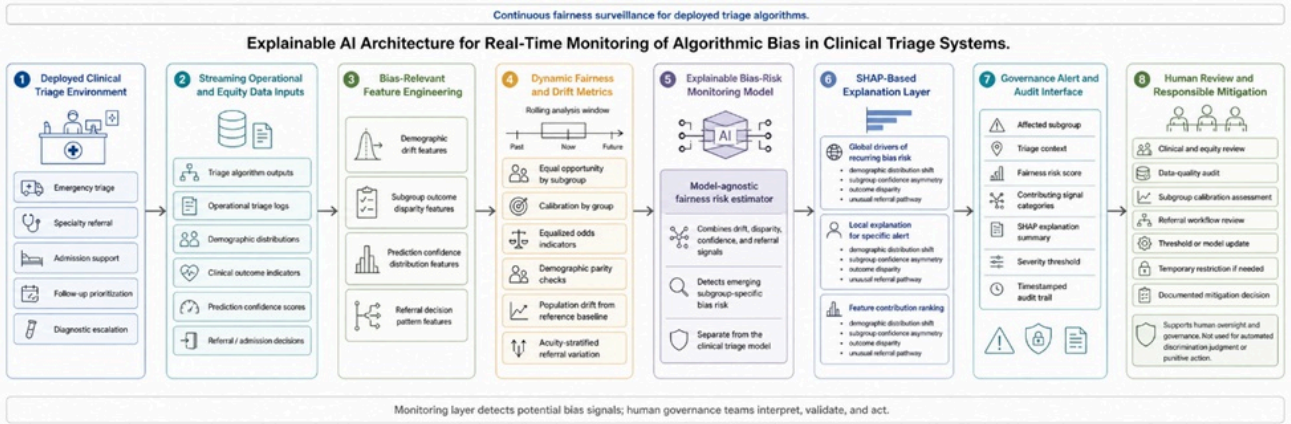


Figure 1. Explainable AI Architecture for Real-Time Monitoring of Algorithmic Bias in Clinical Triage Systems

Core input features

The core input features would represent four categories of bias evidence: demographic shift metrics, subgroup outcome disparities, prediction confidence distributions, and referral decision counts. Demographic shift features would compare current patient distributions with reference distributions, while outcome features would summarize whether triage-related outcomes and errors appear uneven across demographic groups [5, 8]. Confidence features would characterize whether the model is systematically more or less certain for particular groups, and referral features would track whether similar acuity levels lead to different escalation pathways across groups. This multi-signal design reflects evidence that bias can emerge through data representation, model behavior, clinical workflow, and institutional response patterns rather than through a single observable metric [1, 2, 20].

Table 1 presents the proposed multi-signal architecture for converting demographic drift, outcome disparities, prediction confidence, and referral decision patterns into interpretable bias-monitoring features for clinical triage governance.

Table 1. Multi-Signal Architecture for Real-Time Bias Monitoring in Clinical Triage

Bias-monitoring signal domain	Primary operational data source	Feature engineering logic	Fairness or drift question addressed	Example monitoring indicator	Interpretive value for governance
Demographic drift	Current triage population compared with a reference deployment population	Compare rolling distributions of race, ethnicity, age, sex, language, insurance status, disability status, or other locally approved equity variables against baseline distributions	Has the population served by the triage algorithm changed in ways that may affect subgroup reliability or fairness?	Increasing divergence between current and reference subgroup distributions	Helps determine whether emerging bias may be related to population shift rather than only model malfunction
Outcome disparity	Triage outcomes, clinical endpoints, return visits,	Stratify outcomes and errors by subgroup, acuity	Are clinical outcomes or prediction errors	Widening subgroup gap in false	Identifies whether model-assisted triage may be

	escalation events, admission, intensive care transfer, or follow-up completion	level, and time window	becoming uneven across demographic groups?	negatives, calibration, sensitivity, or adverse outcome rates	associated with unequal downstream clinical consequences
Prediction confidence asymmetry	Model confidence scores, probability outputs, uncertainty scores, risk categories, or acuity recommendations	Compare confidence distributions across demographic groups and triage contexts	Is the model systematically more confident or less confident for particular groups?	Higher confidence in low-risk predictions for one subgroup despite similar acuity or outcome burden	Supports detection of subgroup-specific miscalibration or uneven representation in training data
Referral decision pattern	Referral orders, specialty consult requests, admission decisions, diagnostic escalation, discharge decisions, and follow-up recommendations	Aggregate decisions by subgroup, acuity category, clinical context, and model output level	Are similar patients receiving different access pathways after triage?	Lower specialty referral intensity for one subgroup at comparable triage acuity	Links algorithmic outputs to operational access decisions rather than limiting fairness analysis to prediction metrics
Acuity-stratified comparison	Triage acuity category, chief complaint, risk score, clinical severity indicators, and referral decision	Compare decision patterns within clinically comparable strata	Do observed subgroup differences persist after accounting for triage severity?	Referral gap among patients assigned the same acuity level	Reduces overinterpretation of raw demographic differences and supports clinically contextualized fairness review
Temporal fairness trend	Rolling windows of model outputs, outcomes, and decisions	Track fairness indicators over time rather than as a single retrospective snapshot	Is a disparity newly emerging, persistent, worsening, or resolving?	Progressive increase in calibration gap over sequential monitoring windows	Helps governance teams distinguish transient variation from sustained equity risk
Composite fairness risk signal	Integrated drift, outcome, confidence, and referral features	Combine multiple bias-relevant indicators into a monitored risk score	Do several weak signals together suggest a meaningful fairness concern?	Elevated fairness risk score driven by drift plus referral divergence	Prioritizes alerts for governance review while avoiding reliance on any single fairness metric
Explainability-ready signal structure	Engineered features prepared for SHAP attribution	Maintain interpretable feature categories aligned with clinical	Can the system explain why a bias alert was generated?	SHAP attribution identifying confidence asymmetry	Converts fairness surveillance from opaque detection into actionable

		governance questions		and referral divergence as top contributors	governance intelligence
--	--	----------------------	--	---	-------------------------

Design principles

The model is designed to be real-time, model-agnostic, explainable, clinically governed, and equity-oriented. Real-time monitoring supports earlier recognition of emerging bias, while model-agnosticism allows the same oversight framework to monitor different triage algorithms, including proprietary or externally supplied tools [7, 14]. Explainability ensures that alerts are not merely statistical warnings but interpretable signals that can be discussed by clinicians, health equity officers, informatics teams, and compliance leaders [9, 11]. Alignment with clinical governance cycles is essential because fairness monitoring should lead to review, recalibration, workflow modification, or policy action rather than passive reporting [22, 23].

Data Sources and Feature Engineering for Bias Monitoring

Demographic and outcome data streams

The model would draw demographic data, triage outputs, clinical outcomes, and operational decisions from electronic health records, triage logs, decision-support systems, and referral workflows. Demographic variables may include race, ethnicity, age, sex, language, insurance type, and other locally governed equity-relevant fields, while outcome streams may include admission, intensive care escalation, return visits, mortality, or other triage-relevant endpoints. Because electronic health record data can reflect missingness, measurement bias, documentation inequity, and structural access differences, feature engineering must treat these streams as imperfect indicators rather than neutral facts [2, 13]. Governance review is therefore required to define which variables are appropriate for monitoring and how they should be interpreted in context [3, 21].

Drift and confidence features

Drift features would quantify how current demographic and clinical distributions differ from a reference period, allowing the monitoring layer to detect whether the served population has changed in ways that may affect fairness. Confidence features would summarize the distribution of prediction certainty across demographic groups and triage categories, enabling the system to identify potential over-confidence or under-confidence in specific subgroups. These signals are important because clinical AI degradation can occur gradually and may be missed when only aggregate performance is monitored [6, 8]. By combining drift and confidence features, the model could detect situations where a population shift coincides with a change in model certainty or fairness behavior [5, 24].

Referral decision pattern encoding

Referral decision pattern features would aggregate operational decisions such as discharge, admission, specialty consultation, diagnostic escalation, and follow-up recommendation by demographic group and triage acuity level. Encoding these decisions by acuity context is important because raw referral differences may reflect legitimate clinical variation, while persistent differences among otherwise similar triage categories may suggest inequitable workflow effects. The monitoring model would not assume that every difference is biased, but it would flag patterns that warrant clinical and equity review [22, 23]. This approach recognizes that algorithmic bias in triage may appear not only in predicted risk but also in how predictions shape access to subsequent care [4, 14].

Explainable Model Architecture for Bias Detection

Learning to predict disparities

The proposed architecture uses a supervised explainable monitoring model to estimate an aggregated fairness risk signal from demographic drift, outcome disparity, prediction confidence, and referral decision features. This monitoring model is distinct from the triage model itself; it learns patterns that would be expected to indicate possible bias in operational behavior rather than predicting patient acuity directly. Gradient-boosted models are conceptually suitable because they can capture nonlinear interactions among drift, confidence, outcome, and referral signals while remaining compatible with SHAP-based explanation [10, 12]. The goal is not to automate judgments of discrimination, but to prioritize cases where governance teams should investigate whether a deployed triage system is behaving inequitably [11, 25].

SHAP attribution of disparity drivers

SHAP attribution would decompose each fairness risk alert into feature-level contributors, such as a demographic distribution shift, a subgroup-specific confidence pattern, an outcome disparity signal, or an unusual referral pathway. Global explanations would show recurring drivers of bias risk across time, while local explanations would describe why a specific alert was generated for a specific subgroup or triage context [9, 10]. This interpretive layer is essential because a fairness alert without an explanation may be difficult for clinicians and governance teams to trust or act upon. By connecting disparity signals to operational features, SHAP explanations could help distinguish model drift, data-quality problems, referral workflow changes, and subgroup-specific miscalibration [7, 24].

Threshold-based alerting and escalation rules

The alerting layer would compare the explainable fairness risk signal with locally defined governance thresholds, generating alerts when bias indicators warrant review. Each alert would include the affected subgroup, the relevant triage context, the contributing signal categories, and SHAP-based explanations so that reviewers can understand why the system elevated the concern. Thresholds should be adjustable because acceptable sensitivity to bias signals may vary by clinical setting, patient population, model role, and institutional equity priorities [3, 22]. Escalation rules would route alerts to appropriate stakeholders, supporting structured responses such as data review, calibration assessment, workflow audit, or temporary restriction of model use [11, 23].

Dynamic Fairness Metrics and Bias Signal Generation

Computing sliding-window fairness metrics

The monitoring model would compute fairness metrics over rolling operational windows so that disparities can be examined as evolving patterns rather than isolated retrospective summaries. Equal opportunity, demographic parity, equalized odds, and calibration-by-group would be treated as complementary indicators because each captures a different aspect of fairness in triage decision-making [3, 20]. Sliding-window analysis would allow governance teams to identify whether disparities are newly emerging, persistent, or worsening after a workflow or population change. This design is especially important in clinical environments where aggregate model performance may remain stable while subgroup-specific performance becomes inequitable [5, 6].

Detecting demographic drift and its impact

Demographic drift detection would compare current patient distributions with a reference baseline and then examine whether observed shifts coincide with changes in subgroup performance, confidence, or referral behavior. The model would be expected to flag situations in which population changes are not merely statistical variation but appear connected to widening fairness gaps [7, 8]. Such an approach recognizes that drift is clinically meaningful when it affects model reliability, patient access, or equity across groups. Recent work on harmful data shifts in clinical AI supports the need to connect drift detection with responsible post-deployment monitoring and remediation rather than treating drift as a purely technical phenomenon [24, 26].

Analyzing referral decision patterns for disparities

Referral decision analysis would examine whether admission, discharge, specialty consultation, or diagnostic escalation patterns differ across demographic groups within comparable triage acuity contexts. The model would use clinical covariates and triage-level stratification to help determine whether a referral pattern appears explainable by clinical factors or remains an equity concern requiring review [22, 23]. This analysis would not establish discrimination by itself, but it could detect operational patterns that are consistent with biased triage pathways. In this framework, referral signals complement outcome and confidence metrics because they show how model-assisted triage may influence access to downstream care [4, 14].

Explainability and Actionability for Stakeholders

Global explanations for equity officers and governance

Global explanations would summarize the most frequent contributors to fairness risk across the institution, allowing health equity officers and algorithm governance committees to see whether concerns are driven by demographic drift, confidence asymmetry, outcome disparities, or referral patterns. SHAP summary views could show how specific features repeatedly contribute to bias risk, helping reviewers prioritize systemic interventions rather than reacting to isolated alerts [9, 10]. Such explanations would be expected to support institutional learning by revealing whether a triage algorithm is vulnerable in particular patient groups, time periods, or operational contexts. This governance-facing layer aligns with calls for accountable clinical AI oversight that connects technical monitoring with organizational responsibility [11, 22].

Local explanations for a specific alert

Local explanations would provide a focused account of why a particular alert was generated, including the subgroup affected, the triage context, the signal categories involved, and the features contributing most strongly to the fairness risk score. For example, an alert might indicate that a subgroup with similar acuity is receiving lower specialty referral intensity while the triage model expresses higher confidence in low-risk predictions for that group, prompting review of both model calibration and clinical workflow [16, 19]. The purpose of the explanation is not to assign blame but to help clinicians and governance teams understand where to investigate. Local explanations are especially valuable in high-stakes triage settings because they translate abstract fairness metrics into operational patterns that can be reviewed and corrected [12, 27].

Counterfactual explanations for mitigation guidance

Counterfactual explanations would support mitigation planning by estimating how the fairness risk signal might change under alternative operational or modeling assumptions, such as subgroup recalibration, revised referral thresholds, or improved documentation completeness. These counterfactuals should be framed conceptually rather than as guaranteed causal effects because observed disparities may reflect multiple interacting clinical and social mechanisms [21, 25]. Their value lies in helping stakeholders compare plausible response options and identify which intervention would be expected to reduce a monitored disparity. When combined with SHAP attribution, counterfactual reasoning could make fairness alerts more actionable by connecting detected bias signals with candidate mitigation pathways [9, 10].

Transparent audit trail for regulatory review

The monitoring system would maintain a timestamped audit trail of fairness metrics, drift indicators, confidence summaries, referral patterns, alerts, explanations, reviewer actions, and mitigation decisions. Such documentation would support retrospective accountability by showing not only whether bias signals occurred, but also how the institution interpreted and responded to them [7, 11]. Transparent audit trails are important because clinical AI governance requires evidence that monitoring is continuous, interpretable, and connected to action rather than limited to pre-deployment

validation. Model documentation principles such as structured reporting, accountability records, and operational oversight provide a foundation for this audit-oriented approach [22, 23].

Integration Into Clinical Governance and Workflow

Embedding in health equity governance committees

The proposed monitoring model would feed real-time alerts and periodic summary reports into an institutional committee responsible for health equity, clinical AI oversight, and patient safety. This committee would review recurring bias signals, examine SHAP explanations, request technical audits, and coordinate operational responses when disparities appear clinically meaningful [3, 22]. Embedding the model within governance structures is essential because fairness monitoring cannot be reduced to a technical dashboard; it must be tied to institutional authority and accountability. Such integration reflects broader recommendations that responsible clinical AI requires multidisciplinary oversight, clear escalation pathways, and continuous evaluation after deployment [7, 11].

Response protocols for bias alerts

Response protocols would define how the organization acts when a bias alert meets predefined severity criteria, including review of input data, subgroup calibration, referral workflows, documentation practices, and model update needs. Depending on the alert, governance teams could request recalibration, restrict model use in a specific context, adjust triage thresholds, retrain the model, or revise clinical workflow guidance [6, 24]. These actions should be reviewed by clinicians and equity leaders because the monitoring model identifies potential bias signals rather than proving causal discrimination. Predetermined protocols would reduce alert fatigue and help ensure that fairness concerns lead to timely, proportionate, and documented responses [22, 25].

Table 2 outlines a governance response framework linking explainable bias-alert patterns to review actions, mitigation options, and documentation requirements for responsible clinical triage oversight.

Table 2. Governance Response Framework for Explainable Bias Alerts in Clinical Triage Systems

Alert interpretation category	Typical signal pattern	SHAP explanation pattern	Primary governance question	Recommended review action	Possible mitigation pathway	Documentation requirement
Population drift with stable performance	Demographic distribution has shifted, but subgroup calibration, outcomes, and referral patterns remain stable	SHAP attribution dominated by demographic drift features, with weak contribution from outcome or referral disparity	Does the new population profile require closer surveillance even without current evidence of harm?	Continue intensified monitoring and review subgroup sample adequacy	Update monitoring baseline or expand subgroup-specific validation	Record drift event, review date, monitoring decision, and rationale for no immediate intervention
Drift-associated fairness degradation	Demographic shift coincides with worsening	SHAP attribution shows joint contribution	Is the model becoming unreliable for a changing	Conduct subgroup calibration assessment and	Recalibrate model, revise thresholds, retrain with updated	Document drift evidence, affected subgroup,

	subgroup calibration, sensitivity, confidence, or referral access	from drift features and subgroup disparity features	patient population?	compare current performance with reference period	data, or restrict use in affected context	technical review findings, and chosen mitigation
Confidence asymmetry without clear outcome harm	One subgroup receives systematically higher or lower confidence despite comparable acuity, but outcome disparity is not yet established	SHAP attribution highlights prediction confidence distribution features	Could uneven confidence amplify future inequity or inappropriate clinician reliance?	Review calibration curves, uncertainty behavior, subgroup representation, and clinician use of confidence outputs	Modify confidence display, add uncertainty warning, recalibrate subgroup confidence, or adjust user guidance	Record confidence concern, uncertainty interpretation, and decision about display or threshold changes
Referral divergence at comparable acuity	Referral, admission, specialty consult, or diagnostic escalation rates differ across groups within similar triage levels	SHAP attribution highlights referral decision pattern features and acuity-stratified comparison	Are model-assisted triage decisions associated with unequal access pathways?	Audit referral workflows, review clinician override patterns, and examine local policy or capacity constraints	Revise referral protocols, adjust triage decision-support prompts, provide equity-focused workflow guidance	Document operational pathway reviewed, clinical justification assessment, and workflow response
Outcome disparity with unclear mechanism	Subgroup outcomes or error rates differ, but drift, confidence, and referral explanations are mixed or weak	SHAP attribution is distributed across several weak contributors	Is there an unmeasured clinical, documentation, or access factor driving the disparity?	Convene multidisciplinary review including clinicians, equity officers, informatics, and data-quality specialists	Improve data capture, revise feature definitions, conduct deeper subgroup analysis, or request external audit	Record uncertainty, additional analyses requested, and limits of causal interpretation
Recurring alert for same subgroup or pathway	Similar alerts recur across multiple rolling windows or operational settings	Global SHAP summaries show repeated contribution from the same subgroup, feature	Does the pattern represent a systemic equity risk rather than an isolated alert?	Escalate to institutional AI governance committee and patient-safety leadership	Formal mitigation plan, model update, workflow redesign, or temporary model-use limitation	Maintain longitudinal audit trail linking repeated alerts, review actions, and mitigation outcomes

		class, or referral pathway				
High-severity composite fairness alert	Multiple signals align: drift, outcome disparity, confidence asymmetry, and referral divergence	SHAP attribution shows strong contributions from several signal domains	Is continued routine use of the triage algorithm appropriate in the affected context?	Initiate urgent governance review and consider temporary restrictions while investigation proceeds	Pause model use for specific subgroup/context, require human-only review, recalibrate, retrain, or revise deployment policy	Record severity basis, interim safeguards, responsible decision-makers, and follow-up timeline
Resolved or improving alert pattern	Previously elevated fairness risk declines after mitigation or workflow adjustment	SHAP attribution shows reduced contribution from prior dominant drivers	Did the mitigation plausibly reduce the monitored fairness concern?	Review pre/post monitoring trends and assess whether surveillance intensity can return to routine level	Maintain mitigation, update governance threshold, or continue targeted monitoring	Document mitigation assessment, residual risk, and decision to close or continue monitoring

Evaluation Strategy

Detection sensitivity and specificity

Evaluation of the monitoring framework should focus on whether it could detect known or plausibly simulated bias events while avoiding excessive false alarms that would undermine clinical trust. Historical episodes of workflow change, demographic shift, or subgroup performance degradation could be used conceptually to assess whether the monitoring system would have generated timely alerts [5, 8]. Because this article proposes a model architecture rather than reporting experiments, such evaluation should be described as a future validation requirement rather than as evidence of demonstrated performance. Studies of temporal AI degradation and clinical model monitoring show why evaluation must include post-deployment behavior, not only development-stage validation [7, 26].

Explanation quality and stakeholder trust

Explanation quality should be evaluated through structured review by clinicians, health equity officers, informaticians, and patient-safety leaders who can assess whether alerts are understandable, relevant, and actionable. The review should examine whether SHAP summaries and local explanations help stakeholders identify plausible contributors to disparity signals without overstating causal certainty [9, 10]. Trustworthiness would depend not only on technical explanation fidelity but also on whether the system communicates uncertainty, distinguishes statistical association from bias adjudication, and supports human oversight. Prior work on explainable clinical AI emphasizes that explanations must be evaluated in relation to user needs, clinical context, and governance decisions rather than as purely mathematical outputs [27, 28].

Impact on organizational fairness

The organizational impact of the monitoring model should be evaluated by examining whether it improves the health system's ability to identify, review, and respond to potential algorithmic bias in triage workflows. Relevant evaluation domains would include timeliness of review, quality of governance response, evidence of recalibration or workflow

adjustment, and the institution's ability to document fairness-related decisions [11, 22]. Any comparison of fairness before and after deployment should avoid simplistic causal claims unless supported by appropriate prospective evaluation. The model would be expected to contribute most meaningfully when paired with accountable governance, clear response protocols, and a commitment to equity-centered clinical AI implementation [23, 29].

Limitations

Data limitations and causality

The proposed monitoring system would detect signals of potential bias, but it would not by itself prove discrimination, causality, or inappropriate clinical decision-making. Observed disparities may reflect unmeasured clinical complexity, missing data, documentation differences, access barriers, or historical inequities embedded in the health system [2, 21]. Human adjudication is therefore essential to interpret alerts, determine whether they represent bias, and decide what mitigation is appropriate. This limitation is especially important because algorithmic fairness metrics can identify inequitable patterns but cannot replace ethical, clinical, and organizational judgment [3, 25].

Privacy and stigmatization risk

Continuous monitoring of demographic outcomes requires careful privacy protection, role-based access, transparent governance, and safeguards against stigmatizing patients or clinicians. Demographic attributes are necessary for equity monitoring, but they must be handled in ways that prevent misuse, inappropriate profiling, or punitive interpretation of subgroup patterns [13, 22]. The system should therefore emphasize aggregate surveillance, accountable review, and documented mitigation rather than individual-level labeling. Responsible implementation must balance the need to detect inequity with the obligation to protect patients, support clinicians, and maintain trust in clinical AI governance [11, 21].

Conclusion

The proposed explainable AI model reframes algorithmic bias monitoring in clinical triage as a continuous surveillance and governance function. By integrating demographic drift, outcome disparities, prediction confidence, and referral decision patterns, the model offers a structured way to detect potential inequity as clinical conditions change.

Its key strength is the combination of multi-signal bias detection with explainable attribution. Rather than producing an opaque warning, the system would show which features, subgroups, and operational patterns contributed to a fairness alert, supporting more targeted institutional review.

Important challenges remain, including distinguishing correlation from causation, protecting patient privacy, preventing alert fatigue, and ensuring that monitoring results lead to responsible human oversight. The model should therefore be understood as an accountability tool that supports governance rather than as an automated judge of discrimination.

Pilot implementation within health systems committed to equity would help clarify how such monitoring can be integrated into clinical operations, governance committees, and regulatory expectations. As clinical AI becomes more common in triage, continuous and explainable fairness monitoring should become a core standard for safe and equitable deployment.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 04 Aug 2025

Revised: 08 Sep 2025

Accepted: 12 Oct 2025

Published online: 25 February 2026

Rights and permissions

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

References

- Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447–53.
- Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*. 2018;178(11):1544–7.
- Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866–72.
- Char DS, Shah NH, Magnus D. Implementing machine learning in health care: Addressing ethical challenges. *N Engl J Med*. 2018;378(11):981–3.
- Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385(3):283–6.
- Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: Implications and potential remedies. *Br J Radiol*. 2023;96(1150):20220878.
- Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: Towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit Med*. 2022;5(1):66.
- Vela D, Sharp A, Zhang R, Nguyen T, Hoang A, Pianykh OS. Temporal quality degradation in AI models. *Sci Rep*. 2022;12(1):11654.
- Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655.

- Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell.* 2020;2(1):56–67.
- Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. *Nat Med.* 2019;25(9):1337–40.
- Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat Commun.* 2020;11(1):3852.
- Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics.* 2019;21(2):E167–E179.
- Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17(1):195.
- Panch T, Mattie H, Celi LA. The “inconvenient truth” about AI in healthcare. *NPJ Digit Med.* 2019;2(1):77.
- Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med.* 2021;27(12):2176–82.
- Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci U S A.* 2020;117(23):12592–4.
- Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen LC, et al. AI recognition of patient race in medical imaging: A modelling study. *Lancet Digit Health.* 2022;4(6):e406–e414.
- Khor S, Haupt EC, Hahn EE, Lyons LJ, Shankaran V, Bansal A. Racial and ethnic bias in risk prediction models for colorectal cancer recurrence when race and ethnicity are omitted as predictors. *JAMA Netw Open.* 2023;6(6):e2318495.
- Huang J, Galal G, Etemadi M, Vaidyanathan M. Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review. *JMIR Med Inform.* 2022;10(5):e36388.
- Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. *Patterns.* 2021;2(10):100347.
- Chin MH, Afsar-Manesh N, Bierman AS, Chang C, Colón-Rodríguez CJ, Dullabh P, et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. *JAMA Netw Open.* 2023;6(12):e2345050.
- Jain A, Brooks JR, Alford CC, Chang CS, Mueller NM, Umscheid CA, et al. Awareness of racial and ethnic bias and potential solutions to address bias with use of health care algorithms. *JAMA Health Forum.* 2023;4(6):e231197.
- Subasri V, Krishnan A, Kore A, Dhalla A, Pandya D, Wang B, et al. Detecting and remediating harmful data shifts for the responsible deployment of clinical AI models. *JAMA Netw Open.* 2025;8(6):e2513685.
- Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. *Commun Med (Lond).* 2021;1(1):25.
- Kore A, Abbasi Babil E, Subasri V, Abdalla M, Fine B, Dolatabadi E, et al. Empirical data drift detection experiments on real-world medical imaging data. *Nat Commun.* 2024;15(1):1887.
- Boultsakis-Logothetis S, Green D, Holland M, Al Moubayed N. Predicting acute clinical deterioration with interpretable machine learning to support emergency care decision making. *Sci Rep.* 2023;13(1):13563.
- Choi A, Choi SY, Chung K, Chung HS, Song T, Choi B, et al. Development of a machine learning-based clinical decision support system to predict clinical deterioration in patients visiting the emergency department. *Sci Rep.* 2023;13(1):8561.

Cheng CY, Hsu TH, Hung YL, Hsu TY, Cheng FJ, Pan HY, et al. Early prediction of in-hospital deterioration after emergency department admission using machine learning models. *BMC Emerg Med.* 2026;26:1–12.