Integrating Group and Individual Fairness in Clinical AI: A Post-Hoc, Model-Agnostic Framework for Fairness Auditing

Javen Xu[†1], Yeon-Mi Hwang[*†1], Inés Dormoy[1], Serena Liang Jing[2], Malvika Pillai[1,3], Catherine Curtin[4], Tina Hernandez-Boussard[1,5]

[1]Department of Medicine (Biomedical Informatics), Stanford University, Stanford, CA, USA
[2]School of Medicine, Stanford University, Stanford, CA, USA
[3]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA
[4]Department of Surgery, Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA
[5]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

[*]Corresponding author: ymh@stanford.edu [†]These authors contributed equally.

Abstract

Ensuring fairness across diverse patient populations is a fundamental challenge for clinical AI systems, yet current fairness evaluation approaches create critical blind spots. Group-level metrics capture systemic disparities but miss patient-level variations, while individual fairness frameworks ensure consistency but potentially obscure structural biases. In this paper, we propose EquiLense, a post-hoc, model-agnostic framework that bridges these perspectives through clinical similarity matching and comprehensive fairness auditing. Our method introduces the Mean Predicted Probability Difference (MPPD), which quantifies prediction inconsistencies between clinically similar patients across demographic groups, integrating both individual-level consistency and group-level equity assessment. Moreover, we provide flexible similarity matching using clinical features and comprehensive visualization tools that support practical deployment in healthcare settings. Applied to electronic health record data from over 59,000 surgical patients, our framework revealed disparities in prediction consistency even when overall model performance appeared strong. EquiLense identified differences in predicted probabilities between clinically similar patients from different racial groups, disparities that were substantially reduced when sensitive attributes were excluded from model training. Our method provides a clinically relevant and interpretable approach to fairness auditing that enables healthcare practitioners to identify, understand, and address algorithmic disparities in real-world deployment settings.

Data and Code Availability This study used electronic health record (EHR) data of Stanford University. Due to institutional and patient privacy restrictions, these data are not publicly available. Code is available at `https://github.com/su-boussard-lab/mppd_fairness`.

Institutional Review Board (IRB) The study was approved by the institutional review board at Stanford University (Stanford University, Stanford, CA, USA; IRB-34551)

1 Introduction

As artificial intelligence (AI) becomes increasingly integrated in clinical decision-making, ensuring algorithmic fairness is both an ethical priority and a practical challenge.[1–4] The fundamental challenge of determining when algorithmic unfairness is present in clinical AI remains inadequately addressed due to fragmented fairness evaluation approaches.[5] Most current fairness evaluations focus on group-level disparities (e.g., across racial or gender categories) or individual-level inconsistencies.[2, 6] This divide creates blind spots: group metrics can obscure important patient-level variations, while individual metrics may fail to reveal systemic biases affecting entire demographic groups.[7]

Group-level metrics are valuable for identifying systemic disparities that disadvantage marginalized populations, yet they cannot reveal whether clinically similar individuals across demographic groups receive consistent predictions.[2, 8, 9] In contrast, individual fairness frameworks, grounded in the principle that similar patients should receive similar outcomes, ensure prediction consistency for comparable patients but often obscure inequities rooted in broader structural bias.[5, 10, 11]

To address these limitations, several integrated fairness frameworks have been proposed. Speicher et al.[12] introduced a generalized entropy-based metric that decomposes total unfairness into between-group and within-group components, offering a unified mea-

sure of disparity. However, its practical application can be less intuitive in clinical settings. Xu and Strohmer [6] presented a theoretical framework that characterizes the conditions under which group and individual fairness are compatible, focusing on post-processing trade-offs but without operational implementation. In the context of graph neural networks, FairGI jointly optimizes group fairness and within-group individual fairness using adversarial learning and similarity matrices.[13] Panda and Louis[14] proposed a probabilistic matching method that aims to ensure individuals are likely to receive favorable outcomes while also satisfying group-level fairness constraints. While these approaches contribute important conceptual and algorithmic advances, they often fall short in delivering the clinical relevance, interpretability, and usability required for deployment in healthcare settings.

We introduce EquiLense, a clinical fairness auditing tool that bridges the gap between group-level equity and individual-level consistency in a clinically interpretable way. Its core innovation is the Mean Predicted Probability Difference (MPPD), a post-hoc metric that quantifies prediction inconsistencies between clinically similar patients across demographic groups. MPPD captures how inconsistencies differ across groups, reflecting both patient-level variation and structural disparities. Designed for healthcare, EquiLense offers practical advantages over existing frameworks: it operates on clinically interpretable features for similarity matching and provides adjustable code-based functions and visual summaries of fairness. To our knowledge, this is among the first approaches in the clinical domain to combine cross-group patient matching with an interpretable metric and visualization, supporting more comprehensive fairness auditing in clinical AI.

2 Method

2.1 Study design and population

To demonstrate the application of our tool, we used EHR data from an integrated healthcare organization comprising an academic medical center, a primary and specialty care alliance, and a community medical center. All facilities operate on a unified Epic-based EHR. We examined EHR data from 59,047 surgical patients treated between January 1, 2012 to December 1, 2022. The cohort included patients aged 50 years or older who had inpatient stays of 90 days or less and survived at least 30 days post-surgery. Our analysis incorporated all post-surgicial clinical documentation, including progress and nursing notes. For patients with multiple surgeries during the study timeframe, only data from the initial procedure was included. Cohort details are available in Figure S1 and described in detail elsewhere.[15]

2.2 Model Development

2.2.1 Outcome

The primary outcome was post-surgical delirium, identified through a comprehensive multi-modal approach. We used ICD diagnostic codes (Table S1) and documented positive Confusion Assessment Method (CAM) scores, a validated delirium assessment tool. To enhance case ascertainment, we supplemented this phenotype using natural language processing of free-text clinical notes (Method S1), as delirium is frequently underreported in structured data fields. The secondary outcome was 30-day hospital readmission, defined as any inpatient admission occurring within 30 days of discharge.

2.2.2 Features

We categorized features into sensitive and clinical attributes. Sensitive attributes included sex, race/ethnicity, and insurance type (as a proxy for socioeconomic status). While sex can serve as either a sensitive or clinical attribute depending on context, we classified it as sensitive for this study. Clinical attributes included age, body mass index (BMI), pre-surgical pain score, Charlson Comorbidity Index (CCI), opioid-naïve status, mental health diagnoses, anesthesia type, and surgery type. We used the most recent BMI recorded prior to surgery. Pre-surgical pain scores were collected from a 180-day window before surgery through the day of surgery. The CCI was calculated using diagnoses from the two years preceding surgery. For descriptive analyses, we categorized CCI scores as low (0–2), moderate (3–4), or high (≥ 5), while continuous scores were used in machine learning models. We defined opioid-naïve status as having no opioid prescription in the 6 months before surgery. Mental health diagnoses were identified using the five most frequent ICD-10 F-code categories in our cohort: mental organic disorders, substance use disorders, mood disorders, anxiety disorders, and behavioral symptoms (Table S1). Anesthesia type was determined through rules-based natural language processing of operative and anesthesia procedure notes, with cases involving both general and regional anesthesia categorized as hybrid. Surgery type was extracted from surgical notes (Method S2).

All features were preprocessed using standard normalization and one-hot encoding. Missing values

were imputed using median values.

2.2.3 Models

We used a 80/20 train-test split. To address class imbalance, we applied the Synthetic Minority Oversampling Technique (SMOTE) to the training set.

Models were developed using logistic regression and extreme gradient boosting (XGB). Model performance was assessed using accuracy, precision, recall, and area under the receiver operating characteristic curve (AUROC). All analyses were conducted using scikit-learn (v1.3.0) and XGBoost (v2.1.1) in Python. We also evaluated the contribution of each sensitive attribute to model performance and fairness by training models with and without individual sensitive attributes and examining the resulting disparities. We note that the models were not optimized for predictive performance or clinical deployment but serve as demonstration cases to illustrate the application of the EquiLense fairness auditing framework.

2.3 Fairness Evaluation

2.3.1 Fairness evaluation visualization framework

Group fairness was evaluated using Demographic Parity Difference (DPD), Equal Opportunity Difference (EOD), Equalized Odds (EO), and the Error Distribution Disparity Index (EDDI).

Individual fairness was assessed by measuring prediction consistency among clinically similar patients, with similarity defined using distance-based metrics. We calculated individual fairness metrics with both a narrow set of clinical variables (age, CCI, BMI) and the full feature set to examine how the definition of clinical similarity influences fairness estimates. For the full feature set, comorbidities included in the CCI were one-hot encoded. All metrics were computed on the test set.

Currently, there is currently no consensus on how to define clinical similarity in individual fairness assessments. Similarity may vary depending on the context, population, and prediction task. To support flexible exploration, our codebase allows users to specify their own similarity criteria, including variables, distance metrics, and thresholds. This enables customized fairness analyses across diverse clinical use cases.

2.3.2 Mean Predicted Probability Difference (MPPD)

To evaluate fairness at both the group and individual levels, we developed a novel integrated metric, the Mean Predicted Probability Difference (MPPD). This metric quantifies inconsistencies in predicted probabilities between clinically similar patients who belong to different sensitive groups.

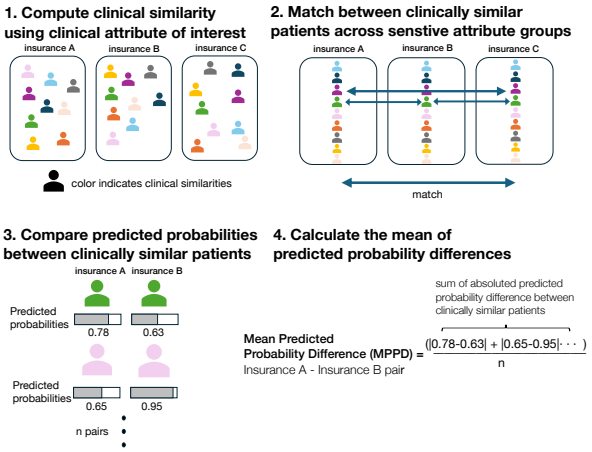

Figure 1: Conceptual Overview of Mean Predicted Probability Difference (MPPD) metric.

We had two approaches: 1) reference-based and 2) pairwise. The reference-based approach begins by identifying the largest group within each sensitive attribute (e.g., Non-Hispanic White patients for race/ethnicity) as the reference group. For each patient in the reference group, we identify clinically similar individuals purely based on cosine similarity computed on clinical attributes. The similarity threshold, set to 0.01 by default but user-selectable, defines matched pairs. The set of clinical and social attributes used to define similarity can also be tailored based on user judgment. For example, for a Non-Hispanic White patient, they could have several matched pairs as long as the other patient's clinical similarity distance is within 0.01 (the similar patients could be Non-Hispanic White patients, Non-Hispanic Asian patients, Hispanic patients, or from any of the race/ethnicity groups). They might also have no matched pair if there is no other patient within 0.01 of clinical similarity distance.

For each matched pair, we calculate the absolute difference in model predicted probabilities, and we aggregate these differences by group (for all of reference and non-reference groups) to obtain the average MPPD value for each group. This process captures whether patients from non-reference groups re-

ceive systematically different predictions despite having similar clinical profiles, thereby integrating elements of group fairness (via inter-group comparison) and individual fairness (via intra-clinical similarity).

To complement the reference-based approach and avoid assumptions about any one group being "fair" by default, we extended MPPD to a pairwise version. In this version, we performed bi-directional matching between all combinations of sensitive attribute groups, allowing identification of clinically similar patients across group-pairs (e.g., Black–White, Asian–Hispanic). For each group-pair, we computed the average absolute difference in predicted probabilities between matched patients, producing a matrix of pairwise MPPD scores. This symmetric comparison allows for a more comprehensive understanding of prediction consistency across all groups, including disparities between minority populations that would be overlooked in a reference-only framework.

Together, these two implementations of MPPD provide a flexible and interpretable method for uncovering fairness concerns in clinical prediction models. The interpretability of the framework comes from its use of clinically familiar features already involved in model development, while its flexibility lies in allowing users to define clinical similarity according to their own context (e.g., age, comorbidity burden, surgical type). This alignment with routine clinical reasoning makes the outputs easier for practitioners to understand and apply in practice.

3 Result

3.1 Descriptive Statistics

Our cohort comprised 59,047 elderly patients who underwent 75,179 surgeries at a large academic medical center (Table S2 and Figure S1). The prevalence of post-surgical delirium was 14.9% and 30-day readmission was 8.8%. The median age at surgery was 66 years, with 52% of patients being female. The majority of patients identified as Non-Hispanic White (63%), followed by Non-Hispanic Asian (13%) and Hispanic (12%). At the time of surgery, 59% had private insurance (Table S2).

3.2 Model Performance

For delirium prediction, the XGBoost model achieved an accuracy of 0.84, AUROC of 0.78, precision of 0.45, and recall of 0.26 (Table S3). Removing individual sensitive attributes (race/ethnicity, sex, or insurance) had minimal impact on model performance, with accuracy and AUROC remaining stable at 0.83 and 0.77, respectively, representing only minor changes in precision (0.40–0.43) and recall (0.27-0.29). The XGBoost model demonstrated higher overall performance for 30-day readmission prediction compared to LR, achieving an accuracy of 0.91, AUROC of 0.61, precision of 0.19, and recall of 0.02. Excluding sensitive attributes again had negligible effects on performance metrics, with only small fluctuations noted. Logistic regression models performed worse than XGBoost across both outcomes. For delirium prediction, logistic regression achieved an accuracy of 0.69 and AUROC of 0.78, while for readmission prediction, accuracy was 0.64 and AUROC was 0.68. In both models, sensitive attributes had low-to-moderate feature importance and minimal effect on predictive performance when removed.

In the delirium model (Figure S2), key predictors included CCI, pre-operative pain scores, mental health diagnoses, and specific surgery types (e.g., coronary artery bypass grafting [CABG], hernia repair, prostatectomy), which showed moderate importance. In the 30-day readmission model (Figure S3), important features included anesthesia type, comorbidity burden, and mental health conditions. Sensitive features such as race/ethnicity and insurance type had low feature importance in both models, while sex showed moderate importance in the readmission model.

3.3 Individual and Group Fairness Metric Visualization

Figure 2a shows a radar plot summarizing group fairness metrics for the delirium prediction model, using race as the sensitive attribute. Both maximum and average values are presented for DPD, EOD, EO, and EDDI, with disparity thresholds set at 0.01 (low), 0.05 (moderate), and 0.1 (high). EDDI shows the largest disparity, reflecting differences in prediction error distributions between groups. While EDDI captures an aggregate measure of error distribution, Figure 2b focuses on EO by showing its true positive rate and false positive rate components. EO was selected for this breakdown because the true positive rate component is directly related to clinical sensitivity, making it especially relevant in a biomedical context.

Figure 2b presents EO difference values for each race and ethnicity group. Bars are color-coded according to disparity thresholds, with values closer to zero indicating lower disparity. Asian and non-Hispanic other patients show the highest EO differences (0.0389), both falling in the moderate disparity range, while White patients have the lowest EO dif-

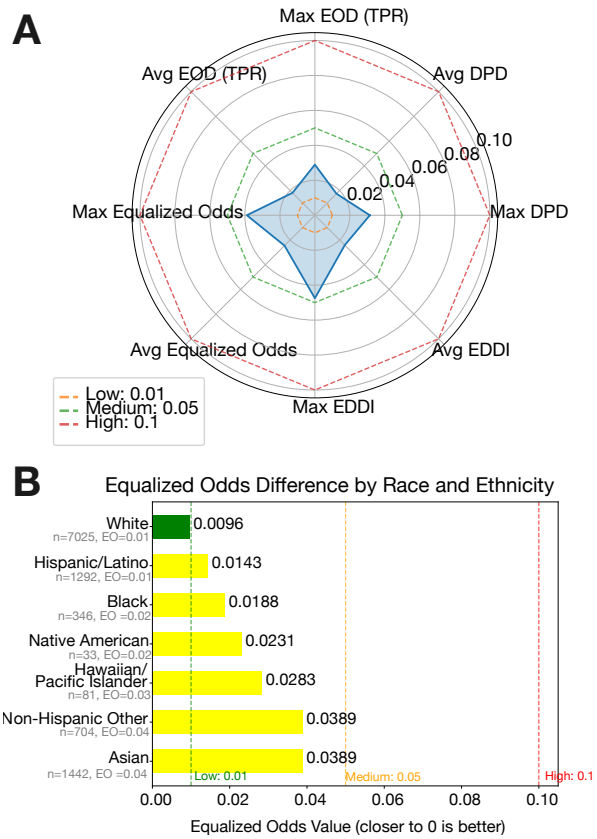

Figure 2: Visual Summary of Group Fairness Metrics in the Delirium Prediction Model

ference (0.0096), within the low disparity threshold. This group-level view highlights where disparities in EO are most pronounced.

Figure 3 shows individual fairness gaps across different clinical feature sets and distance thresholds. Individual fairness was measured as the average difference in predicted probabilities between clinically similar patients, where lower scores indicate better fairness (that is, smaller differences in predictions among similar individuals). Models using the full set of clinical features, including key predictors such as surgery type and anesthesia type, consistently achieved lower individual fairness gaps compared to narrower feature sets based on selected variables (such as age, BMI, and CCI) or comorbidity features alone. Using a more comprehensive set of clinically relevant features leads to more consistent predictions among similar patients and improves individual fairness across distance thresholds. However, clinical relevance may vary depending on the outcome and modeling context, so fairness assessments should be grounded in the specific clinical use case.

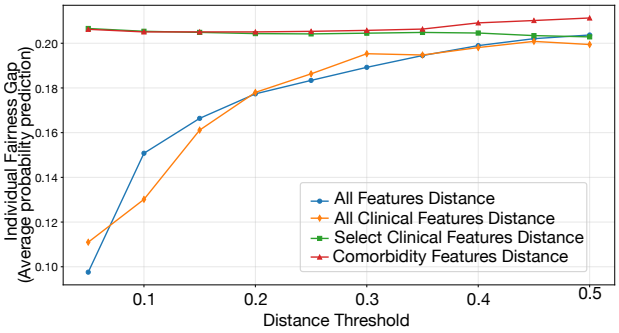

Figure 3: Individual Fairness Gaps by Feature Set and Distance Threshold

3.4 Mean Predicted Probability Difference

Figure 4a presents an integrated group fairness-individual fairness evaluation of the delirium model by visualizing MPPD among clinically similar patients across sensitive attributes. This analysis highlights how prediction inconsistencies, reflecting deviations from individual fairness, vary by demographic group. For instance, among patients clinically similar to a White patient, those identifying as Asian exhibited the largest disparity in MPPD (MPPD White-Asian = 0.119, MPPD White-White = 0.074, $\Delta = 0.045$), indicating lower individual fairness for that group. When the race/ethnicity variable was excluded from model training, these disparities were reduced, suggesting improved fairness in prediction consistency (Figure S4). Figure 4b displays pairwise MPPD values across racial and ethnic groups in the delirium prediction model. Each cell shows the average absolute difference in predicted probabilities between clinically similar patients from two different groups.

We note that within-group MPPD values (such as White–White) are not zero because each patient is compared to another clinically similar patient from the same group. When race and ethnicity were excluded from the feature sets used in model prediction (Figure S5), MPPD scores decreased substantially across most group-pairs, with the exception of the Black–Native American pair, likely due to the small test sample size of Native American patients (n = 32). Similar trends were observed in the 30-day readmission model.

4 Discussion

We present EquiLense, a model-agnostic post-hoc auditing tool designed to evaluate both group and in-

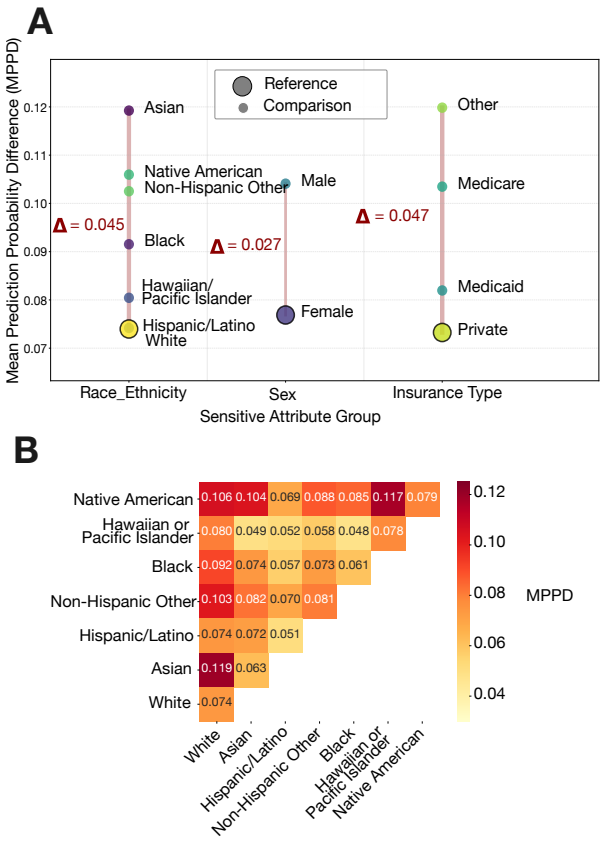

Figure 4: Mean Predicted Probability Differences (MPPD) Across Demographic Groups in the Delirium Prediction Model (A) Reference-Based MPPD (B) Pairwise MPPD for Race-Ethnicity Group

dividual fairness in clinical prediction models using real-world EHR data. Applying it to post-surgical delirium and readmission models, we demonstrate that models with strong overall performance can still yield uneven outcomes across demographic groups. EquiLense introduces the Mean Predicted Probability Difference (MPPD), an integrated metric that bridges group fairness and individual fairness to provide a more holistic evaluation. While we highlighted racial and ethnic disparities as one illustrative example, similar patterns were observed across other sensitive attributes, including insurance type. Our findings show that both model inputs and similarity definitions influence fairness outcomes, underscoring the importance of context-specific and interpretable fairness evaluations in clinical AI.

Compared to previous fairness frameworks, EquiLense addresses a key gap in clinical machine learning: the need for a post-hoc tool that combines individual-level consistency with group-based disparity assessment. Traditional group fairness metrics, while widely used, are aggregate and insensitive to individual-level variation within groups.[16, 17] Conversely, individual fairness metrics are rarely applied in healthcare due to challenges in defining clinically meaningful distances.[2, 11] Our method builds on the individual fairness principle by matching patients using clinical features directly relevant to the task, and surfaces how discrepancies in prediction outcomes align with sensitive attributes. By doing so, it bridges population- and patient-level perspectives, enabling stakeholders to pinpoint fairness issues that are both measurable and framed in clinically meaningful terms.

A growing body of work has sought to unify group- and individual-level fairness, including entropy-based decompositions,[12] theoretical compatibility frameworks,[6] and adversarial training methods in graph models.[13] Others have approached this integration through probabilistic assignment, such as the matching-based method, which balances fairness across individuals and groups during resource allocation.[14] While these methods contribute important conceptual and algorithmic insights, most are designed for benchmarks or allocation problems, with limited attention to interpretability or clinical applicability. In contrast, our method is tailored to evaluating model predictions in real-world clinical data, offering a model-agnostic tool for identifying disparities across and within groups based on clinically meaningful similarity. Importantly, EquiLense produces values (e.g. MPPD) that can be directly interpreted as prediction differences between clinically similar patients, allowing users to judge both whether disparities exist and their potential magnitude and impact in practice. The degree to which specific magnitudes of MPPD represent clinically meaningful or harmful disparities remains an open question that future studies should establish.

At first glance, our method may resemble counterfactual fairness approaches, but the underlying logic and assumptions differ fundamentally. Counterfactual fairness[18] evaluates whether a model's prediction would remain the same for a hypothetical version of the same individual with a different demographic identity. While conceptually appealing, this approach requires strong assumptions about the underlying causal structure,[19, 20] assumptions that are rarely verifiable in complex clinical settings due to unmeasured confounders, treatment heterogeneity, and documentation biases in EHR data.[21, 22] In contrast, EquiLense uses real, observed patients to assess prediction consistency across demographic groups, avoiding counterfactual inference altogether.

By matching clinically similar individuals from different groups and comparing their predicted probabilities, our method offers a practical, interpretable, and model-agnostic alternative grounded in real-world data. The use of pairwise comparisons across all group combinations further enables fairness evaluation without relying on a fixed reference group.

We observed that excluding sensitive attributes, such as race/ethnicity and insurance group, from model training resulted in lower MPPD scores, indicating improved fairness in prediction consistency across demographic groups. In our case, removing these attributes reduced disparities between clinically similar patients without materially affecting overall model performance, though we did observe increased MPPD scores for the Black-Native American pair due to limited Native American sample size. These sensitive attributes also had low feature importance, suggesting they contributed little to predictive accuracy. However, algorithmic bias can arise from multiple sources, including data imbalance, labeling variation, and broader structural inequities. There is rarely a single solution. Removing a sensitive attribute is one possible approach, but it does not guarantee improved fairness in all cases. EquiLense helps users evaluate these trade-offs by allowing them to test how different modeling choices, such as modifying feature sets, reweighting data, or applying post-processing methods, affect both fairness and model performance. Because fairness goals and clinical priorities vary by setting, EquiLense is designed to be flexible. Users can define fairness objectives, choose relevant metrics, and assess trade-offs in a way that aligns with the specific context of their model's application.

A key strength of EquiLense is its adaptability. Because fairness goals and clinical priorities vary across healthcare settings, the framework allows users to tailor evaluations to their context. For example, age may be treated as a clinical predictor in surgical risk stratification but as a sensitive attribute in fairness assessments, while sex may serve as a biological covariate or as a protected attribute depending on the analytic objective. EquiLense accommodates these choices by enabling users to specify which variables define clinical similarity and which are considered sensitive attributes, and by allowing adjustment of similarity thresholds and metric selection. This design ensures that fairness evaluation remains adaptable across diverse clinical use cases, supporting both researchers and practitioners in aligning fairness analyses with their specific priorities.

The delirium and readmission models in this study were used solely as illustrative cases to demonstrate EquiLense. While they highlight how fairness auditing can reveal disparities, future work should apply the framework to clinically validated, production-ready models to assess its utility in real-world deployment. These models, which showed relatively low precision and recall, were not intended for clinical use but served as realistic use cases for demonstrating fairness evaluation. Although performance could likely be improved with additional tuning, feature engineering, and incorporation of more clinical variables, model optimization was outside the scope of this work, which focused on showcasing the fairness evaluation capabilities of the EquiLense.

Despite its advantages, our approach has several limitations. First, MPPD is not a causal measure and cannot determine whether observed differences in predicted probabilities reflect unfair bias or clinically appropriate distinctions. However, MPPD builds on existing models and provides a systematic way to surface inconsistencies that warrant further investigation.

Second, the accuracy of fairness evaluation depends on how well clinical similarity is defined. Incomplete or noisy feature data may lead to mismatches that distort fairness estimates. Users should carefully consider which clinical features best capture meaningful similarity for their specific task. Furthermore, fairness estimates can be sensitive to the similarity threshold used to define matches. While users can adjust this threshold, small changes may affect who is considered similar, which can influence the resulting fairness scores. Future studies should investigate how different thresholds affect both fairness estimates and clinical interpretability, and whether standard guidelines can be developed for threshold selection in applied settings.

Third, defining clinical similarity inherently involves subjective decisions, such as which features to include or how to weigh them. EquiLense allows users to make these decisions flexibly, but this subjectivity is a fundamental limitation of individual fairness approaches. We also did not explore potential interactions between socio-demographic and clinical predictors, which may mask subtle biases. Future work could investigate how such interactions influence fairness evaluations.

Fourth, while EquiLense identifies disparities in model outputs, it does not assess the downstream clinical consequences of these disparities, such as differences in care decisions or health outcomes.

Fifth, the delirium and readmission models here were designed for demonstration rather than deployment. As such, some observed disparities may partly reflect calibration or recall/precision limitations of the models rather than true systemic bias. Applying

EquiLense to production-ready models and real-word deployment settings represents an important direction for future work.

Finally, our framework does not include mitigation strategies; it is designed as a diagnostic tool to support model auditing, transparency, and informed decision-making. As a future direction, simulated bias scenarios could be introduced to assess how MPPD scores respond to known disparities, helping to further validate and refine its interpretability and utility in different clinical contexts.

5 Conclusion

We introduced EquiLense, a flexible and model-agnostic tool for auditing fairness in clinical prediction models. At its core is the Mean Predicted Probability Difference (MPPD), a novel integrated metric that combines group- and individual-level perspectives to provide a more holistic assessment of disparities. Using post-surgical delirium and readmission models, we demonstrated how EquiLense can help surface fairness concerns, support sensitivity analyses, and guide decisions about model design and feature inclusion. Beyond these case studies, the framework is broadly applicable across prediction tasks, offering clinicians and researchers a practical way to evaluate fairness in real-world data. As algorithmic tools continue to influence care delivery, accessible frameworks like EquiLense will be essential for advancing responsible AI and ensuring that model performance aligns with equity-driven goals in healthcare.

6 Citations and Bibliography

References

[1] Irene Y Chen, Shalmali Joshi, and Marzyeh Ghassemi. Treating health disparities with artificial intelligence. *Nat. Med.*, 26(1):16–17, January 2020.

[2] Qizhang Feng, Mengnan Du, Na Zou, and Xia Hu. Fair machine learning in healthcare: A review. *arXiv [cs.LG]*, June 2022.

[3] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.*, 169(12):866–872, December 2018.

[4] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv [cs.LG]*, August 2019.

[5] Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. An empirical characterization of fair machine learning for clinical risk prediction. *J. Biomed. Inform.*, 113(103621):103621, January 2021.

[6] Shizhou Xu and Thomas Strohmer. On the (in)compatibility between group fairness and individual fairness. *arXiv [math.ST]*, January 2024.

[7] Chouldechova Alexandra and Roth Aaron. The frontiers of fairness in machine learning. *arXiv [cs.LG]*, October 2018.

[8] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019.

[9] Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine*, August 2020.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, New York, NY, USA, January 2012. ACM.

[11] Joshua W Anderson and Shyam Visweswaran. Algorithmic individual fairness and healthcare: a scoping review. *JAMIA Open*, 8(1):ooae149, February 2025.

[12] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. *arXiv [cs.LG]*, July 2018.

[13] Duna Zhan, Dongliang Guo, Pengsheng Ji, and Sheng Li. Bridging the fairness divide: Achieving group and individual fairness in graph neural networks. *arXiv [cs.LG]*, April 2024.

[14] Atasi Panda, Anand Louis, and Prajakta Nimbhorkar. Individual fairness under varied notions of group fairness in bipartite matching - one framework to approximate them all. *arXiv [cs.AI]*, August 2022.

[15] Malvika Pillai, Terri L Blumke, Joachim Studnia, Yuqing Wang, Zachary P Veigulis, Anna D Ware, Peter J Hoover, Ian R Carroll, Keith Humphreys, Thomas F Osborne, Steven M Asch, Tina Hernandez-Boussard, and Catherine M Curtin. Improving postsurgical fall detection for older americans using LLM-driven analysis of clinical narratives. *medRxiv*, page 2024.06.25.24309480, June 2024.

[16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv [cs.LG]*, October 2016.

[17] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[18] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Neural Inf Process Syst*, 30:4066–4076, March 2017.

[19] Judea Pearl. Causality. September 2009.

[20] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv [cs.AI]*, May 2018.

[21] Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2):345–352, April 2020.

[22] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due

to processes within the healthcare system: retrospective observational study. *BMJ*, 361:k1479, April 2018.