

ORIGINAL RESEARCH

Open access

Explainable Artificial Intelligence Model for Detecting Inequitable Specialty Referral Patterns Using Patient Demographics, Insurance Type, Diagnosis Severity, Primary Care Documentation, and Provider Network Structure

George Papadopoulos^{1*}, Eleni Georgiou¹

Abstract

Specialty referral pathways are a critical point at which healthcare inequities can emerge. Referral decisions may be shaped by insurance status, race, language, documentation practices, diagnosis severity, and the structure of available provider networks. Health systems often lack scalable and explainable tools for detecting inequitable referral patterns as they occur. As a result, discriminatory or structurally biased patterns may remain hidden within routine clinical operations. This article develops a conceptual explainable artificial intelligence model for identifying whether demographic or insurance factors unduly influence specialty referral decisions after accounting for clinical severity. The model is designed to support transparent, fairness-oriented referral analytics rather than replace clinical judgment. The proposed model uses a gradient-boosted classification framework trained on referral-eligible primary care encounters. Input features include patient demographics, insurance type, diagnosis severity, primary care note-derived complexity and completeness features, and provider network metrics, with SHAP-based explanation layers used for fairness auditing. Conceptually, the model could flag encounters in which predicted referral likelihood diverges from clinically expected patterns. These flags would be interpreted through explanation methods that attribute potential inequitable influence to insurance, demographic, documentation, or network-related factors. The model could help health systems audit, explain, and intervene on systemic specialty referral bias. Its central contribution is a transparent framework for moving referral equity work from retrospective description toward proactive, data-driven fairness review.

Keywords Health equity, Explainable artificial intelligence, SHAP, Specialty referral, Algorithmic fairness, Provider networks

*Correspondence:

George Papadopoulos

george.papadopoulos@gmail.com

¹ Department of Clinical Informatics and Smart Health Systems, Faculty of Medicine, National and Kapodistrian University of Athens, Athens, Greece

Introduction

Specialty referral decisions are consequential because they determine whether patients gain access to diagnostic evaluation, procedural care, longitudinal subspecialty management, and disease-specific treatment pathways. Evidence on primary care referral behavior suggests that referral patterns can differ by race and insurance status, indicating that inequities may occur before patients ever reach specialist care [1, 2]. These referral gaps are not simply administrative

inconveniences; they can shape downstream diagnosis, treatment timeliness, and continuity of care, especially for patients who already face barriers to navigating complex health systems [3, 4].

Most health systems still rely on periodic equity reports, manual chart review, or complaint-driven investigation to identify referral disparities. These methods may reveal broad differences after the fact, but they are poorly suited for detecting emerging patterns across clinics, payers, diagnoses, and provider networks in near real time [5, 6]. Referral management systems and electronic referral tools can improve process visibility, yet without fairness-aware analytics they may not distinguish ordinary operational friction from inequitable access patterns [7, 8].

Interpretable machine learning offers a way to move beyond descriptive disparity measurement toward patient-level and system-level equity auditing. Fairness-oriented models can be designed to evaluate whether referral likelihood is shaped primarily by clinical need or instead by non-clinical attributes such as insurance type, language, race, or network access [9, 10]. Explainable artificial intelligence is especially relevant in this setting because referral recommendations must be clinically intelligible, auditable, and aligned with equity goals rather than treated as opaque predictions [11, 12].

The thesis of this article is that an explainable artificial intelligence model could estimate referral likelihood for referral-eligible primary care encounters and decompose each prediction into clinical, documentation, demographic, insurance, and network contributions. Such a model would not declare discrimination from prediction alone, but it could identify patterns requiring review when non-clinical factors appear to influence referral probability after severity adjustment. By combining fairness metrics, SHAP explanations, and counterfactual reasoning, the system would support structured referral equity audits while preserving the role of clinical judgment.

Background

Referral disparities and their drivers

Referral disparities may arise from clinician decision-making, patient communication barriers, insurance acceptance, administrative complexity, specialist availability, and structural constraints within referral networks. Studies of primary care and specialty access show that race, insurance type, and care coordination processes can be associated with differential referral scheduling or completion, suggesting that inequity can occur across multiple points in the referral pathway [1, 3]. Insurance-related access barriers are especially relevant for publicly insured or uninsured patients because even when a referral is clinically indicated, specialist availability and acceptance may constrain whether referral pathways are realistically accessible [2, 4].

Diagnosis severity and clinical confounders

Any model designed to detect inequitable referral patterns must distinguish legitimate clinical variation from bias. Diagnosis severity, comorbidity burden, symptom complexity, and disease trajectory may appropriately influence whether a referral is needed, so an unadjusted comparison across demographic groups could either exaggerate or obscure inequity [9, 13]. Fairness-aware healthcare modeling therefore requires careful attention to clinical confounding, because variables used to define “need” may themselves reflect prior inequities in access, documentation, diagnosis, and treatment [10, 14].

Primary care documentation and bias

Primary care documentation can shape referral decisions by communicating symptom burden, diagnostic uncertainty, prior treatment attempts, patient preferences, and urgency to specialists. However, documentation may also encode bias when notes contain stigmatizing descriptors, uneven detail, or language that frames patients differently by race, gender, or social position [15, 16]. If referral appropriateness is inferred from notes, documentation bias could become a mediator through which inequitable clinical attention is converted into lower apparent referral need [17].

Provider network structure and access

Provider network structure can convert clinical need into unequal access when specialists are geographically concentrated, selectively accept certain insurance types, or operate within dense referral relationships that exclude particular clinics or patient populations. Referral access is therefore not only an individual clinician decision but also a network phenomenon shaped by specialist supply, payer participation, distance, scheduling capacity, and institutional affiliations [1, 2]. Enhanced referral management systems may improve coordination, but an equity model must explicitly represent network constraints so that structural access barriers are not mistaken for patient-level clinical differences [6, 7].

Explainable AI and algorithmic fairness in healthcare operations

Explainable artificial intelligence methods such as SHAP, partial dependence, local explanations, and counterfactual analysis can help translate prediction models into auditable evidence about which factors influence a referral probability. In healthcare, explainability must be paired with fairness principles because transparent models can still reinforce inequity if they learn biased labels, rely on proxy variables, or optimize operational efficiency without equity safeguards [18, 19]. Fairness metrics such as equal opportunity, demographic parity, and calibration by group can be useful, but they require contextual interpretation because healthcare decisions often involve legitimate clinical heterogeneity and historically biased data-generating processes [14, 20].

Model Development Overview

High-level equity audit pipeline

For each primary care encounter with potential specialty referral need, the proposed pipeline would estimate the probability that a referral should occur under observed clinical, documentation, insurance, demographic, and network conditions. SHAP explanations would then decompose the prediction into contributions from clinical features, such as diagnosis severity and comorbidity, and non-clinical or structural features, such as insurance type, race, language, and specialist network access [11, 21]. The model would flag encounters or groups of encounters when non-clinical factors appear to contribute disproportionately to lower referral likelihood after accounting for clinical need, making the output an equity audit signal rather than an automatic referral mandate [10, 20].

Figure 1 illustrates the explainable artificial intelligence pipeline for detecting inequitable specialty referral patterns through severity-adjusted prediction, SHAP-based decomposition, and fairness-oriented audit signals.

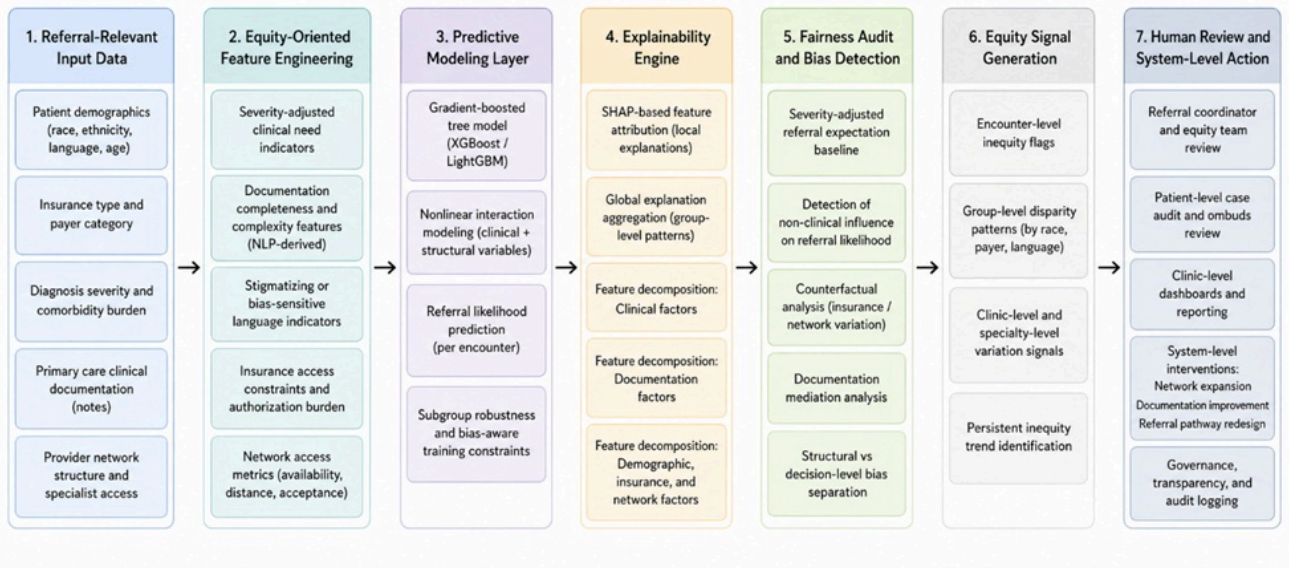


Figure 1. Explainable AI Pipeline for Equity-Oriented Detection of Inequitable Specialty Referral Patterns

Core input features

The core inputs would include structured demographic variables, insurance type, diagnosis severity indicators, note-derived documentation features, and provider network measures that describe access to relevant specialists. Demographic variables such as race, ethnicity, language, age, and socioeconomic proxies should be handled as fairness-audit attributes rather than simplistic biological risk markers, consistent with concerns about race correction and biased clinical algorithms [13, 19]. Documentation and network variables would be included because referral inequity may be mediated through incomplete notes, stigmatizing language, specialist scarcity, payer restrictions, or referral network exclusion rather than direct demographic effects alone [15, 16].

Design principles

The model should be fairness-aware, transparent, clinically interpretable, and explicitly designed to identify patterns rather than punish individual clinicians. This distinction is important because algorithmic tools in healthcare can misclassify systemic inequity as individual behavior if they are implemented without governance, context, and stakeholder review [12, 20]. The design should therefore emphasize accountable auditing, careful subgroup evaluation, and human review of explanations, particularly where EHR-derived features may reproduce historical bias or underrepresentation [18, 22].

Data Sources and Equity-Oriented Feature Engineering

Identifying the referral-eligible cohort

The referral-eligible cohort would be defined from primary care encounters associated with diagnoses, symptoms, or care pathways for which specialty evaluation is commonly considered, using EHR, referral order, scheduling, and referral management data. This cohort definition should capture both completed referrals and clinically plausible missed opportunities, because models trained only on observed referrals may learn existing inequitable referral behavior as if it were appropriate practice [5, 6]. Electronic referral and automation systems provide useful process signals, but the equity-oriented cohort must be constructed to support auditing of who did and did not enter specialty pathways under comparable clinical circumstances [7, 8].

Building demographics, insurance, and severity features

Demographic and insurance features would be drawn from structured EHR fields, including race, ethnicity, preferred language, age, payer category, and other available indicators relevant to access and equity. Severity features would be built from diagnosis codes, comorbidity information, condition-specific staging, symptom descriptors, prior treatment history, and other clinically meaningful indicators that help establish expected referral need [9, 14]. These features must be interpreted carefully because race, payer, and utilization variables may operate as proxies for prior unequal access, making fairness review essential when they appear to influence model predictions [10, 13].

Table 1 presents a structured decomposition of referral decision drivers, distinguishing clinically appropriate factors from documentation, demographic, insurance, and network features that may introduce inequitable influence.

Table 1. Conceptual Decomposition of Referral Decision Drivers into Clinical Necessity, Structural Constraints, and Potential Sources of Inequity

Domain Category	Feature Type	Operational Definition	Role in Referral Decision	Equity Interpretation Risk	Audit Priority Level
Clinical Necessity	Diagnosis severity	Disease stage, acuity, symptom burden	Legitimate driver of referral need	May reflect delayed diagnosis in underserved groups	High
Clinical Necessity	Comorbidity burden	Multimorbidity and complexity	Increases referral probability	May encode prior unequal care access	Medium
Documentation	Note completeness	Length, detail, structured elements	Signals clinical clarity	Underdocumentation may suppress referrals	High
Documentation	Urgency language	Expressions of severity/concern	Influences perceived need	Language bias may vary by patient group	High
Documentation	Stigmatizing descriptors	Bias-sensitive wording	May reduce perceived appropriateness	Direct pathway for inequity	Very High
Demographic Factors	Race/ethnicity	Self-reported or recorded identity	Should not directly influence referrals	Proxy for systemic inequity	Critical
Demographic Factors	Language	Preferred language	May affect communication and referral	Reflects structural access barriers	High
Insurance Factors	Payer type	Medicaid, private, uninsured	Affects access to specialists	Major structural inequity pathway	Critical
Network Structure	Specialist availability	Density and supply	Enables or constrains referrals	Structural inequity driver	Critical
Network Structure	Acceptance patterns	Insurance participation	Determines feasibility of referral	Reinforces payer-based disparity	Critical

Constructing documentation depth and network structure features

Documentation features would be derived from primary care notes using natural language processing methods that estimate note completeness, complexity, urgency, symptom burden, uncertainty, and the presence of potentially stigmatizing language. Referral-related NLP is well suited to extracting structured signals from unstructured clinical text, but such signals must be audited because documentation patterns may differ systematically across patient groups [8, 15]. Provider network features would represent specialist density, travel burden, payer acceptance, scheduling availability, and referral relationship structure so that the model can distinguish decision bias from structural barriers that constrain referral feasibility [1, 3].

Explainable Model Architecture and Fairness-Audit Mechanism

Predictive model – gradient-boosted trees

A gradient-boosted tree model, such as XGBoost or LightGBM, would be appropriate because referral decisions are likely shaped by nonlinear interactions among severity, documentation, insurance, and network access. Tree-based models are also compatible with SHAP explanations, making them suitable for linking predictive outputs to interpretable feature contributions at both the patient and group levels [11, 21]. Model development should prioritize subgroup robustness, proxy assessment, and bias monitoring rather than optimization alone, because high-performing healthcare models can still produce inequitable outputs when trained on historically biased EHR data [18, 22].

Fairness-oriented SHAP analysis

Fairness-oriented SHAP analysis would separate feature contributions into clinically legitimate factors, potentially inequitable non-clinical factors, and structural access factors. At the individual level, this decomposition could show whether a lower predicted referral likelihood is mainly explained by low clinical severity, incomplete documentation, restrictive insurance, or limited specialist network availability [11, 21]. At the clinic or system level, aggregated explanations could reveal recurring patterns in which insurance type, race, language, or network constraints contribute to referral suppression among clinically comparable patients [10, 20].

Counterfactual explanations for bias investigation

Counterfactual explanations would support bias investigation by estimating how the predicted referral likelihood might change if a non-clinical attribute, such as insurance type or network access, were altered while clinical severity remained comparable. The purpose would not be to claim that a patient's identity can be changed, but to test whether the model has learned a referral pattern in which clinically similar patients are treated differently under different payer or access conditions [14, 19]. Such counterfactuals should be reviewed alongside SHAP explanations and organizational context because explainability methods can create false reassurance if they are treated as definitive proof rather than structured evidence for equity investigation [12, 23].

Disentangling Clinical Necessity from Bias

Controlling for diagnosis severity to establish an expected referral rate

The model would first estimate expected referral likelihood from clinical indicators such as diagnosis severity, comorbidity burden, symptom persistence, prior treatment response, and documented functional impact. This severity-adjusted baseline is essential because referral differences are not automatically inequitable when patients differ meaningfully in clinical need, but it is also insufficient if the severity indicators themselves reflect delayed diagnosis or prior underdiagnosis in underserved groups [24, 25]. The equity audit would therefore treat severity adjustment as a necessary but cautious step, recognizing that healthcare algorithms can reproduce hidden bias when labels and clinical measurements are generated within unequal care systems [9, 18].

Identifying undue influence from demographics and insurance

After establishing clinically expected referral likelihood, the model would examine whether race, ethnicity, language, or insurance type contributes to lower referral probability among otherwise comparable patients. This analysis would be grounded in fairness-aware healthcare modeling, where demographic parity, equal opportunity, and calibration by group are interpreted as diagnostic tools rather than universal definitions of justice [10, 14]. Because algorithms can infer sensitive attributes through correlated variables, the audit should also consider proxy pathways in which geography, utilization history, documentation depth, or network participation indirectly encode demographic and payer-based disadvantage [20, 26].

Documentation mediation analysis

Documentation mediation analysis would evaluate whether differences in note completeness, urgency language, symptom detail, and complexity descriptions partially explain referral gaps for minoritized or publicly insured patients. Stigmatizing and biased language in clinical records can influence how future clinicians interpret patient credibility, adherence, need, or appropriateness for specialty care, making documentation both a data source and a potential mechanism of inequity [15, 16]. The model should therefore distinguish between apparent clinical underdocumentation and direct referral decision bias, while recognizing that documentation itself may be shaped by structural racism, gender bias, and differential clinical attention [17].

Provider network as a structural confounder

Provider network structure would be treated as a structural confounder because referral likelihood may depend on specialist availability, travel burden, payer participation, referral relationships, and scheduling capacity. A patient with high clinical need may still face reduced referral probability if local specialists do not accept their insurance or if the primary care clinic has weaker specialist connections, making network access a potential driver of inequity rather than a neutral operational variable [1, 2]. The model could simulate how referral patterns would be expected to change if network constraints were reduced, helping separate clinician-level decision variation from broader structural barriers in specialty access [4, 6].

Explainability and Actionable Equity Insights

Clinic-level audit dashboards

Clinic-level audit dashboards would aggregate explanation patterns to show whether non-clinical features repeatedly influence referral likelihood across diagnoses, specialties, payers, and demographic groups. Instead of presenting isolated predictions, the dashboard would translate SHAP patterns into equity-relevant signals, such as recurring payer-related suppression or language-associated reductions in expected referral likelihood among clinically similar patients [11, 21]. Such dashboards should be embedded in governance processes because explainability alone does not guarantee fairness, and leadership must interpret model outputs in light of institutional context and historical inequity [12, 27].

Patient-level explanations for ombuds and case review

Patient-level explanations would support ombuds, referral coordinators, and health equity reviewers by showing which factors contributed to the model's referral expectation for a specific encounter. The explanation could clarify whether a low referral likelihood was primarily associated with low documented severity, limited specialist network access, incomplete documentation, or potentially inequitable non-clinical factors such as payer status [5, 7]. Because model explanations may be persuasive even when incomplete, case review should include clinical context, patient preferences, and human judgment rather than relying on explanation outputs as definitive determinations [12, 23].

Counterfactual-driven intervention planning

Counterfactual-driven intervention planning would help health systems reason about which system changes could reduce referral inequity. For example, a counterfactual analysis could conceptually examine whether referral likelihood would be expected to improve if specialist access expanded, if documentation completeness improved, or if payer-related barriers were reduced while clinical severity remained similar [2, 4]. This approach aligns with fairness-aware interpretable modeling by shifting the use of AI from individual blame toward identifying modifiable organizational levers, including network expansion, referral navigation, and documentation support [20, 28].

Table 2 translates explainable model outputs into a structured framework linking detected inequity signals to targeted organizational, clinical, and policy-level interventions.

Table 2. Framework for Mapping Explainability Outputs to Actionable Equity Interventions in Specialty Referral Systems

Explainability Signal Type	Observed Pattern	Interpretation Context	Potential Root Cause	Recommended Intervention	Level of Action
SHAP (patient-level)	Insurance negatively impacts referral likelihood	After severity adjustment	Payer acceptance limitations	Expand specialist participation	System
SHAP (group-level)	Lower referral scores for specific racial groups	Across similar diagnoses	Structural or implicit bias	Equity-focused clinical review and training	Clinic
Documentation signal	Lower completeness in certain populations	NLP-derived features	Unequal clinical attention	Documentation standardization tools	Provider
Counterfactual result	Referral likelihood increases with insurance change	Simulated condition	Access inequity	Policy advocacy or payer reform	System
Network constraint signal	Reduced referral likelihood in low-density regions	Geographic analysis	Specialist scarcity	Telehealth expansion / network redesign	System
Persistent disparity pattern	Repeated inequity across clinics	Longitudinal signal	Systemic workflow bias	Referral pathway redesign	Organization
High variance across providers	Referral inconsistency	Same clinical scenarios	Practice variation	Clinical decision support guidelines	Provider
Language-related effect	Lower referrals for non-native speakers	Communication barrier	Interpreter or access issues	Language support services	System

Transparent accountability and oversight

Transparent accountability would require logs of model predictions, explanation summaries, audit flags, review decisions, and organizational responses. These logs would allow internal equity committees and external reviewers to examine whether referral patterns reflect legitimate clinical variation, structural network constraints, biased documentation, or potentially discriminatory decision pathways [10, 19]. Oversight is particularly important because machine learning systems in healthcare can amplify disparities when demographic representativeness, label quality, and subgroup performance are not continuously assessed [22, 29].

Clinical Integration and Organizational Response

Embedding into existing referral coordination workflows

The model would be best integrated as a silent audit layer within existing referral coordination workflows rather than as a directive alert for frontline clinicians. Referral coordinators, access teams, or health equity officers could review flagged patterns, compare explanations with clinical context, and decide whether outreach, case review, or process redesign is warranted [6, 7]. This workflow respects clinical discretion while recognizing that algorithmic fairness tools are most useful when paired with accountable organizational response, transparent governance, and safeguards against punitive interpretation [20, 27].

Systemic interventions informed by model outputs

System-level outputs could identify specialties, clinics, payer groups, languages, or geographic areas where referral likelihood appears lower than clinically expected. These findings could guide targeted interventions such as expanding specialist panels, improving Medicaid-participating networks, strengthening referral navigation, standardizing documentation practices, or training teams to recognize biased language in clinical notes [2, 15]. The model should therefore function as an equity learning system that points to modifiable processes, not as a tool for assigning fault to individual clinicians or patients [19, 28].

Evaluation Strategy

Predictive accuracy and fairness metrics

The evaluation strategy should include conventional predictive assessment, fairness metrics, and subgroup calibration, but the article's conceptual framing avoids reporting fabricated performance values. AUROC, calibration by group, equal opportunity difference, and demographic parity difference could be evaluated to determine whether the model distinguishes clinically expected referrals while avoiding systematic underprediction for protected or disadvantaged groups [10, 14]. These metrics should be interpreted cautiously because fairness definitions may conflict, and a model can appear statistically balanced while still relying on biased clinical labels or proxy variables [20, 22].

Explanation quality and audit utility

Explanation quality should be evaluated by whether clinicians, health equity officers, referral coordinators, and patient advocates find the model's explanations understandable, plausible, and useful for investigation. SHAP summaries, local explanations, and counterfactuals should be assessed for whether they reveal actionable patterns rather than merely restating correlations embedded in EHR data [11, 21]. Because explainable AI in healthcare can produce explanations that are technically coherent but operationally misleading, evaluation should include stakeholder review, failure analysis, and governance checks [23, 27].

Prospective impact on referral equity

A prospective evaluation could examine whether model-informed interventions are associated with improved referral equity over time, while accounting for secular trends, policy changes, staffing shifts, and specialist network changes. The goal would be to determine whether the audit framework helps reduce unexplained referral gaps, improve referral completion pathways, and support more equitable access to specialty care without encouraging inappropriate over-referral [3, 5]. Such evaluation should remain attentive to unintended consequences, including the possibility that AI tools may redistribute attention unevenly or obscure deeper structural inequities if not paired with organizational accountability [25, 29].

Limitations

Proxy and measurement limitations

Race and ethnicity data in EHRs may be incomplete, inconsistent, or misclassified, while insurance type may only partially represent socioeconomic disadvantage. Referral appropriateness is also difficult to label perfectly because observed referrals reflect existing clinical practice, patient preference, specialist availability, payer rules, and documentation quality [13, 18]. These measurement limitations mean the model should be treated as an audit-support system that surfaces possible inequity for review, not as an authoritative classifier of discrimination [12, 23].

Risk of misinterpretation and stigmatization

A major risk is that explanations could be misconstrued as blaming individual clinicians, patients, or demographic groups for inequitable referral patterns. To prevent this, implementation should frame the model as a system-level fairness audit that examines how documentation practices, payer restrictions, referral networks, and institutional routines shape access [19, 27]. Governance should also guard against stigmatizing uses of model outputs, particularly when explanations involve sensitive demographic attributes, insurance status, or language-related access barriers [15, 20].

Conclusion

An explainable artificial intelligence model for detecting inequitable specialty referral patterns could help health systems identify where referral decisions diverge from clinically expected pathways. By incorporating patient demographics, insurance type, diagnosis severity, documentation features, and provider network structure, the model would support a more complete understanding of how inequity enters referral workflows.

The strength of the proposed framework lies in its combination of severity adjustment, transparent SHAP decomposition, counterfactual reasoning, and dual patient-level and aggregate auditing. This structure would allow health systems to examine individual referral concerns while also identifying recurring organizational patterns that require broader intervention.

Important challenges remain, including incomplete demographic data, biased clinical documentation, proxy variables, uncertain labels for referral appropriateness, and the risk that explanations may be overinterpreted. The model would need careful governance to ensure that it is used for systems improvement, patient advocacy, and equitable access rather than punitive surveillance.

Health system collaboratives should pilot and share transparent referral equity audit frameworks that move beyond retrospective disparity reports. With responsible design and governance, explainable AI could support proactive, accountable, and clinically meaningful interventions to reduce inequitable specialty referral patterns.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 07 Dec 2023

Revised: 25 Jan 2024

Accepted: 24 Mar 2024

Published online: 20 July 2024

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

Landon BE, Onnela JP, Meneades L, O'Malley AJ, Keating NL. Assessment of racial disparities in primary care physician specialty referrals. *JAMA Netw Open*. 2021;4(1):e2029238.

Geissler KH, Shieh MS, Krishnan JA, Lindenauer PK, Ash AS, Goff SL, et al. Health insurance type and outpatient specialist care among children with asthma. *JAMA Netw Open*. 2024;7(6):e2417319.

Srinivasan AP, Sullivan E, Desai A, Adiele A, Bell S, Lion KC. Disparities in pediatric specialty referral scheduling and completion. *J Pediatr*. 2023;260:113485.

Schuldt R, Jinnett K. Barriers accessing specialty care in the United States: a patient perspective. *BMC Health Serv Res*. 2024;24(1):1549.

Patel MP, Schettini P, O'Leary CP, Bosworth HB, Anderson JB, Shah KP. Closing the referral loop: an analysis of primary care referrals to specialists in a large health system. *J Gen Intern Med*. 2018;33(5):715-21.

Ramelson HZ, Nederlof A, Karmiy S, Neri PM, Kiernan D, Krishnamurthy R, et al. Closing the loop with an enhanced referral management system. *J Am Med Inform Assoc*. 2018;25(6):715-21.

Odisho AY, Lui H, Yerramsetty R, Bautista F, Gleason N, Martin E, et al. Design and development of referrals automation, a SMART on FHIR solution to improve patient access to specialty care. *JAMIA Open*. 2020;3(3):405-12.

Fudickar S, Bantel C, Spieker J, Töpfer H, Stegeman P, Schiphorst Preuper HR, et al. Natural language processing of referral letters for machine learning-based triaging of patients with low back pain to the most appropriate intervention: retrospective study. *J Med Internet Res*. 2024;26:e46857.

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.

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.

Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J Am Med Inform Assoc*. 2020;27(7):1173-85.

Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745-50.

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

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

Sun M, Oliwa T, Peek ME, Tung EL. Negative patient descriptors: documenting racial bias in the electronic health record. *Health Aff (Millwood)*. 2022;41(2):203-11.

Himmelstein G, Bates D, Zhou L. Examination of stigmatizing language in the electronic health record. *JAMA Netw Open*. 2022;5(1):e2144967.

Beach MC, Saha S, Park J, Taylor J, Drew P, Plank E, et al. Testimonial injustice: linguistic bias in the medical records of Black patients and women. *J Gen Intern Med*. 2021;36(6):1708-14.

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.

Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019;322(24):2377-8.

Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng*. 2023;7(6):719-42.

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.

Bozkurt S, Cahan EM, Seneviratne MG, Sun R, Lossio-Ventura JA, Ioannidis JPA, et al. Reporting of demographic data and representativeness in machine learning models using electronic health records. *J Am Med Inform Assoc*. 2020;27(12):1878-84.

Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Appl Sci (Basel)*. 2021;11(11):5088.

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.

Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med*. 2021;27(1):136-40.

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):e406-14.

Amann J, Blasimme A, Vayena E, Frey D, Madai VI; Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20(1):310.

Liu M, Ning Y, Ke Y, Shang Y, Chakraborty B, Ong MEH, et al. FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare. *Patterns (N Y)*. 2024;5(10):101059.

Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol*. 2018;154(11):1247-8.