AI/ML Trustworthiness for Medical Predictions – Experimental Results

Mohimenul Karim
Virginia Tech
Dept. of Computer Science
mohimenul@vt.edu

Tanmoy Sarkar Pias
Virginia Tech
Dept. of Computer Science
tanmoysarkar@vt.edu

Bimal Viswanath
Virginia Tech
Dept. of Computer Science
vbimal@vt.edu

Danfeng (Daphne) Yao*
Virginia Tech
Dept. of Computer Science
danfeng@vt.edu
**Corresponding author*

***Abstract*—It is important to assess the fairness of machine learning models and understand the impact of demographic information on model performance before deploying them in clinical settings. In this study, we propose a new differential impact testing method, DiffImp, to evaluate how the demographic information of a patient impacts the prediction outcome of medical machine learning models, including large language models. We apply DiffImp to evaluate the impact of demographic information, such as gender and race, on BERT, Transformer, and LSTM models in MIMIC III and eICU data sets. We find that the model's balanced accuracy can fluctuate between 0.65 and 0.75 across models and datasets due to demographic information. Our study also reveals that gender and race information can impact the outcome at the subgroup level. For example, for the MIMIC III dataset, the recall score for death cases among Asian patients decreases by up to 41.9% after including race information in the model. For some subpopulation groups, their demographic information prompts models to alter the mortality prediction result, against the statistical trend observed in the training data. We also discuss the clinical implications of our findings.**

1. Introduction

Machine learning (ML) has shown promise in the field of healthcare in predicting various diseases and mortality risks. Predicting mortality from electronic health record (EHR) data in advance can facilitate clinicians in identifying high-risk patients and taking preventive measures. Several studies applying ML models have been done in predicting mortality risks, and other life-threatening scenarios [1], [2], [3], [4], [5]. With the advent of large language models (LLMs), there has been a remarkable transformation in various natural language processing (NLP) tasks, including the medical field, such as clinical text summarization [6], [7], medical question answering [8], [9], etc. However, it has been shown that there can be performance disparity within coarse race groups in clinical machine learning [10]. Prediction algorithms in healthcare may be affected by race, and at a given risk score, Black patients were found to be sicker than White patients [11]. In dermatology AI, there can be performance differences between dark and light skin tones [12]. In addition, chest X-ray classifiers underdiagnose underserved subpopulations such as Hispanic female patients at a higher rate [13].

On the other hand, LLMs are not fully capable of accurately diagnosing patients across all pathologies and do not follow diagnostic or treatment guidelines [14]. A recent study showed that LLMs can propagate harmful, race-based conceptions [15]. These may pose a serious risk to patients. Demography can cause medical errors, and prediction models can exhibit performance disparity for some underrepresented demographic groups [16], [17], [18], [19]. However, [20] indicated that demographic information can have a two-sided effect on the predictive performance of AI healthcare models. Interestingly, a recent study concluded that simply omitting demographics may decrease prediction accuracy [21]. However, most other studies are not conclusive enough to support either the omission or use of demographic data in clinical decision-making [22], [23], [24]. As a result, till now, this is an open question to the researcher in the field and requires more rigorous investigation. In accordance with this line of research, our paper presents a structured framework to evaluate whether demographic information, such as, gender, race, should be included in LLM and ML models for mortality prediction in time-series data.

Although predictive models are used in U.S. hospitals, only 44% of the models were evaluated for bias [25]. Although [26], [27] suggested that high-quality and large training data are essential to avoid potential bias in AI models, an unfair utilization of demographic features can lead to higher prediction performance [28]. Furthermore, the impact of demographic variables on LLMs in mortality prediction using time series data has not been explored. When performing the mortality risk prediction task, we need to quantitatively assess how the LLMs and ML models' decisions are influenced by demographic variables. Such an influence can lead to wrong predictions and eventually to serious consequences. Attackers may exploit the bias of the models to generate specific predictions. They can change demographic information in the data to manipulate the response, which is a major threat to human health.

In this work, we propose a differential impact testing method for mortality risk prediction algorithms and assess their fairness in terms of different demographic variables such as gender and race. With such a differential testing method, we can answer these problems: 1) *Does demo-*

graphic information change models' mortality risk prediction? 2) *To what extent can a model's predictions deviate from those of a demographic-unaware model, including at the subgroup level?* 3) *Do demographic variables impact a pretrained LLM differently than conventional neural network models when trained on a single dataset?* Our differential impact testing method is called **DiffImp**. We apply our method to analyze models for the in-hospital mortality prediction task. In addition to standard ML models, we assessed the fairness of LLM in mortality prediction using time series data through the lens of demographic variables, which, to our knowledge, has not yet been explored. Our contributions are summarized as follows.

- We present a differential impact testing method, DiffImp, to evaluate the impact of demographic information on medical prediction models, including LLM. DiffImp can be used as a framework to assess the fairness of machine learning and large language models in healthcare. Our tests assess how demographic variables affect prediction accuracy when incorporated individually or in combination. It supports single-attribute tests and joint-attribute tests.
- We apply the DiffImp method to evaluate the performance of BERT, Transformer, and LSTM models in an in-hospital mortality prediction task, using MIMIC III and eICU datasets. The evaluation shows that demographic information of patients impacts prediction accuracy. For the eICU dataset of 4,602 test cases, the average recall score for the death cases of the models varied between the range of 0.4 and 0.57 after the inclusion of gender information, and 0.48 and 0.62 after incorporating both gender and race information. The balanced accuracy of the models remained mostly unchanged, with a minimum and maximum value of 0.65 and 0.75, respectively.
- Our study also reveals the differential impact of demographic variables at the subpopulation level. For some subgroups, their demographic information prompts some models to alter the mortality prediction result, against the statistical trend observed in the training data. For the MIMIC III dataset, the recall score for death cases among Asian and Hispanic patients can decrease by up to 41.9% and 33.3%, respectively, after the incorporation of race information into the model.

Our Differential Impact Testing method is new compared to existing fairness studies as it describes a dynamic testing approach, as opposed to passive subgroup measurement (in typical fairness studies). We actively and systematically adjust how demographic information is fed into the model and then measure the impact and change. This kind of dynamic evaluation workflow by actively perturbing input (exists in the security domain) is new in the digital health domain. The codes for DiffImp and all supplementary materials are available at https://github.com/MohimenulRafi/DiffImp.

2. Methods

This section will describe the data sets, data preprocessing, the models used, and finally, our method, DiffImp.

2.1. Dataset And Preprocessing

For our experiment, we used the Medical Information Mart for Intensive Care III (MIMIC-III) [29] and the eICU data [30], which contains 48-hour time-series data from ICU patients. The MIMIC III and eICU contain 21,139 (13.2% death cases, 86.8% survival cases, 55.01% male, 44.99% female, 2.69% Asian, 82.57% White, 11.03% Black, 3.71% Hispanic), and 30,680 (11.5% death cases, 88.5% survival cases, 54.6% male, 45.4% female, 1.72% Asian, 82.54% White, 11.87% Black, 3.88% Hispanic) patient data instances, respectively. For MIMIC-III and eICU, we used the clinical prediction benchmark of [31] and [32], respectively.

In both datasets, each patient has 17 attributes, including diastolic blood pressure, Glascow coma scale eye opening, Glascow coma scale motor response, respiratory rate, systolic blood pressure, temperature, etc. The datasets have two classes: death (Class 1 and presented in this work as C1) and survival (Class 0 and presented in this work as C0).

The datasets are divided into train, validation, and test sets. After the split, the MIMIC III dataset contains 14,681, 3,222, and 3,236 instances for the train, validation, and test sets, respectively, and eICU contains 21,476, 4,602, and 4,602 instances for the same splits. In each split, we maintained a similar demographic and class proportion to ensure that inherent bias in the data distribution does not cause performance disparity. Further details on the dataset are provided in the supplementary.

Our datasets contain patient information for the first 48 hours of an ICU stay. We used MIMIC III and eICU benchmark data and preprocessed them following [31] and [32], respectively. For missing values, we imputed the attributes with the most recent value if it exists. If not, we use a prespecified normal value similar to that of [31]. A binary mask for each variable was used to indicate the true value vs. the imputed value. Categorical variables were encoded using one-hot encoding, while numerical variables were standardized by z-score normalization. This preprocessed data was used to train, validate, and test the ML models.

The LLMs take texts as input and generate responses. However, our problem involves time-series tabular data with numerical features. To utilize the LLM, we serialized the data into text using the *Table-To-Text* method proposed in [33] so that we can use the model in a conversational way, which is more suitable in medical scenarios. In our dataset, since each patient has data spanning 48 hours with 17 attributes recorded in each hour, the serialization of such data will result in a significantly large text. Therefore, we created 4 windows having 12 hours of length spanning 0-48 hours. We considered the average values of each attribute within each window and then converted the data of those windows into text. This preprocessed input was used for fine-tuning and prediction using the language model.

2.2. Models

In this study, we trained and tested three types of models - LSTM models and Transformer models that are trained from scratch, and pretrained BERT models. We used the Bidirectional Encoder Representations from Transformers (BERT) [34] model as the LLM. We chose the BERT language model because it is good for classification when finetuned on a task-specific dataset. In addition, it is a computationally efficient and interpretable baseline for fairness evaluation. We considered benchmark models from existing literature to select the best-performing models. Long short-term memory (LSTM) performs the best in the MIMIC III benchmark dataset, as indicated in [31], and a recent study [35] demonstrated that Transformer performs similarly (higher in some cases) to LSTM on MIMIC III and eICU datasets. We included the LSTM model architecture from [31] and the Transformer model architecture from [35] as strong baseline ML models. Each model was applied to each dataset (i.e., MIMIC III and eICU) in four different demographic configurations. Each configuration presents a systematic variation of demographic input attributes.

For the prediction task using LLM, we finetuned the BERT model. We finetuned the model as a classifier because we want to achieve higher performance by providing domain- and task-specific knowledge to the model. Empirical studies have shown that fine-tuning pretrained language models outperform models trained from scratch on limited-domain data, as they benefit from prior linguistic knowledge while adapting to the task-specific domain [36], [37]. Moreover, finetuning the model is computationally more efficient than pretraining from scratch. Finetuning the model as a classifier enables more empirical and quantitative analysis of a specific task while maintaining computational efficiency and robustness. The BERT model we finetuned is a multilayer bidirectional transformer encoder architecture. It uses a masked language model (MLM) that uses both left and right context in all layers. The Transformer model is designed for processing time-series data, leveraging multi-head self-attention mechanisms to capture temporal dependencies effectively. The number of parameters for the BERT, Transformer, and LSTM models was 110M, 293K, and 7.5K, respectively (more details in the supplementary).

2.3. Differential Impact Testing

We developed a differential impact testing method, DiffImp, and applied it to assess the influence of demographic information on the mortality risk prediction capabilities of the LLM and ML models. We systematically control the demographic attributes of the patients as input variables to the model and evaluate the change in the predictive outcome.

2.3.1. Test Configurations. We trained the three models (1 LLM and 2 baseline ML models), each with 4 configurations: 1) No demography, 2) Gender, 3) Race, and 4) Gender & Race. In the "no demography" setting, we train the models using the patient diagnostic values with no demographic information. In the other three configurations, we performed the *single-attribute* and *joint-attribute* tests incorporating demographic variables into the data.

In the *single-attribute* test (settings 2 and 3), we consider a single demographic variable, add it as an attribute to the data, and perform training and testing with that additional information. In the *joint-attribute* test (setting 4), we add the combination of two demographic variables as additional attributes to the data, and perform similar training and testing with that additional information. In particular, in the "gender" setting, we add gender as a new attribute and perform the training and testing. In the "race" setting, we introduce race instead of gender as a new attribute and perform a similar experiment. The intuition is to understand the impact of each individual demographic variable alone on mortality prediction. In the "gender & race" setting, which is a joint-attribute test, we add both gender and race as new attributes to the data and perform training and testing. The purpose is to understand how the prediction of the model is affected when a combination of demographic information is provided.

After training each model, we tuned the classification threshold using the validation set to achieve optimal performance and to create a balance in the relevant performance metrics. Then, in the evaluation phase, we assess the impact of demographic information using several metrics.

2.3.2. Testing Procedure. We compared different performance metrics between models in each data set for each setting. For comparison, we considered several metrics that indicate the model's overall performance and also the performance on a specific class. For overall performance, although ROC-AUC, accuracy give an overall idea, they can be misleading for an imbalanced dataset. In such a case, we considered the Matthews Correlation Coefficient (MCC) and balanced accuracy, which gives us a better understanding of the performance quality of the models. Also, since class 1 represents death cases, it is crucial to identify this class accurately. Therefore, we compared the recall, precision, and F1 scores for the minority class (class 1 and represented as Rec_C1, Prec_C1, F1_C1, respectively). We also tested recall, precision, and F1 scores for the majority class (class 0 and represented as Rec_C0, Prec_C0, F1_C0, respectively).

We examine whether there is a change in the performance of the models across the four configurations. A change in the metric indicates the impact of demographic information on the models. We then evaluated the performance of the models for each subgroup to measure the differential impact of demographic variables on the subgroup level. Furthermore, to better understand the effect of demographic variables, we evaluated the percentage of change in the prediction of the model for each subgroup after incorporating demographic information. We compared the percentage of patients for whom the model predicted one class without demographic information, and later altered the prediction to another class with the incorporation of this information. DiffImp utilizes different versions of the model

controlling demographic variables for the same patient input data and assesses fairness beyond feature ablation.

3. Experiment Setup

In this section, we will highlight the model training process, how we tuned the hyperparameters, and the threshold selection process for evaluation.

We train the models using the training and validation sets. In the "no demography" setting, the input of LSTM and transformer models consists of 76 features. When demographic attributes are added, the feature count for the LSTM and transformer model adjusts accordingly: 80 for gender, 82 for race, and 86 for both gender and race, with the model input parameter updated to reflect these variations. For the BERT model, the demographic information is added as text to the input, which eventually increases the input token size.

The LSTM and transformer models were trained for 100 and 20 epochs, respectively. We select the model based on validation AUPRC and validation loss. To select the model, we selected the top 3 epochs based on the highest validation AUPRC and then selected the one with the lowest validation loss. Hyperparameter tuning was performed using grid search over a predefined space, optimizing model performance by systematically exploring different configurations. The search included variations in attention heads (2, 4, 8), number of transformer blocks (1-3), feedforward network dimensions (16, 64, 256), dropout rates (0.1, 0.3, 0.5), batch normalization (enabled/disabled) and hidden unit sizes (16, 32, 64), ensuring optimal model robustness and generalization. The BERT model was finetuned in a similar manner for 10 epochs. For effective convergence, we chose 10 epochs for finetuning and selected the model with the lowest validation loss to avoid possible under-/overfitting. The models were trained to optimize the binary cross-entropy loss function. We used the Adam optimizer to train the LSTM and transformer models and the AdamW optimizer to train the BERT model. For each model and each variable setting, we conducted three independent trials with different initialization of model parameters and reported the mean scores. We trained the models using our local servers and ensured proper data privacy and security.

To achieve optimal and balanced performance from the models, we tuned the threshold for the decision boundary based on the F1 score for class 1 (F1_C1) and balanced accuracy on the validation set. The threshold tuning was performed in the range between 0.0 and 1.0 with a step size of 0.01. We selected the top three thresholds that give the highest F1_C1 score and finally selected the one that gives the highest balanced accuracy on the validation set.

We analyzed the thresholds for different models on MIMIC III and eICU data. The threshold for the positive classes increased in general when demographic information was added. The shift in the threshold after the inclusion/exclusion of demographic information indicates that such information influences the identification of positive classes. For example, in MIMIC III data, the threshold for the LSTM model ranged from 0.17 to 0.28 in four configurations. The increase in the threshold after the inclusion of demographic information is an indicator that the model struggles to optimize the precision and recall with such information.

4. Results

In this results section, we will first evaluate the model's performance based on the four configurations. Then, a comparative analysis will be performed for each subgroup, followed by an analysis of the percentage of change in the prediction of the models after the incorporation of demographic information. Additional details are in the supplementary.

4.1. Demographic Effects on Model Performance

We report the performance metrics of the models based on the configurations of our test. We observe that although finetuning the LLM as a classifier exhibited performance comparable to conventional neural network models, the Transformer and LSTM models performed better than the LLM in terms of precision (class 1), F1 (class 1) and MCC values (Figures 1b, 1c, 1f). In addition, the performance of the models varies when demographic variables are included. In the eICU dataset, the average recall score for death cases (Rec_C1) ranged between 0.4 and 0.62 after the inclusion of demographic information (Figure 1a). The balanced accuracy did not vary much. However, it ranged between 0.65 and 0.75 in the models after the demographic information was added (Figure 1d).

As indicated in Figure 1a, the mean Rec_C1 for BERT in MIMIC III increases from 0.52 to 0.55 after incorporating demographics. However, it decreases for the LSTM (from 0.62 to 0.56) and Transformer models (from 0.58 to 0.52). In contrast, in the eICU data, the mean Rec_C1 for BERT decreases (from 0.49 to 0.4) while it fluctuates for LSTM and Transformer. When both gender and race are included, Rec_C1 increased for both the LSTM (from 0.58 to 0.62) and the Transformer (from 0.51 to 0.55) models. The F1_C1 (Figure 1c) had a trend to decrease slightly except for the LSTM model with MIMIC III, where it increased slightly (from 0.51 to 0.53 with joint attribute).

We observed that gender has a slightly negative impact on balanced accuracy, ROC-AUC, and MCC in all models and datasets, while there was a small increase in MCC from 0.43 to 0.44 in the LSTM model (Figure 1f). Race has a smaller effect on balanced accuracy, ROC-AUC, and MCC scores than gender. After the incorporation of race information, the scores remain almost similar, with slight variation among models and datasets. This shows that demographic variables affect the consistency of the model performance.

4.2. Performance Comparison Based on Subgroups

Figure 2 illustrates the varying performance of each model at the subgroup level in the MIMIC III data.

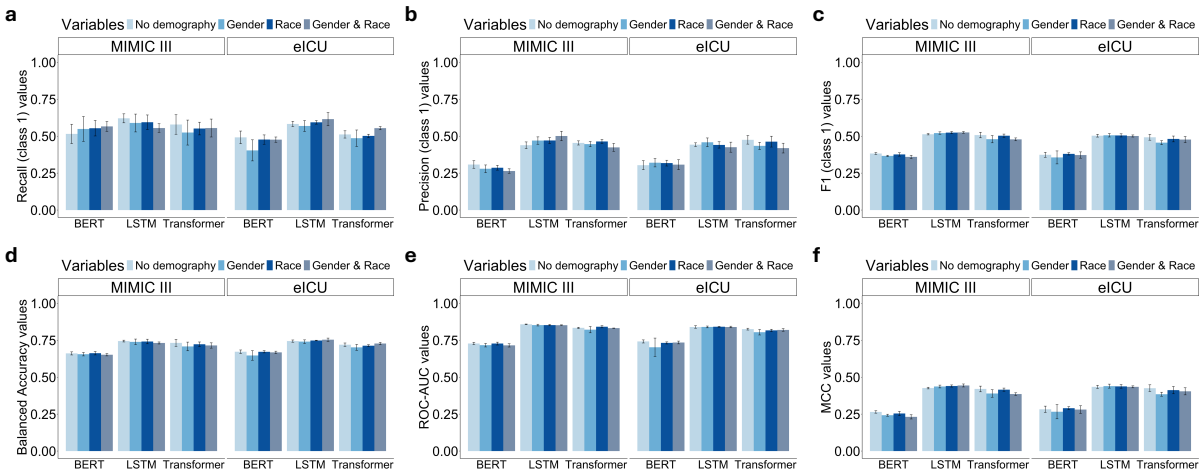

Figure 1. Metric-wise comparison of models on the MIMIC III and eICU dataset based on single and joint demographic variables. Each subfigure (a-f) presents a comparison of the model's performance based on a specific metric for four configurations. Each bar represents one of four configurations, and the mean score was calculated from three independent trials for the corresponding configuration. The standard deviations of the scores in each setting are shown as the error bars. P-values are from one-way ANOVA tests assessing differences in model performance across four settings.

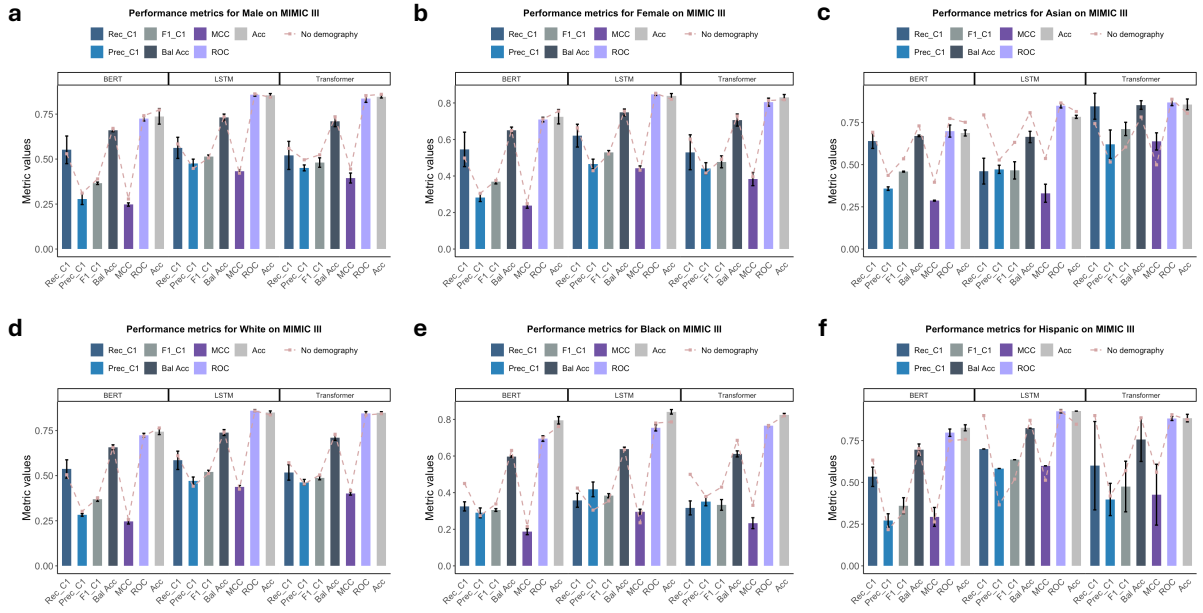

Figure 2. Model performance comparison on MIMIC III data for subgroups when demographic information is added as a single attribute. Subfigures a, b present the performance for male and female patients, respectively, when gender is added. Subfigures c to f present performance for Asian, White, Black, and Hispanic patients, respectively, when race is added. Bars represent the mean metric values (from three independent trials) after adding gender or race information, with error bars showing standard deviations. C1 represents class 1 (death). The dashed line shows results for the "no demography" setting.

The mean Rec_C1 score for the BERT model increased slightly for male, female, and White patients (Figure 2a, 2b, 2d), while the score decreased for Asian, Black, and Hispanic patients (Figure 2c, 2e, 2f). However, in eICU data, the Rec_C1 score of BERT for male and female patients decreased, while the score for Hispanic patients increased. For the LSTM model in MIMIC III data, a decrease in Rec_C1 was observed for all subgroups, with a large decrease of 41.9% for Asian patients (Figure 2c). For the Transformer model in MIMIC III data, a decrease in the score was observed across all groups except Asian patients. Rec_C1 for Hispanic patients decreased by 33.3% (Figure 2f). However, in eICU data, the score for Asian patients slightly decreased (from 0.81 to 0.76), while for Hispanic patients it slightly increased (from 0.37 to 0.42). Such observations indicate that the identification of death cases can be impacted by demographic information of the patients. Overall performance, measured by balanced accuracy, re-

mained almost unchanged or decreased in all datasets across all models, in general, with some exceptions for specific models and subgroups. For example, for the Transformer model, the balanced accuracy increased for Asian patients in MIMIC III (from 0.78 to 0.85). For C0 (survival class), we also observed some variability.

4.3. Percentage of Change in Prediction Analysis

In our study, we analyzed and compared the percentage of change in the prediction of the models for each subgroup after the inclusion of demographic information. We analyzed four cases of change: 1) from survival to death (correct), 2) from death to survival (correct), 3) from survival to death (incorrect), and 4) from death to survival (incorrect).

Gender groups analysis In the MIMIC III dataset, all models showed a higher percentage of correct and incorrect prediction changes from death to survival for female patients than for male patients. For example, for the Transformer model, the percentages of correct change for male and female are 2.69% and 5.07%, respectively. However, the survival rate in the dataset is similar for male and female patients (male: 87.86%, female: 87.44%). In the eICU dataset, we observe the same with some exceptions. For example, the percentages of incorrect change for male and female are 1.36% and 0.9%, respectively, for the Transformer model.

In case of a change from survival to death, the models had a slightly higher error rate (incorrect change) for male patients than for female patients in MIMIC III (for example, for Transformer, male: 3.47%, female: 2.71%), except for the BERT model. However, the original trend in the dataset shows the opposite, where female patients have a higher death rate than males (male: 11.14%, female: 12.56%). In the eICU dataset, we observe the same, except for the Transformer model. For example, the mean error percentage with the LSTM model is 1.71%, and 1.0% for male and female patients, respectively.

Race groups analysis For different race groups, the disparity in the change was also observed. For the correct prediction change from death to survival, BERT showed a higher percentage for Hispanic patients than for other race groups in both datasets (MIMIC III, Hispanic: 10.0%, Black: 7.13%, White: 4.35%, Asian: 3.17%; eICU, Hispanic: 6.4%, Black: 5.86%, White: 5.17%, Asian: 6.02%). In case of incorrect change, the LSTM model showed a higher rate for Asian patients (Hispanic: 1.82%, Black: 1.03%, White: 0.84%, Asian: 6.88%) in MIMIC III, while in eICU, both the BERT and the Transformer model showed a slightly higher tendency for White patients (for example, for BERT, Asian: 0%, White: 1.06%, Black: 0.88%, Hispanic: 0.53%). However, Black patients had the highest survival rate in both training datasets (for example, in the eICU, Asian: 87.68%, White: 88.31%, Black: 90.22%, Hispanic: 88.55%).

We similarly observed inconsistency in the prediction change from survival to death for race groups. In MIMIC III, the percentage of correct prediction change to death was slightly higher for White patients than for others in both the BERT (1.5%) and the LSTM models (0.49%). In eICU, the percentages in the BERT, LSTM, and Transformer models were slightly higher for Hispanic and White patients. However, Asian patients had the highest death rate in both training sets (MIMIC III, Asian: 12.19%, White: 11.05%, Black: 6.96%, Hispanic: 7.38%; eICU, Asian: 12.32%, White: 11.69%, Black: 9.78%, Hispanic: 11.45%). In case of incorrect prediction change to death in the MIMIC III data, the BERT model showed a higher percentage for Asian patients than others, followed by White patients (Asian: 8.47%, White: 6.85%, Black: 1.95%, Hispanic: 2.12%). The lower performance score after including race information for Asian patients (Figure 2c) also reflects this negative effect of the BERT model on performance accuracy. This higher error rate for Asian patients indicates that the data distribution in the training set can cause fairness issues for specific race groups. The LSTM and the Transformer model showed a higher percentage for White patients, followed by Black patients (for example, for the Transformer, Asian: 0.53%, White: 2.16%, Black: 1.61%, Hispanic: 0.3%). In the eICU data, the BERT, LSTM, and Transformer models exhibited a higher percentage of change for Black, Asian, and White patients than for others, respectively. However, Black patients had the lowest (9.78%) and Asian patients had the highest (12.32%) death rate in the training data.

Across all models (BERT, LSTM, Transformer), incorporating demographic information (gender and race) leads to noticeable changes in predictions for some groups despite the trend in the training data. The degree of change varies by both model type and demographic group, highlighting potential concerns about fairness. For example, in MIMIC III, BERT had a higher percentage of incorrect change from survival to death for Asian and White patients than other models (for Asian, BERT: 8.47%, Transformer: 0.53%, LSTM: 0%; for White, BERT: 6.85%, Transformer: 2.16%, LSTM: 1.37%). On the other hand, in the same dataset, LSTM had a higher percentage of incorrect change from death to survival for Asian patients than other models (BERT: 1.59%, Transformer: 1.06%, LSTM: 6.88%). In eICU, BERT had a higher correct prediction shift from death to survival for both male and female patients than other models after adding gender information (for male, BERT: 6.1%, Transformer: 2.98%, LSTM: 1.43%; for female, BERT: 6.97%, Transformer: 2.77%, LSTM: 2.77%). Significant changes in death-to-survive (wrong) suggest that certain racial or gender groups may be at higher risk of misclassification, potentially leading to delayed or incorrect treatments. These observations indicate that the prediction can be influenced by both the data distribution in the training set and the inclusion of demographic information.

5. Discussion

Performance impact We observed that although balanced accuracy remained mostly unchanged, there was a slight decrease in the score when gender or race information was added. This indicates that demographic information might not improve overall performance and may potentially

have a slightly negative impact on mortality risk prediction. In the subgroup level analysis, we sometimes observed a large decrease in Rec_C1 scores. For example, the Rec_C1 scores for Asian (drop from 0.79 to 0.46 with LSTM) and Hispanic patients (drop from 0.9 to 0.6 with Transformer) in MIMIC III data. This decrease in these subgroups might be due to their small sample size in training. However, the decrease suggests that incorporating race information might hurt the performance of identifying death cases for specific race groups. This will eventually cause unequal access to healthcare in critical situations. It is interesting to note that in the eICU data, although the Rec_C1 score of the Transformer and LSTM model decreased with the inclusion of gender and race information as a single attribute, the score improved when they were used in combination. This indicates that a joint attribute can have an advantage over a single attribute in identifying death cases correctly.

Prediction switching Our analysis of percentage change reveals that the models have a higher tendency to change predictions for specific subgroups. However, the original death or survival rate had an opposite trend in the training dataset. Such observation indicates that data distribution alone may not impact the learning or prediction of the model, and demographic variables could influence the prediction outcome. During training, models' learning may be impacted by demographic features than by the diagnostic values. Assessing feature importance and token attention scores can reveal the impact on training. Our findings imply that before their clinical deployment, LLM and ML models need to be carefully assessed to understand the impact of demographic information not only in general, but also on the patient subgroup level. Including demographic information had some negative impacts on model performance, specifically for some subgroups (for example, Asian). However, other factors such as data distribution and socioeconomic conditions should be taken into account to ensure fairness in the evaluation. A continued focus on more evaluations like ours is needed to develop fair machine learning and large language models.

Limitations Our study is limited to two datasets and three models and focuses on time series vitals without incorporating clinical notes. Another limitation of this study is the data imbalance in death and survival cases. Additionally, there are small group sizes for some subpopulations. Also, the number of experiment trials per setting was limited (3 per setting), and the performance difference did not reach statistical significance. On the other hand, demographic factors can be embedded in medical history and biomarkers. One subgroup can differ from another in ranges of certain vitals such as heart rate, body temperature, etc., and this difference can impact the model performance. Our study does not differentiate these latent aspects or any inherited biases from the pretrained LLM. We aim to assess the overall bias (both inherited and data-driven) in the models. Some prediction shifts can result from differences among differently trained models rather than from demographic variables alone. Moreover, our proposed framework measures fairness but does not remove bias from the models.

6. Conclusions and Future Work

In this study, we proposed a differential testing method to evaluate the demographic impact on medical prediction models. We demonstrated the impact of gender and race as both single and join attribute in the mortality prediction task. Although there were both positive and negative impacts on specific performance metrics, metrics indicating overall performance remained mostly unchanged or declined slightly. Furthermore, demographic variables impact the pretrained LLM differently from conventional neural network models. The change in prediction after adding demographic information was greater for the LLM in some cases (for example, in MIMIC III, the BERT model's incorrect change from survival to death for Asian: 8.47%).

The vision of precision medicine depends on demographic information. Thus, there is an urgent need to understand how such information (sometimes latent) in datasets and pre-trained models impacts clinical decisions. Our work presents the initial measurement study with the need for more in-depth future investigation. A major future direction is to evaluate how demographic information impacts the performance of generative AI models (e.g., Llama-3, Mistral, Meditron, Phi-3) in digital health settings. Because generative AI models are being rapidly considered for healthcare purposes, we plan to measure their correctness and fairness under various settings, including chain-of-thought prompting. To distinguish between inherited and data-driven biases, one may generate predictions from LLMs without any task-specific finetuning and do the same after task-specific finetuning. One can measure and compare the degrees of change in model behaviors. Prediction without finetuning would inform us about the potential inherited bias.

Acknowledgment

This work was partly supported by the National Science Foundation under Grant No. SaTC-2231002. We thank the HealthSec'25 reviewers for their valuable comments and feedback.

References

[1] C. Li, Z. Zhang, Y. Ren, H. Nie, Y. Lei, H. Qiu, Z. Xu, and X. Pu, "Machine learning based early mortality prediction in the emergency department," *International Journal of Medical Informatics*, vol. 155, p. 104570, 2021.

[2] G. Kong, K. Lin, and Y. Hu, "Using machine learning methods to predict in-hospital mortality of sepsis patients in the icu," *BMC medical informatics and decision making*, vol. 20, pp. 1–10, 2020.

[3] A. Vaid, S. Somani, A. J. Russak, J. K. De Freitas, F. F. Chaudhry, I. Paranjpe, K. W. Johnson, S. J. Lee, R. Miotto, F. Richter *et al.*, "Machine learning to predict mortality and critical events in a cohort of patients with covid-19 in new york city: model development and validation," *Journal of medical Internet research*, vol. 22, no. 11, p. e24018, 2020.

[4] D. Bertsimas, G. Lukin, L. Mingardi, O. Nohadani, A. Orfanoudaki, B. Stellato, H. Wiberg, S. Gonzalez-Garcia, C. L. Parra-Calderon, K. Robinson *et al.*, "Covid-19 mortality risk assessment: An international multi-center study," *PloS one*, vol. 15, no. 12, p. e0243262, 2020.

[5] C. Zang, Y. Hou, D. Lyu, J. Jin, S. Sacco, K. Chen, R. Aseltine, and F. Wang, "Accuracy and transportability of machine learning models for adolescent suicide prediction with longitudinal clinical records," *Translational psychiatry*, vol. 14, no. 1, p. 316, 2024.

[6] D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová *et al.*, "Adapted large language models can outperform medical experts in clinical text summarization," *Nature medicine*, vol. 30, no. 4, pp. 1134–1142, 2024.

[7] D. Van Veen, C. Van Uden, M. Attias, A. Pareek, C. Bluethgen, M. Polacin, W. Chiu, J.-B. Delbrouck, J. M. Z. Chaves, C. P. Langlotz *et al.*, "Radadapt: Radiology report summarization via lightweight domain adaptation of large language models," *arXiv preprint arXiv:2305.01146*, 2023.

[8] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal *et al.*, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

[9] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[10] R. Movva, D. Shanmugam, K. Hou, P. Pathak, J. Guttag, N. Garg, and E. Pierson, "Coarse race data conceals disparities in clinical risk score performance," in *Machine Learning for Healthcare Conference*. PMLR, 2023, pp. 443–472.

[11] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[12] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert *et al.*, "Disparities in dermatology ai performance on a diverse, curated clinical image set," *Science advances*, vol. 8, no. 31, p. eabq6147, 2022.

[13] L. Seyyed-Kalantari, H. Zhang, M. B. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature medicine*, vol. 27, no. 12, pp. 2176–2182, 2021.

[14] P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis *et al.*, "Evaluation and mitigation of the limitations of large language models in clinical decision-making," *Nature medicine*, vol. 30, no. 9, pp. 2613–2622, 2024.

[15] J. A. Omiye, J. C. Lester, S. Spichak, V. Rotemberg, and R. Daneshjou, "Large language models propagate race-based medicine," *NPJ Digital Medicine*, vol. 6, no. 1, p. 195, 2023.

[16] J. P. Cerdeña, M. V. Plaisime, and J. Tsai, "From race-based to race-conscious medicine: how anti-racist uprisings call us to act," *The Lancet*, vol. 396, no. 10257, pp. 1125–1128, 2020.

[17] S. Afrose, W. Song, C. B. Nemeroff, C. Lu, and D. Yao, "Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction," *Communications medicine*, vol. 2, no. 1, p. 111, 2022.

[18] A. Vaidya, R. J. Chen, D. F. Williamson, A. H. Song, G. Jaume, Y. Yang, T. Hartvigsen, E. C. Dyer, M. Y. Lu, J. Lipkova *et al.*, "Demographic bias in misdiagnosis by computational pathology models," *Nature Medicine*, vol. 30, no. 4, pp. 1174–1190, 2024.

[19] T. S. Pias, Y. Su, X. Tang, H. Wang, S. Faghani, and D. Yao, "Enhancing fairness and accuracy in diagnosing type 2 diabetes in young adult population," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[20] C. F. Manski, J. Mullahy, and A. S. Venkataramani, "Using measures of race to make clinical predictions: Decision making, patient health, and fairness," *Proceedings of the National Academy of Sciences*, vol. 120, no. 35, p. e2303370120, 2023.

[21] S. Khor, E. C. Haupt, E. E. Hahn, L. J. L. Lyons, V. Shankaran, and A. Bansal, "Racial and ethnic bias in risk prediction models for colorectal cancer recurrence when race and ethnicity are omitted as predictors," *JAMA Network Open*, vol. 6, no. 6, pp. e2 318 495–e2 318 495, 2023.

[22] K. Ladin, J. Cuddeback, O. K. Duru, S. Goel, W. Harvey, J. G. Park, J. K. Paulus, J. Sackey, R. Sharp, E. Steyerberg *et al.*, "Guidance for unbiased predictive information for healthcare decision-making and equity (guide): considerations when race may be a prognostic factor," *NPJ Digital Medicine*, vol. 7, no. 1, p. 290, 2024.

[23] L. N. Borrell, J. R. Elhawary, E. Fuentes-Afflick, J. Witonsky, N. Bhakta, A. H. Wu, K. Bibbins-Domingo, J. R. Rodríguez-Santana, M. A. Lenoir, J. R. Gavin III *et al.*, "Race and genetic ancestry in medicine—a time for reckoning with racism," pp. 474–480, 2021.

[24] V. L. Bonham, S. L. Callier, and C. D. Royal, "Will precision medicine move us beyond race?" *The New England journal of medicine*, vol. 374, no. 21, p. 2003, 2016.

[25] P. Nong, J. Adler-Milstein, N. C. Apathy, A. J. Holmgren, and J. Everson, "Current use and evaluation of artificial intelligence and predictive models in us hospitals: Article examines uses and evaluation of artificial intelligence and predictive models in us hospitals," *Health Affairs*, vol. 44, no. 1, pp. 90–98, 2025.

[26] V. Muralidharan, J. Schamroth, A. Youssef, L. A. Celi, and R. Daneshjou, "Applied artificial intelligence for global child health: Addressing biases and barriers," *PLOS Digital Health*, vol. 3, no. 8, p. e0000583, 2024.

[27] R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, and J. Zou, "Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review," *JAMA dermatology*, vol. 157, no. 11, pp. 1362–1369, 2021.

[28] C. Meng, L. Trinh, N. Xu, J. Enouen, and Y. Liu, "Interpretability and fairness evaluation of deep learning models on mimic-iv dataset," *Scientific Reports*, vol. 12, no. 1, p. 7166, 2022.

[29] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[30] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU collaborative research database, a freely available multi-center database for critical care research," *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.

[31] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, p. 96, 2019.

[32] S. Sheikhalishahi, V. Balaraman, and V. Osmani, "Benchmarking machine learning models on multi-centre eICU critical care dataset," *Plos one*, vol. 15, no. 7, p. e0235424, 2020.

[33] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, "Tabllm: Few-shot classification of tabular data with large language models," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 5549–5581.

[34] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[35] T. S. Pias, S. Afrose, M. D. Tuli, I. H. Trisha, X. Deng, C. B. Nemeroff, and D. D. Yao, "Low responsiveness of machine learning models to critical or deteriorating health conditions," *Communications Medicine*, vol. 5, no. 1, p. 62, 2025.

[36] K. N. Jensen and B. Plank, "Fine-tuning vs from scratch: Do vision & language models have similar capabilities on out-of-distribution visual question answering?" in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 1496–1508.

[37] A. Bonfigli, L. Bacco, M. Merone, and F. Dell'Orletta, "From pre-training to fine-tuning: An in-depth analysis of large language models in the biomedical domain," *Artificial Intelligence in Medicine*, vol. 157, p. 103003, 2024.