Revealing Treatment Non-Adherence Bias in Clinical Machine Learning Using Large Language Models

Zhongyuan Liang ZHONGYUAN_LIANG@BERKELEY.EDU
UC Berkeley and UCSF

Arvind Suresh ARVIND.SURESH@UCSF.EDU
UCSF

Irene Y. Chen IYCHEN@BERKELEY.EDU
UC Berkeley and UCSF

Abstract

Machine learning systems trained on electronic health records (EHRs) increasingly guide treatment decisions, but their reliability depends on the critical assumption that patients follow the prescribed treatments recorded in EHRs. Using EHR data from 3,623 hypertension patients, we investigate how treatment non-adherence introduces implicit bias that can fundamentally distort both causal inference and predictive modeling. By extracting patient adherence information from clinical notes using a large language model (LLM), we identify 786 patients (21.7%) with medication non-adherence. We further uncover key demographic and clinical factors associated with non-adherence, as well as patient-reported reasons including side effects and difficulties obtaining refills. Our findings demonstrate that this implicit bias can not only reverse estimated treatment effects, but also degrade model performance by up to 5% while disproportionately affecting vulnerable populations by exacerbating disparities in decision outcomes and model error rates. This highlights the importance of accounting for treatment non-adherence in developing responsible and equitable clinical machine learning systems.

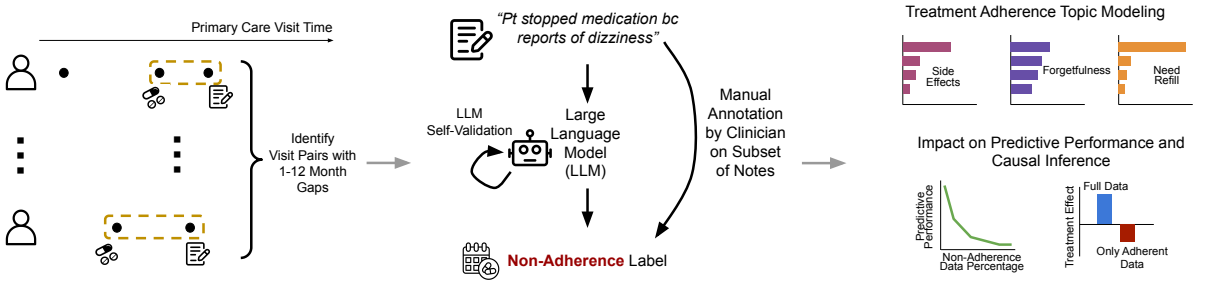
Data and Code Availability Our study utilizes EHR data from UCSF. Details on the data and preprocessing steps are provided in the following sections. While we will make the code publicly available, the data cannot be shared due to data use agreement.

Institutional Review Board (IRB) Our work was reviewed by an IRB and declared exempt as it focuses on retrospective analysis using de-identified data.

1. Introduction

Treatment non-adherence is a pervasive and persistent challenge in healthcare. Researchers estimate that poor medication adherence leads to 125,000 preventable deaths annually in the U.S. and contributes to $100-$300 billion in avoidable healthcare costs (Benjamin, 2012). This issue is particularly prevalent among patients with chronic conditions such as hypertension, with 40-50% failing to take their medications as prescribed (Kleinsinger, 2018; Algabbani and Algabbani, 2020). While researchers have extensively documented this problem through surveys and interviews (Boratas and Kilic, 2018; Fernandez-Lazaro et al., 2019; Algabbani and Algabbani, 2020; Najjuma et al., 2020; Schober et al., 2021), the studies—and ultimately understanding of treatment non-adherence—remain limited by small sample sizes and self-reporting bias (Adams et al., 1999; Stirratt et al., 2015). Physical solutions to monitor and encourage adherence such as electronic pill caps have shown promise in controlled settings but remain impractical for large-scale deployment due to high costs and implementation challenges (Parker et al., 2007; Mauro et al., 2019).

These measurement challenges take on new urgency as healthcare systems increasingly rely on machine learning (ML) models trained on electronic health records (EHRs) to guide treatment decisions (Komorowski et al., 2018; Brugnara et al., 2020; Zheng et al., 2021; Mroz et al., 2024; Yi et al., 2024; Shen et al., 2024; Chen et al., 2022). These machine learning models learn from historical patient data, which assume that prescribed treatments were actually taken. However, this introduces an implicit bias—models trained on non-adherent patients learn



(a) Cohort Creation of 3,623 Patients (b) LLM Extraction of Non-Adherence (c) Treatment Adherence Analysis

Figure 1: Illustration of cohort selection, LLM non-adherence extraction, and non-adherence analysis. (a) We select 3,623 hypertension patients and pair their visits, with hypertension medication prescribed at the first visit and clinical notes extracted from the second. (b) These notes are then processed by an LLM to identify treatment non-adherence, with outputs validated through clinician annotations. (c) We further perform topic modeling to uncover reasons for non-adherence and assess the harmful impact of ignoring this bias on predictive modeling performance and treatment effect estimation.

patterns that misrepresent true treatment effects. This implicit bias may degrade model performance and disproportionately impact underserved populations, who often face greater barriers to treatment adherence (Bosworth et al., 2006; Schober et al., 2021).

Recent advances in large language models (LLMs) have shown that LLMs can advance medical understanding by accurately extracting information from EHRs (Agrawal et al., 2022; Goel et al., 2023). Instead of relying on self-reported treatment adherence from questionnaires and interviews, LLMs could serve as a powerful tool for identifying treatment non-adherence directly from EHRs. By analyzing rich but unstructured clinical notes, LLMs can detect documented instances of missed medications, unfilled prescriptions, and patient-reported barriers to adherence, enabling systematic assessment of treatment non-adherence across large patient populations.

In this study, we examine hypertension treatment non-adherence using EHR data from UCSF by leveraging an LLM to analyze clinical notes, and further investigate its impact on causal inference and ML model performance (Figure 1). With a cohort of 3,623 patients, we identify 786 (21.7%) cases of non-adherence and extract demographic and clinical factors that are statistically significant. Additionally, we apply topic modeling to clinical notes revealing underlying reasons for non-adherence.

To assess the effect of treatment non-adherence bias on downstream model performance, we perform causal inference and build predictive models using EHR with treatment records. Our results show that ignoring treatment non-adherence bias could lead to reversed conclusions in treatment effect estimation, significantly degrade the performance of predictive models up to 5%, and lead to unfair predictions. Furthermore, we highlight the importance of addressing treatment non-adherence bias by showing simply removing patient records with non-adherence, though reducing the size of the training dataset, could improve model performance and lead to fairer predictions.

The contributions of this work include:

1. Conducting a large-scale study on treatment non-adherence in hypertension and identifying statistically significant factors associated with non-adherence.
2. Comparing LLM identification against physician annotations, LLMs perform well with 92% accuracy, precision and recall.
3. Identifying patient-reported reasons for treatment non-adherence including side effects, forgetfulness, difficulties obtaining refills, etc.
4. Demonstrating the harmful impact of ignoring treatment non-adherence bias on causal infer-

ence and predictive modeling, leading to poorer performance and exacerbating racial disparities.

2. Related Work

2.1. Treatment adherence analysis in hypertension

Multiple studies have investigated treatment adherence among patients with hypertension (Boratas and Kilic, 2018; Uchmanowicz et al., 2018; Algabbani and Algabbani, 2020; Najjuma et al., 2020; Schober et al., 2021). These studies are mainly cross-sectional, with a cohort of admitted patients collected at a fixed time point, and treatment adherence is typically measured through questionnaires and interviews. For instance, Algabbani and Algabbani (2020) conducted a study in Saudi Arabia involving 306 hypertensive outpatients, finding that only 42.2% of participants adhered to their antihypertensive medications. Boratas and Kilic (2018) conducted a similar study of 147 hypertensive patients, identifying factors such as age and duration of hypertension to be significant. However, due to their reliance on questionnaire and interview data, they often have small sample sizes (e.g., less than 300 patients) and self-reporting bias (Adams et al., 1999; Stirratt et al., 2015), which limits their representativeness and can even lead to contradictory conclusions. In contrast, our work conducts the first large-scale analysis utilizing EHR, with a significantly larger sample size of 3,623 patients.

2.2. Machine learning and treatment adherence

Machine learning has been used to identify individual risk factors associated with treatment non-adherence (Koesmahargyo et al., 2020; Gichuhi et al., 2023; Burgess-Hull et al., 2023). Gichuhi et al. (2023) developed ML algorithms and found SVM achieved 91.28% accuracy in predicting tuberculosis treatment non-adherence. Instead of predicting treatment adherence, our work focuses on analyzing the impact of treatment non-adherence bias on downstream model performance. Other studies have applied natural language processing (NLP) to analyze surveys to better understand treatment non-adherence (Anglin et al., 2021; Lin et al., 2022; Chan et al., 2024). Chan et al. (2024) applied NLP to free-text responses from questionnaires completed by type 2 diabetes patients, identifying key reasons for non-adherence. Unlike questionnaires, our work leverages treatment adherence information extracted from clinical notes using LLMs. Lastly, Zhong et al. (2022) applied ML while accounting for adherence information when analyzing treatment effects in a randomized controlled trial. To our knowledge, our study is the first to leverage LLMs for extracting treatment adherence information from clinical notes and evaluating its impact on downstream causal inference and predictive model performance.

3. Study Design

3.1. Hypertension cohort selection

We identified 15,002 patients with primary hypertension and extracted their primary care visits occurring on or after January 1st, 2019 following their initial hypertension diagnosis. To assess treatment adherence, consecutive visits for each patient were grouped into pairs. We focused on pairs where a hypertension medical prescription was provided during the first visit, and verified adherence at the second visit by extracting the associated clinical notes.

Our analysis focuses on ten commonly prescribed hypertension medications: amlodipine, losartan, lisinopril, benazepril, carvedilol, hydralazine, hydrochlorothiazide, clonidine, spironolactone, and metoprolol (hea, 2024a). Therefore, we excluded pairs in which the first visit lacked a medication record on this list, as well as pairs with missing or invalid notes during the second visit. We further focus on pairs where the interval between visits is between one month and one year. Lastly, we filtered out patients with unknown demographic information for the purpose of analysis. This resulted in a final cohort of 3,623 patients with 5,952 visit pairs. The cohort selection process is summarized in Appendix A.

Demographic information, including sex, age, race, and marital status, was extracted from patient records. Four clinical factors were further derived from the EHR, many of which have been shown to be associated with hypertension non-adherence (Boratas and Kilic, 2018; Algabbani and Algabbani, 2020). These factors include the duration between the two visits in the pair, the duration of hypertension, the number of primary care visits and the number of comorbidities. We quantified comorbidities using the Charlson Comorbidity Index (CCI) (Charlson et al., 1987) and the Elixhauser Comorbidity Index (ECI) (Elixhauser et al., 1998), which condensed diagnoses

into 17 and 31 well-defined comorbidity categories respectively. The demographic and clinical characteristics of the selected cohort are summarized in Table 1. We detail the comorbidity categories along with other features used in the study in Appendix B.

3.2. LLM configuration and prompt engineering

We used the GPT-4o model (OpenAI et al., 2024) (version 2024-05-13) via the HIPAA-compliant Microsoft Azure API, with the temperature set to 0 and all other parameters left at default. For each pair of visits, we provided the prescription record from the first visit and the clinical notes from the second visit to the model to assess adherence to the prescribed medication.

The model was prompted to identify instances of non-adherence, the type of non-adherence, and extract relevant sections from the notes. We used a zero-shot approach without additional training data or fine-tuning. We also implemented a second round of prompt validation by feeding the model's initial output back into the model, asking it to double-check its response. This additional step significantly reduced hallucinations. The prompt used in the study is provided in Appendix C.

The cost for running all GPT-4o evaluations, including prompt development and inference was $184.77, based on a cost of $0.005 per 1,000 input tokens and $0.015 per 1,000 output tokens.

3.3. Physician validation of LLM detection

To ensure the reliability of the LLM detection, we randomly selected 50 pairs labeled by the model as non-adherence and 50 pairs labeled as adherence for physician validation to assess accuracy. The gold standard was established through physician annotations conducted independently of the model's predictions. Overall, the model achieved an accuracy of 92%, with four instances of physician-labeled non-adherence not detected and four adherent instances mislabeled as non-adherence(92% precision and recall).

We further analyze discrepancies between the model and physician annotations, noting that some mismatches arise from ambiguous notes. For example, cases where patients restarted medication after hospitalization were marked as non-adherent by the LLM, since treatment was paused during hospitalization. Whereas physicians labeled them as adherent, considering the pause as a temporary interruption rather than true non-adherence.

4. Treatment Non-adherence Analysis

We begin by presenting the results of the identified hypertension treatment non-adherence with statistical testing in Section 4.1. In Section 4.2, we apply topic modeling to the extracted clinical notes, uncovering underlying reasons contributing to treatment non-adherence.

4.1. Factors associated with treatment non-adherence

To meet the independence assumption of the statistical tests, we keep only the most recent pair of visits for adherent patients and the most recent non-adherent pair for non-adherent patients when multiple pairs are available for the same patient.

Among 3,623 patients, 786 (21.7%) are identified as non-adherent to their treatment plans. Of these 786 patients, 506 (64.4%) miss their prescribed treatments, 237 (30.2%) take a different dosage than instructed, 53 (6.7%) use a different medication, and 62 (7.9%) take their medication at a time other than instructed. Note that a single patient may exhibit multiple types of non-adherence above.

To identify factors associated with nonadherence to treatment, we begin by performing unadjusted logistic regression, with the results including confidence intervals and p-values presented in Table 1. We find three factors that are statistically significant ($p < 0.05$): age ($p = 0.036$), one-hot encoding for Black race ($p = 0.023$), and the number of comorbidities ($p = 0.045$).

Our findings indicate that younger patients are less likely to adhere to treatment, aligning with previous research (Boratas and Kilic, 2018) that suggests adherence improves with age as patients become more accustomed to managing their diagnoses. Additionally, we find that Black patients exhibit higher rates of non-adherence. In the U.S., Black individuals have a higher prevalence of uncontrolled hypertension than White individuals (Aggarwal et al., 2021), and our finding further highlighting the need for greater attention to prevent further exacerbation of racial disparities in hypertension control. Lastly, a lower number of comorbidities is associated with a higher rate of non-adherence, possibly because patients with fewer

Table 1: Demographic and clinical characteristics of patients in the study and logistic regression results. Age, one-hot encoding for Black race, and the number of comorbidities are found to be statistically significant factors of treatment non-adherence.

Factors	**Total**	**Non-adherent**	**Adherent**	**Bivariate Analysis**		
	$n = 3623$	$n = 786(21.7\%)$	$n = 2837(78.3\%)$	**Unadjusted OR**	**95% CI**	***p*-value**
Demographics						
Sex						
Female	2143	473	1670	Ref.	Ref.	
Male	1480	313	1167	0.95	(0.81 to 1.11)	0.508
Age, mean ± SD	62.03 ± 14.2	61.09 ± 14.7	62.29 ± 14.1	0.94	(0.89 to 1.00)	**0.036**
Race						
Asian	1125	244	881	Ref.	Ref.	
Black	419	114	305	1.35	(1.04 to 1.75)	**0.023**
White	1646	327	1319	0.90	(0.74 to 1.08)	0.244
Other	433	101	332	1.10	(0.84 to 1.43)	0.486
Marital Status						
Divorced	329	69	260	Ref.	Ref.	
Married	1861	370	1491	0.94	(0.70 to 1.25)	0.649
Single	878	220	658	1.26	(0.93 to 1.71)	0.139
Widowed	358	77	281	1.03	(0.72 to 1.49)	0.864
Other	197	50	147	1.28	(0.85 to 1.94)	0.243
Clinical Factors						
Time between visits (days), mean ± SD	116.89 ± 83.4	112.04 ± 85.5	118.23 ± 82.8	1.00	(1.00 to 1.00)	0.066
Total number of comorbidities (ECI), mean ± SD	3.13 ± 2.4	2.98 ± 2.2	3.17 ± 2.4	0.96	(0.93 to 1.00)	**0.045**
Duration of hypertension (years), mean ± SD	5.94 ± 6.5	5.75 ± 6.6	6.00 ± 6.5	0.99	(0.98 to 1.01)	0.341
Number of primary care visits one year prior the visit, mean ± SD	15.75 ± 11.6	15.71 ± 10.8	15.76 ± 11.8	1.00	(0.99 to 1.01)	0.925

Table 2: Multivariate logistic regression results of factors associated with treatment non-adherence. The number of comorbidities and one-hot encoding for Black race remain statistically significant.

Factors	Multivariate Logistic Regression		
	Adjusted OR	**95% CI**	***p*-value**
Age (per 10-year increment)	0.97	(0.91 to 1.02)	0.242
Number of comorbidities	0.96	(0.93 to 1.00)	**0.03**
Race			
Asia	Ref.	Ref.	
Black	1.38	(1.06 to 1.80)	**0.016**
White	0.90	(0.75 to 1.08)	0.266
Other	1.10	(0.84 to 1.43)	0.5

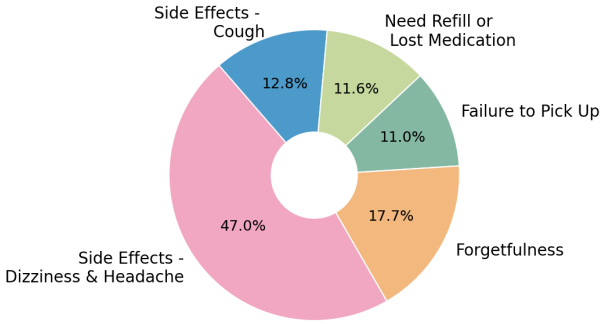


Figure 2: BERT topic modeling results for treatment non-adherence reasons. Side effects are the dominant reason for non-adherence, and 17.7% of reasons are due to forgetfulness, while others are related to not picking up the medication, needing a refill, or losing it.

health conditions may perceive their treatment as less essential.

To avoid potential confounding, we then adjust for the identified significant factors in the multivariate logistic regression model. Results are presented in Table 2, where we see that the number of comorbidities ($p = 0.03$) and race as Black ($p = 0.016$) still remain statistically significant.

4.2. Uncovering reasons for treatment non-adherence

We employ Bidirectional Encoder Representations from Transformers (BERT)-based topic modeling (Grootendorst, 2022) on the extracted non-adherent clinical notes to uncover reasons. This utilizes a BERT architecture (Devlin et al., 2019) to generate embeddings from the extracted notes and ap-

plies Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2020) to reduce the dimensionality of the embeddings. The reduced embeddings are subsequently clustered using HDBSCAN (Campello et al., 2013), and key terms for each cluster are identified using class-based Term Frequency-Inverse Document Frequency (c-TF-IDF).

Across all non-adherent instances, 164 clinical notes fall into clusters. The full clustering with key terms found is in Appendix D, and we summarize the topics in Figure 2. 59.8% of the extracted reasons for non-adherence are related to side effects, with most patients experiencing dizziness and headaches after taking the prescribed medication, while a few report coughing. 17.7% of the extracted reasons indicate non-adherence due to forgetfulness. Additionally, 11.6% are due to needing a refill or losing their medication and 11% of the reasons involve patients not picking up their medication.

5. Impact of Non-Adherence Bias on Downstream Performance

In this section, we evaluate the impact of treatment non-adherence bias on downstream model performance. We first define the outcome of interest and demonstrate the effect of non-adherence on the outcome in Section 5.1. We then analyze its impact on treatment effect estimation for causal inference in Section 5.2 and on ML model performance in Section 5.3.

5.1. Treatment non-adherence leads to worse blood pressure outcomes

Blood pressure reduction is the primary outcome when evaluating hypertension medications. To investigate the impact of treatment non-adherence on downstream performances, we begin by extracting pairs of visits where blood pressure measurements are available for both visits. To ensure a sufficient number of encounters for each medication, we focus on the top five most commonly prescribed medications: amlodipine, lisinopril, losartan, hydrochlorothiazide and metoprolol. We only include pairs where the duration between visits is less than six months to minimize the influence of other factors that could affect blood pressure over longer intervals. This results in 1732 pairs of encounters in total with 303 non-adherent pairs.

We begin by showcasing the impact of treatment non-adherence on blood pressure reduction between visits using t-tests. The results presented in Table 3 indicate that non-adherence leads to smaller blood pressure reduction, with 1.96 mmHg less systolic reduction ($p = 0.011$) and 3.93 mmHg less diastolic reduction ($p = 0.001$) compared to adherence.

Table 3: Results of the t-tests assessing the effect of treatment non-adherence on blood pressure reduction. Treatment non-adherence is statistically significant for both systolic and diastolic reduction, with non-adherence leading to smaller reductions in systolic and diastolic blood pressure.

Outcome	Mean Difference	95% CI	p-value
Systolic Reduction	-1.96	(-3.47 to -0.46)	**0.011**
Diastolic Reduction	-3.93	(-6.23 to -1.63)	**0.001**

5.2. Causal inference for treatment effect estimation

Amlodipine and Lisinopril are the most commonly prescribed medications for hypertension, leading to numerous randomized controlled trials (RCTs) comparing their treatment effects (Cappuccio et al., 1993; Naidu et al., 2000). However, due to the high cost of RCTs, various causal inference methods have been developed to estimate treatment effects from observational data (Pearl, 2009; Austin, 2011; Shalit et al., 2017; Künzel et al., 2019). Among them, Inverse Probability Weighting (IPW) provides an unbiased estimation of the Average Treatment Effect (ATE) by adjusting for confounding (Austin, 2011). Meta-learners such as the S-learner, T-learner and X-learner have been proposed to improve treatment effect estimation by leveraging flexible ML models to learn heterogeneous treatment effects (Künzel et al., 2019). We start by demonstrating the impact of treatment non-adherence bias on ATE estimation using IPW and meta-learners.

Experiment Setup. We compare the ATE estimation with and without including treatment non-adherent data. Demographic and clinical factors are included as confounders and detailed in Appendix B. Patients prescribed lisinopril act as the control group, while those taking amlodipine are considered the treated group. The treatment effect is assessed based on the reduction in diastolic and systolic blood pressure between two visits.

Table 4: Estimated ATE of medication on blood pressure reduction using different causal inference methods. Notably, excluding non-adherent data leads to a lower estimated effect on the diastolic blood pressure reduction and reverses the conclusion on the treatment effect of systolic blood pressure reduction consistently across methods.

	Diastolic Reduction (mmHg)				**Systolic Reduction (mmHg)**			
	IPW	S-Learner	T-Learner	X-Learner	IPW	S-Learner	T-Learner	X-Learner
Full Dataset	1.75	0.77	1.44	1.51	-0.06	-0.05	-0.12	-0.14
Adherent Data Only	1.40	0.57	0.97	0.92	0.11	0.08	0.06	0.07

Results. The results are presented in Table 4. Across all models, excluding non-adherent data leads to a lower estimated effect on the diastolic blood pressure reduction and a reversal in the systolic blood pressure reduction. Specifically when using IPW, without filtering for non-adherent data, amlodipine lowers diastolic blood pressure by 1.75 mmHg but increases systolic blood pressure by 0.06 mm Hg compared to lisinopril. After excluding non-adherent data, amlodipine lowers diastolic blood pressure by 1.40 mmHg and also reduces systolic blood pressure by 0.11 mmHg compared to lisinopril. This result shows a reversal in the estimated treatment effect for systolic blood pressure reduction before and after excluding non-adherent data. The same reversal trend also holds across different meta-learners (e.g., using the X-learner on the full dataset, amlodipine reduces systolic blood pressure by 0.14 mmHg compared to lisinopril, whereas after excluding non-adherent data, the X-learner estimates an increase of 0.07 mmHg in systolic blood pressure reduction). The results demonstrate that non-adherence bias can lead to systematic distortions in treatment effect estimation.

5.3. Supervised learning for treatment outcome prediction

We now demonstrate the impact of treatment non-adherence bias on predictive modeling performance. Following a common setup in the literature (Mroz et al., 2024; Yi et al., 2024), we use patients' EHR data with treatment prescriptions and blood pressure measurements from their first visit as covariates. The target to predict is whether the blood pressure will be normal at their second visit. Following the guidelines of the American Heart Association (hea, 2024b), we define normal blood pressure as having a systolic value of less than 120 and a diastolic value of less than 80. To evaluate model performance in predicting outcomes, we use 500 adherent samples as the test set in all subsequent experiments. We test exclusively on adherent patients since our goal is to evaluate model performance in scenarios where treatments are followed as prescribed, representing the intended clinical use case. A detailed description of the features used is provided in Appendix B.

5.3.1. Effect of varying treatment non-adherence data ratios on model performance and fairness

We begin by showing the harmful impact of treatment non-adherence bias by varying the proportion of non-adherent data in the training set.

Experiment Setup. We fix the training set size at 300 and vary the proportion of non-adherent data in the training set across $\{0\%, 10\%, 30\%, 50\%, 70\%, 90\%\}$ to evaluate the impact of treatment non-adherence bias on model performance. We train logistic regression and random forest models, both commonly used for modeling tabular EHR data and assess performance using AUROC on the test set. Beyond performance, ensuring fair decision-making is also a critical consideration in healthcare. Let A denote the sensitive attribute (e.g., race), Y represent the true outcome, and $\hat{Y}$ denote the predicted outcome. Demographic parity (Dwork et al., 2012) difference measures the disparity in the likelihood of receiving a positive prediction between groups, i.e.,

$$|P(\hat{Y} = 1 \mid A = 1) - P(\hat{Y} = 1 \mid A = 0)|$$

Equal odds (Hardt et al., 2016) difference compares both true positive rates and false positive rates across groups, i.e.,

$$|P(\hat{Y} = 1 \mid Y = 1, A = 1) - P(\hat{Y} = 1 \mid Y = 1, A = 0)|$$
$$|P(\hat{Y} = 1 \mid Y = 0, A = 1) - P(\hat{Y} = 1 \mid Y = 0, A = 0)|$$

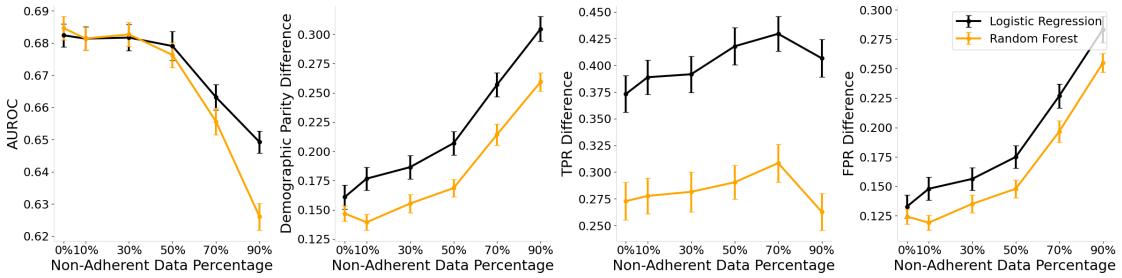


Figure 3: Results of varying treatment non-adherence data percentage on model performance and fairness. Increasing the proportion of non-adherent data in the training set degrades predictive performance and increases fairness disparities between Black and non-Black patients, as measured by demographic parity and the equal odds criterion (true positive rate and false positive rate differences). Results are averaged over 100 seeds, with error bars representing the standard error of the mean.

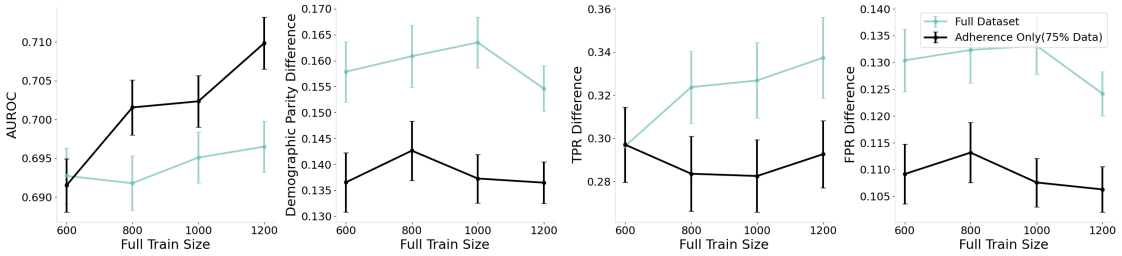


Figure 4: Results of removing non-adherent data on model performance and fairness. The black curve represents training on 75% of the full dataset, which only consists of adherent encounters. Removing non-adherent data improves model performance, with greater gains observed as sample size increases. It also decreases fairness disparities between Black and non-Black patients, as measured by demographic parity and the equal odds criterion (true positive rate and false positive rate differences). Results are averaged over 100 seeds, with error bars representing the standard error of the mean.

We therefore measure the differences in demographic parity and equal odds across racial groups to assess fairness.

Results. We present the results in Figure 3, showing that increasing the percentage of non-adherent data in the training set degrades performance, with a 3% drop for logistic regression and a 5% drop for random forest in AUROC. Additionally, a higher proportion of non-adherent data increases fairness disparities between Black and non-Black patients under both the demographic parity and equal odds criteria. For instance, the false positive rate disparity of the random forest doubles, increasing from 0.125 to 0.25 as the percentage of non-adherent data increases. Similar trends are observed for other racial groups, and we provide full results in Appendix E. These findings consistently highlight the harmful impact of treatment non-adherence bias.

5.3.2. Effect of removing non-adherent data on model performance and fairness

We now emphasize the importance of addressing treatment non-adherence bias by showing that a simple approach to remove non-adherent data can improve predictive performance and lead to fairer predictions.

Experiment setup. We fix the non-adherent data ratio at 25% and compare the performance of random forests trained on the entire dataset versus those trained only on adherent data by removing a

quarter of data that are non-adherent. We report the test AUROC as well as demographic and equal odds differences while varying the full training set size in $\{600, 800, 1000, 1200\}$ before removing non-adherent data.

Results. The results are presented in Figure 4. While the traditional ML perspective suggests that more data generally improves performance, our findings show that using only the adherent 75% of the data leads to better model performance, with the improvement becoming more significant as the sample size increases. For instance, with a training size of 1,200, the model achieves an AUROC of 0.695 when using all data, whereas dropping non-adherent data improves AUROC to 0.71. Additionally, we find removing non-adherent data reduces racial disparities between Black and non-Black patients, as both demographic parity and equal odds differences are consistently smaller across sample sizes. Similar trends are observed for other racial groups, and we provide full results in Appendix E. These findings further highlight the importance of addressing treatment non-adherence bias to achieve better and fairer model performance.

6. Discussion

Treatment non-adherence is a crucial factor in building treatment models but is often overlooked in practice. By leveraging LLMs with clinical notes, we identify non-adherent encounters with hypertension patients and further demonstrate how treatment non-adherence biases can degrade downstream model performance while exacerbating fairness gaps.

While our study focuses on hypertension, the same pipeline can be applied to other disease areas to analyze treatment non-adherence patterns. Beyond adherence, future work can also utilize LLMs to extract insights from clinical notes on other factors such as medication tolerance, side effects, social determinants of health, and patient-provider communications (Guevara et al., 2024; Robitschek et al., 2024; Zink et al., 2024; Antoniak et al., 2024; Miao et al., 2024). Our results show that removing non-adherent data from the training set improves both model performance and fairness. Instead of excluding non-adherent data entirely which could lead to insufficient samples, future research could also explore strategies to better integrate non-adherent data into modeling or develop models that are more robust to treatment adherence biases. Additionally, understanding the degree of non-adherence remains a significant challenge due to the absence of standardized quantitative metrics in clinical notes. Future work could explore approaches to estimate non-adherence severity, which may offer deeper insights into non-adherence patterns and enhance more effective strategies for downstream modeling.

Although leveraging LLMs with clinical notes enables large-scale analysis of treatment non-adherence, our study holds keys limitations. First, our work relies on the premise that non-adherence is explicitly documented in the notes, which means cases not mentioned may be missed, potentially leading to an underestimation of the true non-adherence rate. Furthermore, while physician validation confirms that the LLM's output is largely accurate, the use of ML models such as LLMs in shifting and censored patient populations may yield changing performances such that the automated extraction should be scrutinized (Pollard et al., 2019; Yuan et al., 2023; Finlayson et al., 2021; Chen et al., 2022). In conclusion, our work demonstrates the impact of treatment non-adherence bias in predictive modeling and causal inference through a real-world study on hypertension medications. We hope this study raises awareness of treatment non-adherence bias for future research on clinical machine learning.

References

Types of Blood Pressure Medications — heart.org. https://www.heart.org/en/health-topics/high-blood-pressure/changes-you-can-make-to-manage-high-blood-pressure/types-of-blood-pressure-medications, 2024a. [Accessed 06-02-2025].

Understanding Blood Pressure Readings — heart.org. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings, 2024b. [Accessed 06-02-2025].

Alyce S Adams, Stephen B Soumerai, Jonathan Lomas, and Dennis Ross-Degnan. Evidence of self-report bias in assessing adherence to guidelines. *International Journal for Quality in Health Care*, 11 (3):187–192, 1999.

Rahul Aggarwal, Nicholas Chiu, Rishi K Wadhera, Andrew E Moran, Inbar Raber, Changyu Shen,

Robert W Yeh, and Dhruv S Kazi. Racial/ethnic disparities in hypertension prevalence, awareness, treatment, and control in the united states, 2013 to 2018. *Hypertension*, 78(6):1719–1726, 2021.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*, 2022.

Fahad M Algabbani and Aljoharah M Algabbani. Treatment adherence among patients with hypertension: findings from a cross-sectional study. *Clinical hypertension*, 26:1–9, 2020.

Kylie L Anglin, Vivian C Wong, and Arielle Boguslav. A natural language processing approach to measuring treatment adherence and consistency using semantic similarity. *AERA Open*, 7: 23328584211028615, 2021.

Maria Antoniak, Aakanksha Naik, Carla S Alvarado, Lucy Lu Wang, and Irene Y Chen. Nlp for maternal healthcare: Perspectives and guiding principles in the age of llms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1446–1463, 2024.

Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

Regina M Benjamin. Medication adherence: helping patients take their medicines as directed. *Public health reports*, 127(1):2–3, 2012.

Selma Boratas and Hulya Firat Kilic. Evaluation of medication adherence in hypertensive patients and influential factors. *Pakistan Journal of Medical Sciences*, 34(4):959, 2018.

Hayden B Bosworth, Tara Dudley, Maren K Olsen, Corrine I Voils, Benjamin Powers, Mary K Goldstein, and Eugene Z Oddone. Racial differences in blood pressure control: potential explanatory factors. *The American journal of medicine*, 119(1): 70–e9, 2006.

Gianluca Brugnara, Ulf Neuberger, Mustafa A Mahmutoglu, Martha Foltyn, Christian Herweh, Simon Nagel, Silvia Schönenberger, Sabine Heiland, Christian Ulfert, Peter Arthur Ringleb, et al. Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning. *Stroke*, 51(12):3541–3551, 2020.

Albert J Burgess-Hull, Caleb Brooks, David H Epstein, Devang Gandhi, and Enrique Oviedo. Using machine learning to predict treatment adherence in patients on medication for opioid use disorder. *Journal of Addiction Medicine*, 17(1):28–34, 2023.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

Francesco P Cappuccio, Nirmala D Markandu, Donald RJ Singer, and Graham A MacGregor. Amlodipine and lisinopril in combination for the treatment of essential hypertension: efficacy and predictors of response. *Journal of hypertension*, 11(8): 839–847, 1993.

Jill Cates. GitHub - topspinj/medcodes: A Python package for standardizing medical data — github.com. https://github.com/topspinj/medcodes, 2019. [Accessed 06-02-2025].

Juliana CN Chan, Jean Claude Mbanya, Jean-Marc Chantelot, Marina Shestakova, Ambady Ramachandran, Hasan Ilkova, Lucille Deplante, Melissa Rollot, Lydie Melas-Melt, Juan Jose Gagliardino, et al. Patient-reported outcomes and treatment adherence in type 2 diabetes using natural language processing: Wave 8 of the observational international diabetes management practices study. *Journal of Diabetes Investigation*, 2024.

Mary E Charlson, Peter Pompei, Kathy L Ales, and C Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383, 1987.

Irene Y Chen, Rahul G Krishnan, and David Sontag. Clustering interval-censored time-series for disease phenotyping. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6211–6221, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep

bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

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*, pages 214–226, 2012.

Anne Elixhauser, Claudia Steiner, D Robert Harris, and Rosanna M Coffey. Comorbidity measures for use with administrative data. *Medical care*, 36(1): 8–27, 1998.

Cesar I Fernandez-Lazaro, Juan M García-González, David P Adams, Diego Fernandez-Lazaro, Juan Mielgo-Ayuso, Alberto Caballero-Garcia, Francisca Moreno Racionero, Alfredo Córdova, and Jose A Miron-Canelo. Adherence to treatment and related factors among patients with chronic conditions in primary care: a cross-sectional study. *BMC family practice*, 20:1–12, 2019.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385 (3):283–286, 2021.

Haron W Gichuhi, Mark Magumba, Manish Kumar, and Roy William Mayega. A machine learning approach to explore individual risk factors for tuberculosis treatment non-adherence in mukono district. *PLOS Global Public Health*, 3(7):e0001466, 2023.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. Llms accelerate annotation for medical information extraction, 2023. URL https://arxiv.org/abs/2312.02296.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022. URL https://arxiv.org/abs/2203.05794.

Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, et al. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, 7 (1):6, 2024.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Fred Kleinsinger. The unmet challenge of medication nonadherence. *The Permanente Journal*, 22, 2018.

Vidya Koesmahargyo, Anzar Abbas, Li Zhang, Lei Guan, Shaolei Feng, Vijay Yadav, and Isaac R Galatzer-Levy. Accuracy of machine learning-based prediction of medication adherence in clinical research. *Psychiatry research*, 294:113558, 2020.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, February 2019. ISSN 1091-6490. doi: 10.1073/pnas.1804597116. URL http://dx.doi.org/10.1073/pnas.1804597116.

Wei-Chun Lin, Jimmy S Chen, Joel Kaluzny, Aiyin Chen, Michael F Chiang, and Michelle R Hribar. Extraction of active medications and adherence using natural language processing for glaucoma patients. In *AMIA Annual Symposium Proceedings*, volume 2021, page 773, 2022.

Joseph Mauro, Kelly B Mathews, and Eric S Sredzinski. Effect of a smart pill bottle and pharmacist intervention on medication adherence in patients with multiple myeloma new to lenalidomide therapy. *Journal of Managed Care & Specialty Pharmacy*, 25(11):1244–1254, 2019.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

Brenda Y Miao, Christopher YK Williams, Ebenezer Chinedu-Eneh, Travis Zack, Emily Alsentzer, Atul J Butte, and Irene Y Chen. Identifying reasons for contraceptive switching from real-world data using large language models. *arXiv preprint arXiv:2402.03597*, 2024.

Thomas Mroz, Michael Griffin, Richard Cartabuke, Luke Laffin, Giavanna Russo-Alvarez, George Thomas, Nicholas Smedira, Thad Meese, Michael Shost, and Ghaith Habboub. Predicting hypertension control using machine learning. *Plos one*, 19 (3):e0299932, 2024.

MUR Naidu, PR Usha, T Ramesh Kumar Rao, and JC Shobha. Evaluation of amlodipine, lisinopril, and a combination in the treatment of essential hypertension. *Postgraduate Medical Journal*, 76(896): 350–353, 2000.

Josephine Nambi Najjuma, Laura Brennaman, Rose C Nabirye, Frank Ssedyabane, Samuel Maling, Francis Bajunirwe, and Rose Muhindo. Adherence to antihypertensive medication: an interview analysis of southwest ugandan patients' perspectives. *Annals of Global Health*, 86(1), 2020.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, et al. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Catherine S Parker, Zhen Chen, Maureen Price, Robert Gross, Joshua P Metlay, Jason D Christie, Colleen M Brensinger, Craig W Newcomb, Frederick F Samaha, and Stephen E Kimmel. Adherence to warfarin assessed by electronic pill caps, clinician assessment, and patient reports: results from the in-range study. *Journal of general internal medicine*, 22:1254–1259, 2007.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Tom J Pollard, Irene Chen, Jenna Wiens, Steven Horng, Danny Wong, Marzyeh Ghassemi, Heather Mattie, Emily Lindemer, and Trishan Panch. Turning the crank for machine learning: ease, at what expense? *The Lancet Digital Health*, 1(5):e198–e199, 2019.

Emily Robitschek, Asal Bastani, Kathryn Horwath, Savyon Sordean, Mark J Pletcher, Jennifer C Lai, Sergio Galletta, Elliott Ash, Jin Ge, and Irene Y Chen. A large language model-based approach to quantifying the effects of social determinants in liver transplant decisions. *arXiv preprint arXiv:2412.07924*, 2024.

Daniel J Schober, Moranda Tate, Denise Rodriguez, Todd M Ruppar, Joselyn Williams, and Elizabeth Lynch. High blood pressure medication adherence among urban, african americans in the midwest united states. *Journal of racial and ethnic health disparities*, 8(3):607–617, 2021.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms, 2017. URL https://arxiv.org/abs/1606.03976.

Judy Hanwen Shen, Inioluwa Deborah Raji, and Irene Y Chen. The data addition dilemma. In *Machine Learning for Healthcare Conference*. PMLR, 2024.

Michael J Stirratt, Jacqueline Dunbar-Jacob, Heidi M Crane, Jane M Simoni, Susan Czajkowski, Marisa E Hilliard, James E Aikens, Christine M Hunter, Dawn I Velligan, Kristen Huntley, et al. Self-report measures of medication adherence behavior: recommendations on optimal use. *Translational behavioral medicine*, 5(4): 470–482, 2015.

Bartosz Uchmanowicz, Anna Chudiak, Izabella Uchmanowicz, Joanna Rosińczuk, and Erika Sivarajan Froelicher. Factors influencing adherence to treatment in older adults with hypertension. *Clinical interventions in aging*, pages 2425–2441, 2018.

Jiayi Yi, Lili Wang, Jiali Song, Yanchen Liu, Jiamin Liu, Haibo Zhang, Jiapeng Lu, and Xin Zheng. Development of a machine learning-based model for predicting individual responses to antihypertensive treatments. *Nutrition, Metabolism and Cardiovascular Diseases*, 2024.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023.

Hua Zheng, Ilya O Ryzhov, Wei Xie, and Judy Zhong. Personalized multimorbidity management for patients with type 2 diabetes using reinforcement learning of electronic health records. *Drugs*, 81: 471–482, 2021.

Yongqi Zhong, Maria M Brooks, Edward H Kennedy, Lisa M Bodnar, and Ashley I Naimi. Use of machine learning to estimate the per-protocol

effect of low-dose aspirin on pregnancy outcomes: a secondary analysis of a randomized clinical trial. *JAMA network open*, 5(3):e2143414–e2143414, 2022.

Anna Zink, Hongzhou Luan, and Irene Y Chen. Access to care improves ehr reliability and clinical risk prediction model performance. *arXiv preprint arXiv:2412.07712*, 2024.

Appendix A. Cohort Creation

We detail the cohort selection in Figure 5.

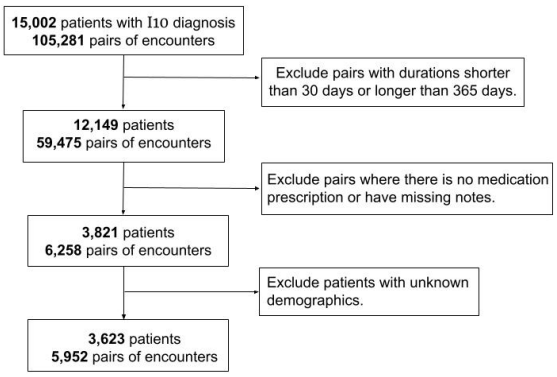


Figure 5: Cohort Selection

Appendix B. Feature Definitions

We use the following features throughout the study. The time between visits is excluded from causal inference and predictive modeling in Section 5, as we instead restrict the analysis to a smaller time window. ECI and CCI are derived using medcodes package (Cates, 2019).

Feature Name	Description
Sex	The sex of the patients.
Age	The age of the patients at the time of the visit.
Race	The race of the patients.
Marital Status	The marital status of the patients at the time of the visit.
Time Between Visits	The time in days between the two visits in each pair.
Charlson Comorbidity Index (CCI)	A measure of comorbidity based on the following conditions: myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, rheumatic disease, peptic ulcer disease, mild liver disease, diabetes without chronic complications, diabetes with chronic complications, hemiplegia/paraplegia, renal disease, any malignancy, moderate/severe liver disease, metastatic solid tumor, and AIDS/HIV.
Elixhauser Comorbidity Index (ECI)	A measure of comorbidity based on the following conditions: cardiac arrhythmias, congestive heart failure, valvular disease, pulmonary circulation disorders, peripheral vascular disorders, hypertension (uncomplicated or complicated), paralysis, other neurological disorders, chronic pulmonary disease, diabetes (uncomplicated or complicated), hypothyroidism, renal failure, liver disease, peptic ulcer disease, AIDS/HIV, lymphoma, metastatic cancer, solid tumor without metastasis, rheumatoid arthritis, coagulopathy, obesity, weight loss, fluid and electrolyte disorders, blood loss anemia, deficiency anemia, alcohol abuse, drug abuse, psychoses, and depression.
Duration of Hypertension	The duration (in years) between the onset of hypertension and the time of the visit.
Number of Primary Care Visits	The number of primary care visits one year prior to the visit.

Table 5: Descriptions of features included in the study

Appendix C. Prompt Description

The prompt used in the study is given in Figure 6.

You will be provided with clinical notes and a medication record for a patient. Your task is to identify treatment adherence problems specifically for the treatments listed in the provided medication record. Treatment adherence problems can include any of the following:

1. Not taking medication as prescribed: The patient is not following the prescribed regimen in the medication record, including missing medication.
2. Dosage discrepancy: The dosage that the patient is taking as mentioned in the clinical notes differs from the instruction in the medication record.
3. Alternative medication use: The patient is using an alternative medication to replace one listed in the medication record.
4. Taking medication at the wrong time: The patient does not adhere to the recommended timing, such as before meals or at bedtime.

Inputs:

- Medication record: *medication_record*
- Clinical notes: *note_text*
- Patient key: *patient_key*
- Deidentified note key: *deid_note_key*
- Date: *note_date*
- Notes: *note_record*

Output:

Return a JSON object representing a row in a dataframe with the following structure (no additional text):

{
patient_key: The provided patient key,
deid_note_key: The provided deidentified note key,
current_medication_record: The provided medication record,
treatment_mentioned: true/false, // Whether a treatment from the medication record is mentioned in the notes
treatment_adherence_presented: true/false, // Whether a treatment adherence problem was identified for the medications in the record
missing_treatment_presented: true/false, // Whether the patient is not taking a listed medication as prescribed (including missing medication)
dosage_discrepancy_presented: true/false, // Whether the dosage that the patient is taking as mentioned in the clinical notes differs from the instruction in the medication record
alternative_medication_presented: true/false, // Whether the patient is taking an alternative medication for a listed treatment
wrong_time_presented: true/false, // Whether the patient is taking medication at the wrong time
notes_mention_adherence: Include all parts of the clinical notes that identified the adherence issue, if applicable,
adherence_reason_notes: Include parts of the clinical notes mentioning reasons for the identified adherence issues, if applicable,
adherence_reason_summary: Summarize the reasons for adherence issues in a few words, if applicable
}

Special Instructions:

- If no treatment adherence problems are found for the listed medications, set *treatment_adherence_presented* to false and all other discrepancy-related fields to false or empty as appropriate.
- Only evaluate adherence issues for medications listed in the provided medication record, ignoring references to medications or treatments not listed in the medication record.
- Do not include adherence problems or treatment changes discussed during the visit; focus only on past adherence issues.
- Avoid flagging vague or implied adherence issues. Only flag adherence problems explicitly stated in the clinical notes.
- The dosage discrepancy should be determined exclusively based on the instruction information provided in the medication record. Do not infer discrepancies using medication names alone.
- If instruction information is nan in the medication record, do not consider dosage discrepancies even if dosage details are mentioned in the notes, unless an explicit adherence issue is stated in the notes.
- Ensure all fields accurately reflect the provided inputs and satisfy the special instructions.

Figure 6: Prompt used in the study

Appendix D. Topic Modeling

Here we show the full BERT topic modeling results with key terms found in clusters and we summarize the topics in Figure 2.

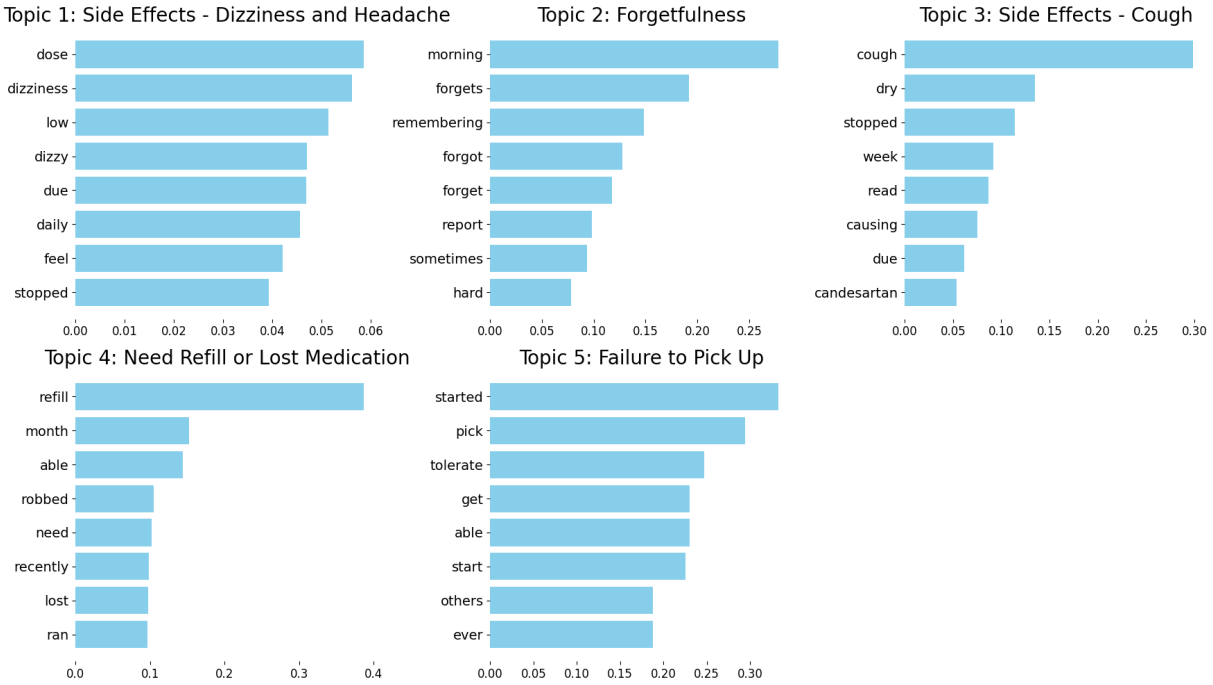


Figure 7: BERT topic modeling identified five clusters of treatment non-adherence reasons. The most common words in each cluster are highlighted in the plot. Key topics identified include side effects, forgetfulness, failure to pick up medications, need for refills, and lost medications. When applying BERT topic modeling, we set a minimum cluster size of 15 notes and used UMAP with 5 components and 15 neighbors for dimensionality reduction.

Appendix E. Additional Results of Non-adherence Bias on Predictive Modeling

Here, we present the complete results on racial disparities for the experiments in Section 5.3.

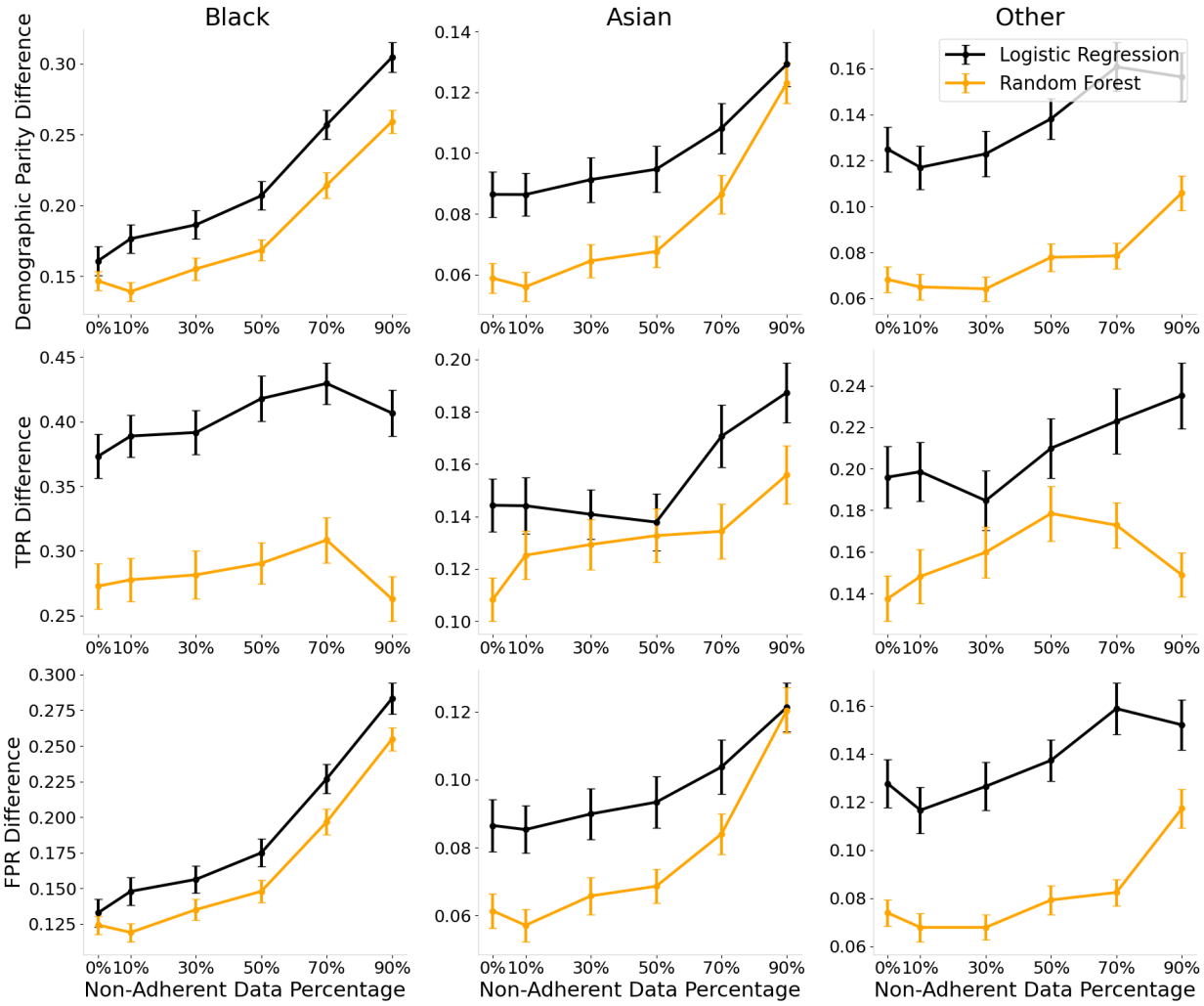


Figure 8: Increasing the proportion of treatment non-adherent data in the training set increases the fairness disparity between different races as measured by demographic parity and the equal odds criterion. Results are averaged over 100 seeds, varying the sampling of the train and test sets. Error bars represent the standard error of the mean.

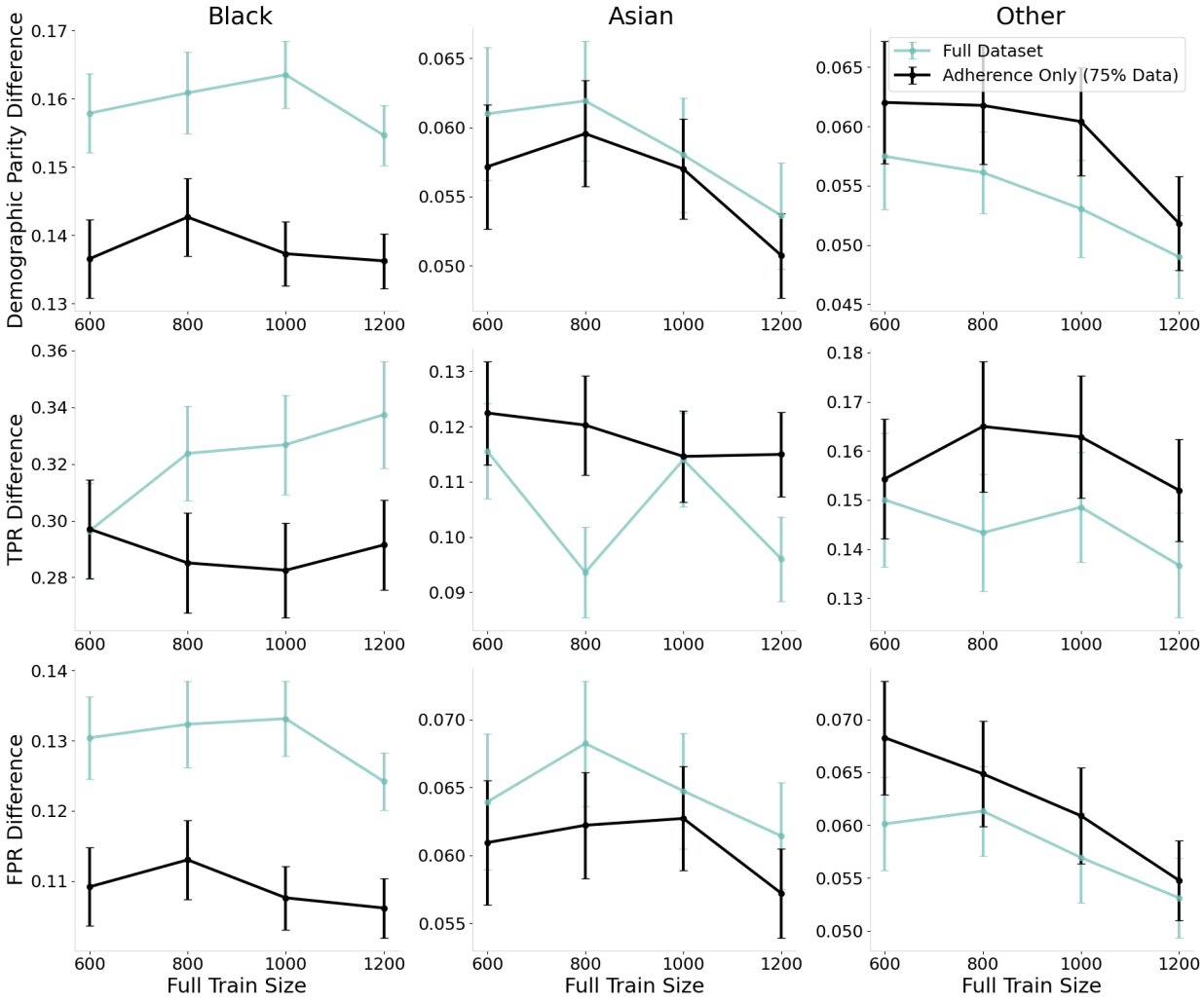


Figure 9: The black curve represents training on 75% of the full dataset consisting of adherent encounters only. Removing treatment non-adherent data from the training set decreases fairness disparities as measured by demographic parity and the equal odds criterion particularly for Black and Asian. Results are averaged over 100 seeds, varying the sampling of the train and test sets. Error bars represent the standard error of the mean.