Uncovering Bias in Clinical Text: Leveraging LLMs for Fairness in EHR

Arshia Mubias Shaik, Divij Gera, Samarth Agarwal, Shireen Chand, Shreya Sethi
{ arshiamu, dgera, samartha, shireenc, shreyase }@usc.edu
University of Southern California

Abstract

Electronic Health Records are deeply enmeshed with modern healthcare, but often carry the prejudice that emanates from social inequalities and the biases of clinicians. When such biases are conflated with AI decision-making systems, they become a serious risk to existing inequalities.

In this project, we try to quantify and analyze biases in EHR-based predictive models across demographic attributes such as race, ethnicity, gender, and others. Large Language Models (LLMs) offer a possible avenue for both identifying as well as reducing biases. To quantify imbalances between model predictions, we evaluate fairness across a range of metrics including Demographic Parity, Equalized Odds, Mean Parity Disparity (MPD), Theil Index etc. In addition to testing for bias, we also explore the mitigation strategies to reduce these gaps, such that model performance is as balanced and equitable across multiple patient populations. Ultimately, the goal of this research is to enhance the integrity and fairness of AI-driven decision-making in medicine to render predictive models more actionable and representative.

1 Introduction and Motivation

Electronic Health Records (EHRs) contain inherent clinical and societal biases that risk being amplified by language models like BioGPT and ClinicalBERT in critical healthcare predictions. As these models see increasing clinical use, their potential for unequal performance across demographic groups (race, ethnicity, gender, and marital status) raises significant equity concerns. Our work addresses this challenge by developing a systematic framework to detect and quantify these biases in real clinical data (MIMIC-III) using fairness measures including Demographic Parity, Equalized Odds, and Mean Parity Disparity. Despite the critical importance of this problem, comparative studies examining bias across different models, prediction tasks, and demographic groups remain scarce, and standardized pipelines for fairness assessment in clinical NLP are notably lacking. This project aims to bridge these gaps through comprehensive bias measurement and mitigation approaches, advancing the development of more equitable healthcare AI systems.

2 Related Work and Contributions

A 2023 study by Chin et al. (2023) in the Journal of American Medical Association (JAMA) found clinical data-based LLMs required Black patients to present with more severe symptoms than White patients to recommend identical interventions and diagnosis. A 2024 article by Zack et al. (2024) in the Lancet journal tested the racial and gender biases of GPT-4 using clinical vignettes, revealing disparities in diagnostic and treatment recommendations.

A 2023 blog from the Radiological Society of North America Wahid (2023) outlined three primary approaches to measuring algorithmic fairness: demographic parity, equalized odds, and equalized opportunity. Demographic parity ensures that "the likelihood of a specific patient prediction should be independent of attributes such as gender, race, or age".

More recently, "fairness-aware prompts" were shown to be a viable way to reduce biases "without significantly compromising model performance" in a thorough bias review of LLMs in mental health analysis that was published in 2023 Wang et al. (2024). "FC prompting," which "consistently achieves the lowest EO scores across all models and datasets," was very successful and led to notable improvements in fairness.

Our research's main goal is to measure and analyze biases in EHR predictive models across demographic factors directly addresses gaps identified in current research. By focusing on calculating

model prediction disparities using fairness metrics such as demographic parity and equalized odds, this work aligns with recent recommendations for appropriate fairness evaluation methods.

Our proposed use of LLMs for bias identification is a new approach that expands on current research on the efficiency of quick engineering methods for bias reduction. By applying these techniques specifically to EHR data analysis, the project extends existing work on fairness-aware prompting strategies to a new domain with significant clinical implications.

3 Problem Description

Clinical language models like ClinicalBERT Huang et al. (2019) and BioGPT Luo et al. (2022) are increasingly used for healthcare prediction tasks, but concerns remain about their potential to amplify demographic biases in electronic health records (EHRs). These biases often appear as performance disparities across subgroups and may worsen existing healthcare inequalities. However, key gaps remain: limited comparison across model types (generative vs. discriminative), lack of standardized evaluation for intersectional bias, and little validation of whether fairness techniques from general NLP apply in clinical contexts.

Our work introduces a comprehensive framework to evaluate and mitigate bias in clinical LMs, highlighting why standard techniques often fall short and offering insights to guide the development of fairer, more reliable clinical AI systems.

4 Methods

4.1 Data and Annotations

We use the MIMIC-III clinical dataset, a large publicly available corpus of de-identified EHRs from ICU admissions. We primarily used unstructured clinical documents such as discharge summaries and progress notes for the purpose of this study. In addition to the textual data, we extracted demographic attributes including ethnicity (39 categories), religion (20 categories), gender (binary: Male/Female), and marital status (7 categories).

For supervised learning operations, we utilized the following clinical prediction labels:

- **Hospital Mortality** - Binary outcome indicating whether the patient survived or died.
- **Length of Stay (LOS)** - Continuous variable representing the number of days a patient stayed in the hospital.
- **Treatment Decisions** - A derived task where treatment outcomes were inferred from model outputs using structured treatment-related information in the EHR.

The demographic attributes served as sensitive attributes for our bias examination, and outcome labels to train and evaluate model performance.

4.2 Data Processing and Feature Extraction

To prepare the data for model training and fairness evaluation, we followed several steps of preprocessing and feature extraction:

- **Text Preprocessing:** Clinical notes were tokenized, lowercased, and truncated/padded to a fixed length (eg. 512 characters in ClinicalBERT).
- **Demographic Encoding:** Demographic attributes (e.g., religion, ethnicity) were encoded as categorical/binary variable and binarization as needed for fairness metrics.
- **Model Input:** Contextual embeddings were extracted using pretrained ClinicalBERT and BioGPT models.
- **Feature Extraction & Fairness Analysis:** Final-layer outputs were used for fairness evaluation across demographic groups using metrics like Demographic Parity and Equalized Odds.

4.3 Model Architecture and Training Procedure

Clinical Bert Fine-tuned using a classification head for *Hospital Mortality* (binary classification) and *Length of stay* (regression). For treatment prediction, we fine-tuned the model to generate treatments for patients based on their diagnosis and demographic attributes. Fine-tuning was done using cross-entropy loss (classification) and mean squared error loss (regression). Fine-tuning was done with the following: Adam optimizer, batch size = 16, learning rate = 2e-5, with early stopping based on validation loss.

BioGPT Fine-tuned using a causal language modeling objective for all tasks. For *Hospital Mortality* and *Length of stay*, prompts were constructed such as: "Patient diagnosed with X and is Y years old, belonging to ethnic group Z. Predict whether

the patient will survive." For treatment prediction, prompts included demographic and diagnosis info with generation of likely treatments. Due to high compute requirements, we limited training to 3–5 epochs with gradient accumulation and mixed precision (FP16) to optimize GPU memory usage.

5 Experimental Results

5.1 Experimental Setup

The major objective was to assess the fairness of two large language models BioGPT and ClinicalBERT. Both models were optimized for our prediction tasks with 3 training epochs.

An 80/20 split was used to randomly divide the dataset into training and testing sets. Using structured inputs that comprised patient diagnoses and demographic characteristics including gender, race, religion, and marital status, fine-tuning was carried out separately for every task. For every model, post-hoc fairness assessments and mitigation techniques were used independently.

5.2 Baselines and Evaluation Metrics

BioGPT and ClinicalBERT served as comparative baselines for one another. We evaluated each model across three critical clinical prediction tasks: hospital mortality, length of stay (LoS), and treatment generation. For each task, both predictive performance and fairness across demographic groups were analyzed.

Task	Metric (Description)
Mortality	Demographic Parity(DP): Equal survival prediction across groups Equalized Odds(EO): Compare true/false positive rates
Length of Stay	Mean Absolute Error(MAE): Avg. error between predicted and actual LOS Mean Prediction Disparity(MPD): Error gap across demographic groups Dispersion Ratio(DR): Variability in predictions across groups
Treatment	Disparate Impact(DI): Relative outcome favorability Theil Index(TI): Inequality in outcome distribution

Table 1: Fairness Metrics Used Across Prediction Tasks

To mitigate observed biases, three post-processing techniques were applied on each of the 3 tasks, respectively: (1) Threshold Optimization using Fairlearn, (2) group-wise mean correction, and (3) counterfactual data augmentation (CDA).

5.3 Results and Analysis

5.3.1 Bias Detection Results

The results of the evaluation of the presence of demographic bias in clinical predictions are summarized in Tables 2, 3, and 4.

Across all tasks, we observe that:

Attribute	BioGPT		ClinicalBERT	
	DP	EO	DP	EO
Ethnicity	0.0435	0.5000	0.1667	0.5000
Gender	0.0004	0.0028	0.0004	0.0087
Religion	0.0217	0.2000	0.0518	0.2000
Marital Status	0.0320	0.1231	0.0843	0.3077

Table 2: Fairness Metrics for Hospital Mortality Prediction

Attribute	BioGPT			ClinicalBERT		
	MAE	MPD	DR	MAE	MPD	DR
Ethnicity	11.92	5.97	15.97	12.12	6.64	17.68
Gender	0.20	0.14	1.11	0.25	0.27	1.10
Religion	5.27	6.47	3.05	5.90	2.68	3.17
Marital Status	2.34	0.78	2.18	3.40	0.98	3.18

Table 3: Fairness Metrics for Length of Stay Prediction

- Ethnicity and Marital Status exhibited the highest bias, especially in ClinicalBERT across all three tasks.
- Gender showed minimal bias consistently across models and metrics.
- BioGPT generally demonstrated lower disparities in demographic parity, prediction error, and outcome distribution than ClinicalBERT.
- Equalized Odds for ethnicity in mortality prediction remained high (0.5) across both models, highlighting significant variation in error rates across subgroups.

5.3.2 Bias Mitigation Results

Each model underwent three bias mitigation approaches: Counterfactual Data Augmentation (CDA), group-wise mean correction, and Fairlearn's Threshold Optimization. These methods significantly reduced fairness disparities while preserving predictive performance.

Threshold Optimization eliminated group disparities in mortality prediction, achieving perfect Demographic Parity (0.00) for Ethnicity and Religion, while Equalized Odds improved from 0.50 to 0.00 (Ethnicity) and 0.20 to 0.00 (Religion).

For Length of Stay, Group-wise Mean Correction reduced Mean Prediction Disparity across all demographics without affecting MAE, though some groups showed slightly increased Dispersion Ratio, revealing a fairness-variance tradeoff.

In treatment generation, CDA improved both Disparate Impact (+0.06 for Ethnicity) and Theil

Attribute	BioGPT		ClinicalBERT	
	DI	TI	DI	TI
Ethnicity	0.89	0.12	0.95	0.09
Gender	0.98	0.05	0.99	0.03
Religion	0.93	0.10	0.97	0.07
Marital Status	0.85	0.15	0.90	0.10

Table 4: Fairness Metrics for Treatment Generation

Index across all attributes, demonstrating more equitable outcome distributions. These results prove clinical LLMs can achieve fairness without sacrificing accuracy.

Attribute	Mortality (EO)	Length of Stay (MPD)	Treatment (DI)
Ethnicity	0.50 → 0.00	5.97 → 2.98	0.89 → 0.95
Religion	0.20 → 0.00	6.47 → 3.23	0.93 → 0.97
Marital Status	0.15 → 0.0056	0.78 → 0.39	0.85 → 0.92
Gender	≈0.00 → ≈0.00	0.14 → 0.05	0.98 → 0.99

Table 5: Fairness Metrics (BioGPT): Before → After Mitigation

Attribute	Mortality (EO)	Length of Stay (MPD)	Treatment (DI)
Ethnicity	0.50 → 0.00	6.64 → 3.32	0.95 → 0.98
Religion	0.11 → 0.00	2.68 → 1.33	0.97 → 0.99
Marital Status	0.11 → 0.035	0.98 → 0.49	0.90 → 0.95
Gender	0.01 → ≈0.00	0.27 → 0.13	0.99 → 1.00

Table 6: Fairness Metrics (ClinicalBERT): Before → After Mitigation

6 Conclusions and Future Work

6.1 Conclusions

We found quantifiable differences in model behavior across demographic groups, especially marital status and ethnicity.

When compared to ClinicalBERT, BioGPT continuously showed more fair behavior across the majority of metrics. However, there were issues with fairness in both models, particularly in Disparate Impact and Equalized Odds. On the other hand, both models handled gender equally, with little variation across all tasks.

Additionally, we used three fairness mitigation strategies to lessen disparities: Counterfactual Data Augmentation (CDA), group-wise mean correction, and threshold optimization. Each technique greatly increased fairness metrics without appreciably lowering predicted accuracy.

Overall, our results highlight how crucial it is to include mitigating techniques and fairness assessments when implementing LLMs in healthcare settings with significant stakes. Even high-performing models run the risk of escalating or maintaining current healthcare inequalities in the absence of such safeguards.

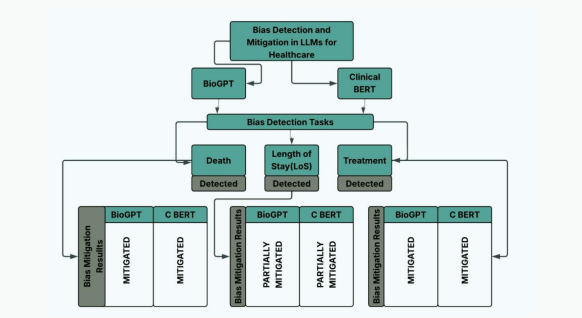

Figure 1: Summary of our experiments for detection and mitigation of bias with respect to demographic data

6.2 Future Work

Although our results show significant advancements in fairness-aware clinical NLP, there are still a number of areas that need further research:

- **Evaluation of Additional Domain-Specific LLMs:** Future work can explore more advanced or specialized biomedical LLMs such as BluBERT or Med-PaLM.
- **In-Processing Debiasing Methods:** Including fairness requirements directly during model training through regularization approaches, adversarial debiasing, or fairness-aware loss functions.
- **Pre-Processing Approaches:** Fairness can be enhanced before the model encounters the data by proactively lowering bias in the data using techniques like reweighting and resampling.

These approaches would improve the generalizability and reliability of LLMs in practical healthcare applications while also advancing the development of inclusive clinical NLP systems.

By tackling these issues, we may help create clinical decision-support systems that are not just strong and precise but also fair, open, and reliable.

7 Individual Contributions

- **Shreya Sethi** - Designed and implemented the hospital mortality prediction pipeline using ClinicalBERT, including fine-tuning and demographic subgroup evaluation. Conducted fairness assessments using Demographic Parity and Equalized Odds, revealing key bias patterns. Integrated ThresholdOptimizer for bias mitigation and evaluated its effectiveness across all sensitive attributes. Also contributed significantly to organizing the overall fairness evaluation framework, writing documentation, and visualizing model disparities pre- and post-mitigation.

- **Shireen Chand** - Led the treatment prediction task across both BioGPT and ClinicalBERT models. Designed and implemented a fairness evaluation strategy using Disparate Impact and Theil Index to assess inequality in treatment-related predictions. Applied Counterfactual Data Augmentation (CDA) to mitigate biases across sensitive attributes like ethnicity and religion. Additionally, coordinated comparison of post-mitigation performance across models and tasks, and contributed to synthesizing results for presentation and reporting.

- **Samarth Agarwal** - Developed the LOS prediction pipeline using ClinicalBERT, focusing on bias measurement and fairness-performance trade-off analysis. Computed key fairness metrics such as MPD, DR, and MAE to identify model bias. Implemented group-wise mean correction for mitigating disparities, and analyzed the effects on model output consistency. Took lead on comparing fairness outcomes between ClinicalBERT and BioGPT, and helped standardize experimental evaluation protocols across models.

- **Divij Gera** - Implemented the end-to-end pipeline for length of stay (LOS) prediction using BioGPT, handling data preprocessing, model training, and evaluation. Focused on analyzing model fairness using Mean Prediction Disparity (MPD), Disparate Representation (DR), and MAE across demographic subgroups. Applied group-wise mean correction for bias mitigation, demonstrating fairness improvements while preserving accuracy. Also contributed to generating interpretability plots and subgroup-level fairness visualizations to support analytical conclusions.

- **Arshia Mubias Shaik** - Developed the pipeline for fine-tuning BioGPT on the hospital mortality prediction task. Conducted extensive fairness evaluations using metrics such as Demographic Parity and Equalized Odds, identifying disparities across ethnicity, gender, and marital status. Played a key role in integrating post-processing bias mitigation through ThresholdOptimizer from Fairlearn, ensuring reduced disparities without compromising model performance. Additionally, contributed to literature review on bias mitigation strategies and their applicability to clinical LLMs.

References

Marshall H. Chin, Nasim Afsar-Manesh, Arlene S. Bierman, Christine Chang, Caleb J. Colón-Rodríguez, and et al. 2023. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. *JAMA Network Open*, 6(12):e2345050–e2345050.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Kareem A. Wahid. 2023. Algorithmic fairness in machine learning. Radiology: Artificial Intelligence Blog, RSNA.

Yuqing Wang, Yun Zhao, Sara Alessandra Keller, Anne de Hond, Marieke M. van Buchem, Malvika Pillai, and Tina Hernandez-Boussard. 2024. Unveiling and mitigating bias in mental health analysis with large language models. *arXiv preprint arXiv:2406.12033*. Version 2, submitted on June 19, 2024. License: CC BY-NC-SA 4.0.

Travis Zack, Eric Lehman, Mirac Suzgun, and et al. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: A model evaluation study. *The Lancet Digital Health*, 6(1).