When to Use Demographic Data in Healthcare AI Models: A Bias-Responsible Approach

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Sebrina Zeleke,

Graduate Program in Department of Computer Science and Engineering

The Ohio State University

2024

Master's Examination Committee:

Tanya Berger-Wolf, Advisor

Xia Ning

Abstract

Given AI's increasing role in healthcare, it is vital to ensure that created models neither perpetuate nor introduce new biases. A naive approach to bias reduction is to exclude demographic data during model training. However, in the healthcare sector, this approach may not produce the most effective models as these features could hold vital information related to care. This dissertation investigates the balance between optimal performance and algorithmic bias associated with the use of demographic data.

The dissertation begins with a case study illustrating the necessity of examining the balance between performance and bias resulting from the use of demographic information. The analysis reveals a clear trade-off when using demographic information, necessitating a structured method for evaluating such a trade-off. Consequently, the dissertation introduces a framework to quantify this trade-off and make decisions regarding the type and extent of demographic information to be used for model training. This framework offers a mechanism to decide whether to use no, some, or all available demographic information by providing a quantified method to identify the bias-performance trade-off. The framework is then tested on two healthcare applications, and a dashboard is developed to analyze the pattern of the results.

The findings indicate a trade-off when including demographic features for model performance, and also suggest a more equitable alternative than using all demographic information. Lastly, the dissertation discusses some of the significant results and ethical considerations of the general use of demographic information observed from the outputs of the two models. By examining various model outcomes in detail, the research offers valuable insights into the intricate relationship between demographic information and model performance in healthcare applications.

Acknowledgments

I would like to express my heartfelt gratitude to Dr. Tanya Berger-Wolf, my advisor, for her unwavering support and invaluable guidance throughout my journey. Over the last couple of years, I have learned so much from you, and I am truly grateful for the opportunity to have been mentored by you.

Additionally, I am deeply thankful to Dr. Xia Ning, my committee member, for her hands-on mentorship and continuous support.

I would also like to thank my friends who have been with me every step of the way. Their support has helped me through the ups and downs of graduate school and I am extremely grateful to God for having them all in my life.

Last but certainly not least, I want to express my deepest appreciation to my family. Thank you for the sacrifices you have made to help me reach this point. Your advice, support, and unwavering presence are what helped me get to where I am today. I am forever grateful for your love and this would not have been possible without your support.

Praise be to the Lord!

Vita

2017 - 2020	B.S. Computer Science, Calvin University
2022 - present	Graduate Research Associate, The Ohio State University.

Fields of Study

Major Field: Department of Computer Science and Engineering

Table of Contents

Page

List of Tables

Table **Page**

List of Figures

Figure **Page**

Chapter 1: INTRODUCTION

1.1 AI and Healthcare

AI systems such as machine learning (ML) models are transforming various industries, and healthcare is no exception. ML models have been utilized to perform various tasks in the healthcare sector, such as diagnosing conditions, assisting doctors in making decisions, and patient monitoring. Although it is not feasible to list all applications of AI in healthcare within the scope of this work, the following list provides insight into some of the most significant and emerging applications of AI in the healthcare sector.

Diagnosis

ML models have been effectively applied in diagnosing various medical conditions such as breast and skin cancer, COVID-19, and various heart and kidney conditions. For example, studies have used approaches such as Convolutional Neural Networks (CNN) and overall deep learning for cancer detection using mammography and ultrasound images ([35]; [19]). Additionally, studies have also demonstrated the utilization of CT images and medical history for diagnosing COVID-19 ([26]; [3]). Furthermore, a study by Ozsahin et al. [37] revealed a substantial increase in publications related to AI/ML and heart, rising from 559 in 2010 to 5697 in 2022, marking a percentage increase of approximately 900%, highlighting the growing use of AI in healthcare solutions.

Assitance in decision making

Another area of application in healthcare is assisting doctors in making decisions beyond diagnosis. For example, AI has been used to predict the readmission risk of patients back to the hospital and ICU, a crucial aspect in reducing mortality risk and ensuring extended care [23]. Additionally, hospitals are ranked based on the quality of care they offer and face government fines for falling below a specified level of care, with patient readmission within a designated time frame serving as one of the key metrics. The utilization of AI in predicting patient readmission not only mitigates mortality risks, but also supports hospitals in maintaining a certain standard of care, ultimately benefiting both patients and healthcare providers[25]. Similarly, ML models have been used to predict patients' mortality risk and anticipate length of stay for proactive interventions allowing healthcare providers to assess resource allocation effectively[21].

Patient monitoring

In times when there is a tremendous shortage of medical staff and equipment, AI models play a crucial role in easing the continuous monitoring workload for healthcare professionals. Hence, ML models have been utilized for ongoing remote monitoring of patients, promptly alerting healthcare professionals in emergencies, and ensuring timely and appropriate care for patients[31]. Continuous monitoring can take various forms, including video-based methods where images and videos capture changes in patients' condition. Alternatively, it may involve IoT devices, allowing real-time monitoring of patients [34].

1.2 AI and Bias

In every context where AI systems are used, they raise the concern of bias against different groups of people. One of the most reported cases of ML bias is found in the criminal justice sector, involving the COMPAS algorithm. The COMPAS algorithm is a 'risk assessment tool' that is used by different courts in the USA to assist in decision making. A study by ProPublica found that this algorithm, which was utilized by courts to assess the risk of recidivism among defendants, inaccurately predicts future criminal behavior for black defendants at twice the rate compared to their white counterparts. In addition, this study also found that, "only 20 percent of the people predicted to commit violent crimes actually went on to do so" [5].

Machine learning (ML) bias is not an issue exclusive to the criminal justice system, as various studies have indicated. ML models have also been found to be biased against women in recruiting, especially in fields such as engineering. This bias often stems from data bias, where these models are trained predominantly on resumes of male employees, leading the model to associate being male as a criterion for success [10]. Furthermore, ML models have also been found to be biased in advertising, healthcare, and education. ([2]; [1]; [27]).

1.3 AI, Healthcare and Bias

The integration of AI into healthcare has the potential to significantly increase the number of lives saved. Hence, it is crucial to thoroughly evaluate the process to ensure that all patients receive equitable benefits from the use of such systems. The evaluation of AI systems in healthcare has been a slow and challenging process because researchers have difficulty accessing healthcare data or the algorithms. It is understandable that healthcare data, due to its sensitive nature, must not be publicly accessible. However, until a pipeline is developed where researchers are allowed to evaluate such systems, the AI system in the healthcare sector can remain a black box that perpetuates the unconscious and historical bias embedded within healthcare. [27]

Studies have shown that ML models in healthcare can contribute to underdiagnosing patients from underserved populations, with an exacerbation of this problem when patients belong to multiple underserved groups simultaneously ([33]). Furthermore, studies have also shown that individuals from various underserved groups experience more inaccurate model predictions and unfavorable outcomes ([29]; [32]; [27]). To ensure the sustained integration of machine learning into healthcare in the long run, it is crucial to actively combat these biases.

ML biases can be caused by various reasons, such as the utilization of incorrect proxies for prediction tasks, as shown by [27], where the wrong features are used to train different ML models. Additionally, biases embedded in historical data can be magnified and result in biased models [33]. Furthermore, data availability bias, where data not originally collected for ML purposes used in these models, is another contributing factor for bias ([30]; [16]).

1.4 Addressing AI Bias in Healthcare

One naive approach to mitigate ML biases, especially one that results from historically biased data, is to exclude features that could aid in identifying an individual, such as race, gender, and insurance type, from the training data to achieve "Fairness through Blindness" ([28]). "Fairness through Blindness" is a bias mitigation technique used in model training across diverse sectors, healthcare included. This approach assumes that being unaware of sensitive attributes while making decisions leads to fair decision making. While this method is easy to implement and check, it has its limitations. First, absolute blindness is usually impossible. Even though sensitive attributes are not explicitly provided, other features could have correlations with the attributes, therefore allowing for the same grouping as if the sensitive attributes were provided ([15]). This phenomenon has been observed in different sectors, such as criminal justice ([4]) and the advertising industry ([14]). Second, it might not yield the most optimal models. Especially in healthcare where some of these sensitive attributes might provide meaningful insight, excluding them in the model training process might not produce the best-performing models. A study by [23] demonstrated that incorporating all provided demographic information improved the model's performance for predicting Intensive Care Unit(ICU) readmission risk. While the inclusion of sensitive attributes, as shown by [23], might improve model performance, it concurrently raises the risk of introducing additional bias. Therefore, it is crucial to analyze the trade-off between performance and the potential for additional bias to determine the optimal circumstances for leveraging such features.

1.5 Problem Statement

This paper introduces and tests a framework that investigates when it is appropriate to use demographic information in model training to get optimal performance while minimizing additional introduced bias. The framework is then used to analyze whether there is a more fair alternative between using no demographic information and using all provided demographic information. Two healthcare application models were used to test the framework. Lin et al. [23] presents the first model and predicts patients' unplanned ICU readmission risk. The second model, presented by Harutyunyan et al. [21], predicts patients' In-Hospital Morality (IHM) risk.

Our findings provide crucial insight into both machine learning and the healthcare domains. Firstly, this work demonstrates that an increase in a model's performance does not always result in a fairer model, which highlights the necessity for a formalized metric to assess the additional bias introduced. Second, this framework is tested with two healthcare applications, illustrating its adaptability to a broader range of applications. Lastly, the results reveal inherent biases that cannot be addressed by altering model architectures or removing all demographic features. This intrinsic bias underscores the prevailing inequities in today's healthcare system.

The remainder of the thesis is structured as follows: Chapter 3 introduces the two models used for the rest of the analysis, while Chapter 4 elaborates on the data used for these models. Chapter 5 discusses the importance of conducting a trade-off analysis by showing a case study, while Chapter 6 introduces and explores the framework developed in depth. Chapter 7 then talks about the visualization tool created to understand the patterns of the results. Chapter 8 uses the visualization tool developed to highlight key findings for the two models tested, introduced in 3, showcasing the process involved in decision making regarding the use of demographic data. Following this, Chapter 9 discusses the main takeaway from this work. Lastly, Chapter 10 addresses the limitations of the work and outlines potential future research directions.

Chapter 2: RELATED WORK

2.1 The Use of Demographic Data for Healthcare Models

The use of sensitive attributes in healthcare models has been a contentious topic ([38]; [8]). This originates from weighing the benefits these features provide versus the potential bias they introduce. Borrell et al. [7] argue that the responsible use of race in healthcare models is beneficial in providing useful information, and should be used until a better alternative is found. This article further discusses how the use of race and ethnicity for biomedical research captures information other than genetics, such as socioeconomic status and environmental exposure, which could contain important epidemiological data. Borrell et al. [7] also acknowledge the complexity behind using information such as ethnicity, as it is self-identifiable information. For example, if a biracial patient comes to a hospital, they could identify themselves as either white or black, when medically they are considered both, potentially leading to misdiagnosis of the patient. Nonetheless, the paper argues that race and ethnicity, when complemented by ancestry data, can provide a comprehensive understanding, with ethnicity offering environmental exposure information and ancestry providing genetic data.

On the contrary, Cooper et al. [12] argue that the minimal advantages of using race are overshadowed by the possible negative consequences stemming from the extensive legacy of racism in the field of medicine [38]. This perspective highlights concerns that the use of race-specific drugs could divert the attention of healthcare providers from existing effective drugs. Furthermore, it emphasizes the genetic diversity present within a single ethnicity. For example, if a patient is of African descent, there is still a big genetic variation among patients in this group. Therefore, more information, such as educational level and lifestyle, would need to be provided to make any inference about a given patient. Similarly, Vyas et al. [38] discuss biases resulting from using race as a decision factor. This article discusses how, as a result of using race, black patients could be considered lower risk, which reduces their chance of admission to cardiology services, and delays referrals to specialists for kidney problems.

However, Lin et al. [23] have shown that the use of demographic features such as age, gender, ethnicity, and insurance can increase the performance of the model used to predict ICU readmission risk, demonstrating the importance of demographic information for healthcare ML models. This study predicts ICU readmission risk using patients' medical records and experiments with the use of demographic information mentioned above. The results indicated an improved performance of the model when demographic features were incorporated compared to models without them. The authors tested demographic information using either all or none of them in their experiments.

Our work analyzes the model of Lin et al. [23] in Chapter 5 to understand the trade-off between increased performance as a result of using the demographics model and the additional bias introduced. The result shows that there is indeed a trade-off, signifying the importance of performing such an analysis before choosing to use demographic information. This work then provides, in scenarios where these features have to be used, a framework that evaluates the trade-off between additional gained performance and additional bias introduced. The framework is discussed in detail in Chapter 6.

2.2 Fairness Analysis of Healthcare Models

Previous studies have shown the bias in machine learning models within healthcare contexts by examining various models developed for different applications. Röösli et al. [29] evaluated the In-Hospital Mortality (IHM) risk prediction model presented by Harutyunyan et al. [21], the same model used to test the framework in this thesis. Röösli et al. [29] presented a three-stage framework that starts with internal validation followed by external validation and concludes with internal validation after retraining the model with a different dataset. Their findings revealed that this model exhibits classification parity violations for patients with Medicaid insurance during internal validation. Additionally, the results indicated poorer performance for black patients, a pattern observed consistently across all three stages of the evaluation.

A work by Chen et al. [9] also studied the bias across different demographic groups for the application of 30-day psychiatric and ICU mortality readmission using patient notes from New England Hospital and MIMIC-III, a dataset that will be explained in depth later, respectively. When doing bias analysis, the authors found that female patients and patients with public insurance have a higher error rate for the ICU readmission model. They also found that private insurance has the highest error rates for psychiatric readmission.

Similarly, Seyyed-Kalantari et al. ([33]; [32]) investigated the bias of AI-based chest X-ray (CXR) prediction models. The results of the analysis show a disparity in the True Positive Rate (TPR) for female patients, Hispanic patients, and Medicaid-insured patients [32]. Similarly, Seyyed-Kalantari et al.'s [33] evaluation found that female patients and Hispanic female patients have a higher underdiagnosis bias compared to other patient groups. Furthermore, Daneshjou et al. [13] examine the performance of dermatology AI models using datasets containing patients with diverse skin colors and observe poorer performance for darker skin tones.

The models discussed above have all been trained without the inclusion of demographic information. Our work tries to complement the studies mentioned above by training models with demographic information and investigating whether additional bias is introduced as a result. To accomplish this, we introduce a framework for analyzing the bias-performance trade-off associated with the use of demographic information, enabling informed decisions regarding its utilization.

Chapter 3: BASE MODELS

The following subsections introduce the two models utilized throughout this paper. The first model is a readmission risk prediction model from the work of Lin et al. [23]. The second model predicts the mortality risk of patients and this work is done by Harutyunyan et al. [21]. The readmission risk model is used in Chapter 5 for a case study to conduct a thorough trade-off analysis, showcasing the necessity for such an evaluation. Both models are then utilized to test the framework developed in Chapter 6.

3.1 Readmission Risk Prediction

Lin et al. [23] used supervised ML models to predict the ICU readmission risk of patients. The target variable is a binary variable that indicates whether a patient is at high risk for unplanned ICU readmission. Various model architectures and data combinations were tested to enhance the models' performances. The architectures examined included LSTM, CNN, LSTM+CNN, and CNN+LSTM. For input data, low-dimensional diagnosis codes ([11]) were used and experiments involved different time series windows of chart events recorded hourly, such as F48 (first 48 hours after admission) and L48 (last 48 hours before discharge). Furthermore, the use of demographic information was experimented with. In this case, demographic information such as age, gender, ethnicity, and insurance were all used to train the model, or none of them were used. For our experiments, we took the LSTM and

LSTM+CNN models with L48 data, as they were the most explored model and highest-performing model, respectively. The performance of the models was reported using True Positive Rate (TPR) and Area Under the Curve (AUC).

The first model selected for our analysis is the LSTM model. It is a bidirectional model with an additional LSTM layer followed by an output layer that is a one-neuron Sigmoid. In contrast, in the LSTM+CNN model, LSTM outputs hidden units which are then passed into a CNN. The CNN will then compute the feature maps based on the hidden units without zero padding. Similar default parameters as the base model in the paper were used to train these models. For example, the learning rate was set to 0.001 and the Adam optimizer was used with a beta of 0.9. More details about these models can be found in [23].

While debugging the base model's code, we discovered an error in calculating the mean and standard deviation(SD) age of patients. The mean and SD age of the patients was calculated horizontally across a single row instead of vertically. Consequently, the model took the mean and standard deviation of age and other variables of an individual patient's records instead of computing the mean and standard deviation age of all patients. This issue was corrected before using the base model to test the framework.

3.2 Mortality Prediction

The second base model, [21], looks at the clinical data of patients to predict the risk of In-Hospital Mortality (IHM) using the first forty-eight-hour (F-48) data after admission. The target label is a binary output that shows if a patient is at high risk of death before discharge. Out of the models experimented with by the authors, standard LSTM and channel LSTM were selected for our analysis. Standard LSTM was the most basic neural network tested, and channel LSTM was the best-performing model experimented with. The standard

LSTM model is a bidirectional LSTM model with an output sigmoid neuron. In contrast, the channel LSTM model is a “modified version of the standard LSTM where all variables are first independently pre-processed with an individual bidirectional LSTM layer instead of working directly on the full matrix of clinical events as usual” ([29]). Similarly to the readmission risk prediction work mentioned in the preceding section, the code for this study also had an error in computing the mean and standard deviation for age. However, this issue was resolved prior to the application of the model.

The models’ performance was evaluated using Area Under the Curve (AUC) and Area Under the Precision-Recall Curve (AUPRC). The default model parameters, as described in the paper by Harutyunyan et al. [21], were used for the application of this thesis. More details about these models can be found in [21].

Chapter 4: DATASETS

The data used for both models, and therefore our analysis is the MIMIC-III data ([22]). MIMIC-III is a deidentified, comprehensive clinical data of patients admitted to the Beth Israel Deaconess Medical Center ICU in Boston, Massachusetts. The version III data spans over the time frame between 2001 and 2012. This data includes demographic information, imaging reports, vital sign measurements made at the bedside (1 data point per hour), laboratory test results, caregiver notes, procedures, medications, diagnosis codes and mortality (both in and out of the hospital).

The specific details of the data used for each base model and any modifications made to the data on our end are described below.

4.1 Readmission Risk Prediction

For this application, Lin et al. [23] remove patients who are under the age of 18 and patients who passed away in hospitals. The details of a positive binary outcome are extensively explained in the paper [23]. The data was split into 80%,10%,10% for training, validation, and testing. The work conducts a five-fold cross-validation, therefore we also use the average of the five-fold as our final result. Features classified into three distinct categories were used. The first category is chart events, which records the physiological conditions of the patients. Seventeen chart events were used; some of them were continuous data and others were categorical. The categorical data was one-hot encoded to prepare the data for model input. Some of the clinical data in this category include heart rate, glucose, oxygen saturation, body temperature, and pH. The second category is low-level embedding of ICD codes, where different chronic diseases could be found ([11]). The third category is demographic information such as age, ethnicity, insurance, and gender of patients. Table 4.1 shows the subcategories of each demographic feature used to train the models. This feature category was used as part of an experiment in which both its inclusion and exclusion were tested to obtain a higher performing model.

Lin et al. [23] experimented with the use of all demographic information listed above all at once or not at all. Even though including demographic features increased the model's performance, a trade-off analysis was not done to understand if additional bias was introduced due to using such information. In this work, in addition to providing the trade-off between using all demographic information and no demographic data, we also experimented with using some demographic information, where we tested every combination of it to find the best-performing and fairest model.

4.2 Mortality Prediction

Harutyunyan et al. [21] remove patients with ICU transfers, patients with two plus ICU stays per admission, and pediatric patients. The base model did not include cross-validation, so we added a five-fold cross-validation with 80%,10%,10% train, validate, and test split to match the splitting process of the readmission model.

This model only used one category of data mentioned in the readmission risk model, the chart events. Neither diagnosis codes nor demographic information was used to train this model. Given the pivotal role of demographic information in analysis, we have integrated

Demographic Data	**Subgroups**
Age	18-120
Gender	F, M
Ethnicity	White, Black, Hispanic, Asian, Other, No Information
Insurance	Medicare, Medicaid, Private, Self-Pay

Table 4.1: Demographic Information

the functionality to train the model using different demographic features of patients and conduct analyses for such groups. The demographic data category and subcategory used for this model are listed in table 4.1. Moreover, given that the base model doesn't explore the integration of demographic features, this study's experimentation with incorporating all, some, and no demographic features sheds light on the correlation and trade-offs associated with using demographic features in the context of mortality prediction.

Chapter 5: EXAMINING PERFORMANCE BIAS TRADE-OFF: CASE STUDY

The goal of this chapter is to perform an in-depth analysis of trade-offs between bias and performance when demographic data is used to train a model. Specifically, this chapter explores the models presented by Lin et al. [23] to investigate whether the increased performance after using demographic data is consistent across all patients. We systematically explored the trade-off for each demographic variable and their combinations by comparing two identical models that differ only in whether they used particular demographic information.

5.1 Method

As discussed in detail in Chapter 3, Lin et al. [23] used supervised machine learning models to predict ICU readmission risk using patients' clinical data. For our analysis, we took two LSTM models with L-48 data from the work done by [23] as base models where the only difference between the two is the incorporation of demographic data. For this case, the first model did not include any demographic information, whereas the second model included all demographic data provided, such as age, gender, insurance, and ethnicity. The LSTM model was used because it was the most explored and showed the third highest performance improvement with the inclusion of demographic data in the original work. We refer to the model with demographic information as `WD` and the one without it as `WOD`.

The original model by Lin et al. [23] utilizes True Positive Rates (TPR) as one of the metrics for reporting results, and we adopt the same metric to examine performance and bias for two primary reasons. First, it is used to maintain consistency with the original work because it allows us to measure disparity using the originally intended metric. Second, assuming that a true positive prediction gets the benefit of extended care due to the high risk of readmission, TPR allows us to gauge the classification effectiveness of the models and assess whether the inclusion of demographic data has increased or decreased the disparity of such benefit.

To examine the introduced bias resulting from the use of demographic data, the TPR of model `WOD` is computed for different demographic subgroups, and compared to the TPR of model `WD` for the same groups. The TPR for each model is derived by averaging the TPR values obtained through a 5-fold cross-validation. The difference of these TPRs between the models `WOD` and `WD` is then used to measure the disparity of benefit for each demographic group that happens as a result of using demographic data.

To explore further, we extend our analysis to include intersectional demographic groups. This entails repeating the same analysis for patients who belong to different categories of demographic groups, simultaneously. For example, we evaluate how model `WOD` performs for female patients with Medicaid insurance and compare it to how model `WD` performs for the same group of patients.

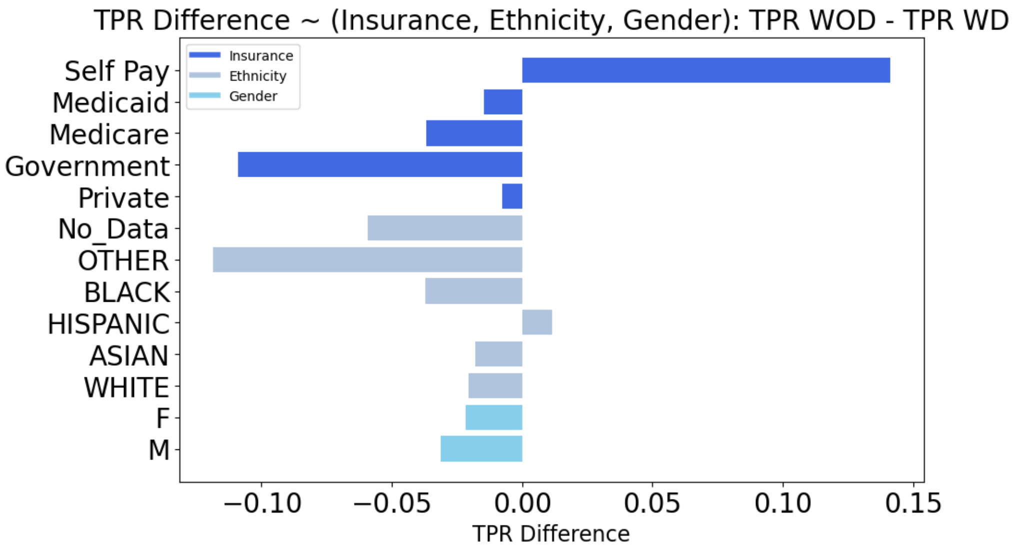


Figure 5.1: TPR difference of model WOD and WD for gender, ethnicity, and insurance separately

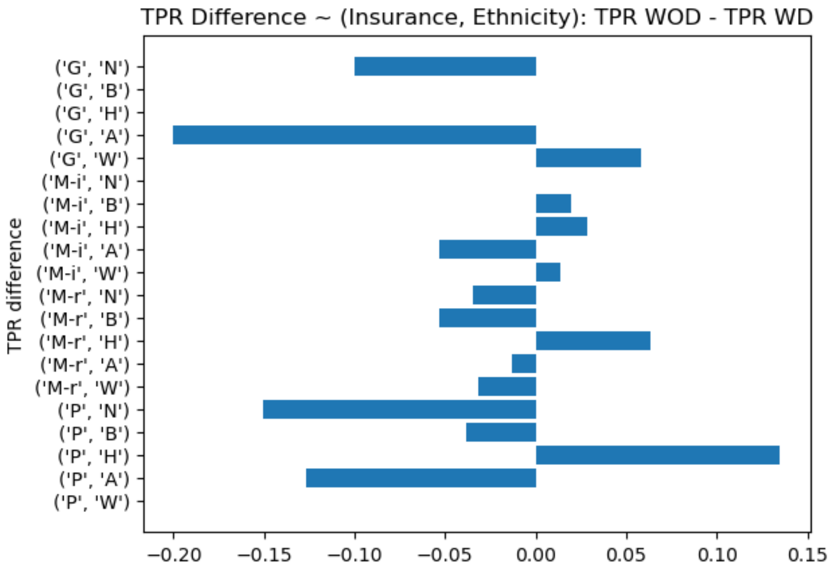


Figure 5.2: TPR of model WOD and WD for intersectional groups (Insurance, Ethnicity) where G: Government, M-i: Medicaid, M-r: Medicare, P: Private are different insurance groups and N: No Data, B: Black, H: Hispanic, A: Asian, W: White are different ethnicity groups

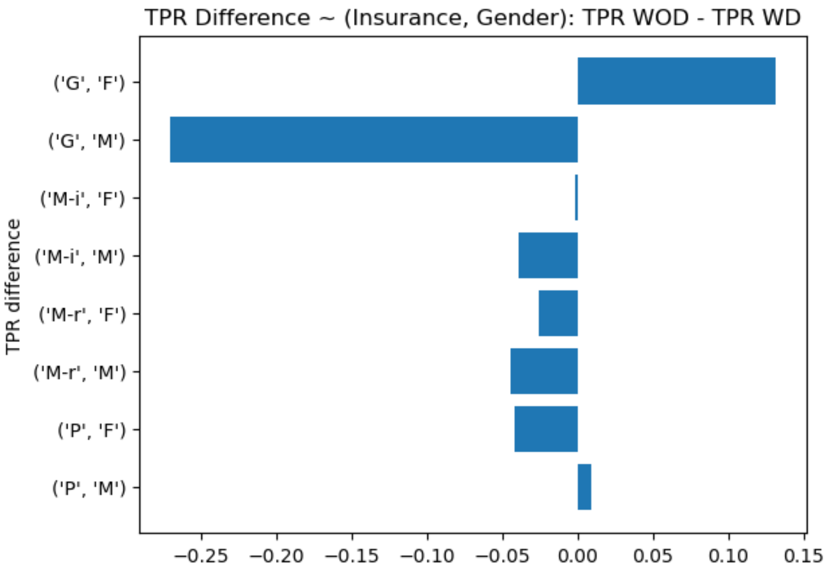


Figure 5.3: TPR of model `WOD` and `WD` for intersectional groups (Insurance, Gender) where F: Female, M: Male, G: Government, M-i: Medicaid, M-r: Medicare, P: Private are different insurance groups

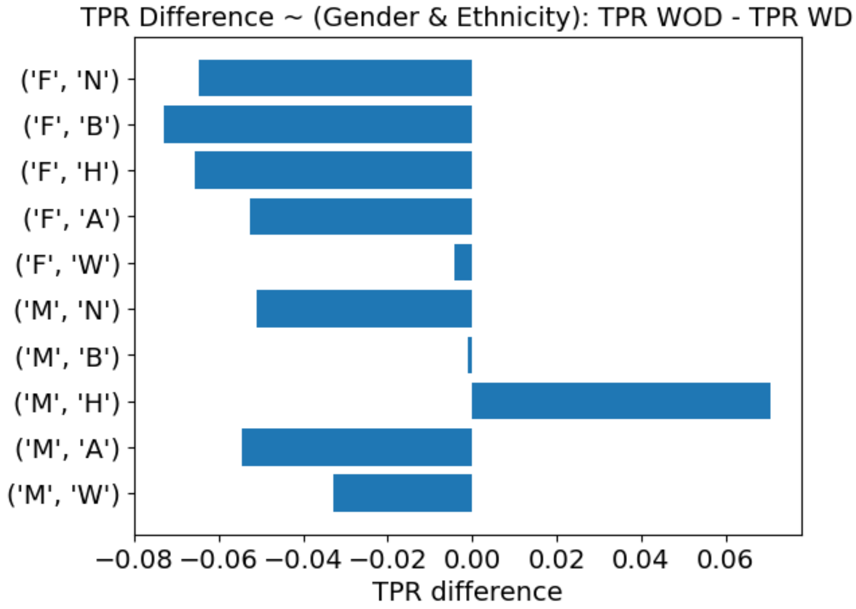


Figure 5.4: TPR of model `WOD` and `WD` for intersectional groups (Gender, Ethnicity) where F: Female, M: Male and N: No Data, B: Black, H: Hispanic, A: Asian, W: White are different ethnicity groups

5.2 Results

When observing the results, bias could be noticed in two ways. First, when the TPR difference is negative for some demographic groups and positive for others, it implies varying benefits from the use of demographic information. Second, when there is a noticeable gap in the magnitude of the TPR difference among different groups, it suggests that the magnitude of benefit from the use of demographic information varies across such groups.

Fig.5.1 to Fig. 5.4 present the TPR difference for individual subgroups and their intersection. Each figure is centered at 0 with positive `WOD` minus `WD` to the right and negative `WOD` − `WD` to the left of the center. The magnitudes of the bars show the extent to which demographic information contributed to the improvement. Figure 5.1 shows the TPR difference for all subgroups across gender, ethnicity, and insurance. Additionally, Figure 5.2, 5.3 and 5.4 show the difference for all the intersectional subgroups.

Figure 5.1 shows that the addition of demographic data increased the benefit of all subgroups except for patients with self-pay and Hispanic patients compared to the model `WOD`. It can also be seen that there is a magnitude difference among both the positive and negative bars. All of the figures illustrate both kinds of biases discussed above. For example, figure 5.3's first type of bias is noticeable when observing the performance bar for female patients with government insurance, where the bar is to the right of the center axis. It can be inferred that the addition did not help this demographic group, resulting in an average performance decrease of approximately 13%. For the second bias, the noticeable comparison is the big difference between females and males with government insurance, where there is a benefit disparity of roughly 40%, although more disparities can be observed. Such inference can be made about all the other figures as well, but it is important to note that as the number of patients decreases in the group, the fluctuations in benefit could be higher and that needs to be kept in mind when making decisions.

5.3 Discussion

As shown throughout this paper, depending solely on performance metrics for reporting can obscure nuanced information, especially in the area of algorithmic fairness. For an increased overall performance of roughly 2 percent TPR, the figures above show the kind of benefit disparity that could be introduced. Depending on the application, the acceptable trade-off and bias could differ, but these kinds of analyses allow us to understand such trade-offs before making decisions.

This chapter presented the result of an analysis that examined the trade-offs between optimal performance and algorithm bias linked to using demographic data. It is important to understand that the use of demographic information does not always increase benefits for all protected groups uniformly. This analysis is key to assessing the trade-off between performance and bias and can be used to decide whether or not to use demographic information.

To generalize the trade-off analysis process to a wider range of healthcare applications and allow developers to apply their data, models, and sets of demographic groups for analysis, we constructed a framework. This framework can be used to quantify the performance bias trade-off associated with the utilization of demographic information. An additional question arising from this analysis is whether there exists a more optimal alternative between utilizing all demographic information and none at all. Consequently, Chapter 6 presents the framework for analyzing trade-offs and identifying a more optimal alternative between using all demographic information and none.

Chapter 6: FRAMEWORK: ANALYZING PERFORMANCE-BIAS TRADE-OFF

Two-stage Framework Introduction

As shown in Chapter 5, the utilization of demographic information, and overall having a better performing model, does not always equate to a fairer outcome. Hence, it is crucial to do an in-depth analysis to understand the possible additional bias introduced due to the use of demographic information before utilizing it. This chapter introduces a framework that can be used to quantify the trade-off and analyze the effect of using demographic information on a case-by-case basis.

The trade-off analysis framework consists of two stages. The first one is model training, followed by trade-off and worst-case disparity analysis. In these two stages, multiple models are trained and subsequently analyzed in the next stage to answer two key questions. The first question aims to understand the performance-bias trade-off between utilizing all demographic information and none for a given healthcare application. The second question aims to explore whether there is a better alternative, in terms of both performance and fairness, to using no demographic data and all demographic information to train healthcare models. This analysis aims to answer whether we can selectively choose different combinations of demographic information groups as features to obtain the most optimal model, which is also fair.

The two stages are explained in depth below:

6.1 Model Training

In any application utilizing this framework, there are two categories of features: those that do not pertain to the demographic information of patients and those that do. We refer to the previous ones as `NonDemographicFeatures` and the latter ones as `DemographicFeatures`. A model trained with any additional `DemographicFeatures` is denoted as `WD` (with demographic) model, while any model trained solely with `NonDemographicFeatures` and without `DemographicFeatures` is denoted as `WOD` (without demographic) model.

Assume that there are N number of `DemographicFeatures` in consideration to be used for training a model. For any given set with N number of demographic features, the cardinality of the power set $P($`DemographicFeatures`$)$ is 2^N, including the empty set. This power set includes every combination of elements of the set, ranging from no elements (the empty set) to all elements included. For model training, while keeping the `NonDemographicFeatures` and architecture of the models similar, the given base model is trained 2^N times. The only difference among these models is the type of `DemographicFeatures` added to train the models, taking one from the power set for each training without replacement. In the end, there will be one `WOD` model and $2^N - 1$ models that are `WD`.

For example, as shown in Chapter 4, $[Age, Gender, Race, Insurance]$ are the four demographic feature types under consideration for the two base models used to test this framework; therefore, $N = 4$. Table 6.1 shows the power set of the above demographic features. Let us take the readmission risk model with LSTM architecture as a sample; then, while keeping the clinical data and architecture of the model similar, the base model would be trained 16 times with a row set of demographic features from Table 6.1 taken for each training round of the base model. Consequently, the resulting models would include 15 WD models and one WOD model.

After the models are trained according to the process described above, the outputs are analyzed in the second stage. This analysis addresses the two key questions posed at the beginning of this chapter. More specifically, training the model with and without the demographic information allows a developer to answer the first question, the trade-off between using no demographic and all demographic information. Additionally, training the model 2^N times with different combinations of demographic information allows for the exploration of the impact of each demographic and the interaction between multiple demographic features for a given application, allowing to answer the second question. The results will indicate whether there are alternatives to using all demographic or no demographic information, resulting in a better performance and fairer outcome.

6.2 Trade-off and Bias Analysis

Two types of analysis are done in this stage: trade-off and worst-case disparity analysis. Trade-off analysis involves comparing a given `WD` model to a `WOD` model of the same application and architecture. This comparison is done to understand whether an additional bias is introduced due to utilizing `DemographicFeatures` to train the `WD` model. In contrast, worst-case disparity analysis is intended to evaluate the bias of each model individually. This means that the bias of each model is calculated and reported separately, rather than in

Age	Gender	Race	Ethnicity
✗	✗	✗	✗
✓	✗	✗	✗
✗	✓	✗	✗
✗	✗	✓	✗
✗	✗	✗	✓
✓	✓	✗	✗
✓	✗	✓	✗
✓	✗	✗	✓
✗	✓	✓	✗
✗	✓	✗	✓
✗	✗	✓	✓
✓	✓	✓	✗
✓	✓	✗	✓
✓	✗	✓	✓
✗	✓	✓	✓
✓	✓	✓	✓

Table 6.1: Power set for the set of demographic features

relation to another model. After the models are trained in the first stage, such trade-off and worst-case disparity analysis could be done for any intended demographic group of patients.

For example, in the two test applications used to evaluate this framework, bias is studied across gender, ethnicity, insurance subgroups and additional intersectional subgroups. Subgroups are distinct categories within a broader group. For example, Medicaid and Medicare are subgroups examined within the insurance category. Intersectional subgroups, as explained in chapter 5, refer to patients who belong to different categories of demographic groups, simultaneously, like black female patients. Subgroups with fewer than 50 patients were removed from the analysis to avoid inflation of the resulting bias report.

6.2.1 Trade-off Analysis

As mentioned above, trade-off analysis aims to understand whether there is an additional bias due to the use of demographic information. This analysis is conducted and reported in two ways. First, the trade-off is assessed using the True Positive Rate (TPR) of the model's performance. Second, the trade-off analysis is reported using the original metric utilized by the model. The two mechanisms of reporting the analysis are explained in depth below:

TPR Disparity

Assuming a true positive prediction provides some benefit, for example, in the case of readmission, extended care, and IHM, extra attention to patients; this metric allows to document the increased or decreased benefit to a specific group due to the use of certain demographic information to train a model. This metric is important because it provides insight into whether the additional incorporation of demographic information evenly distributes the model's capability to identify positive instances accurately. To calculate this difference, we took:

$$WOD(TPR)_g - WD(TPR)_g$$

where g stands for a specific subgroup or intersectional subgroup in question.

If the difference is negative, then the sensitive attribute used to train the model benefited the subgroup being analyzed. On the other hand, if the difference is positive, then the sensitive attribute used to train the model disadvantaged the given group. Rarely, if the difference is zero, it means the sensitive attribute used to train the model neither benefited nor disadvantaged the analyzed group.

Original Metric Disparity

This reporting metric operates similarly to TPR disparity, except in this case, instead of using TPR, the original performance metric of the model is used to document the tradeoff. This means that if the performance metric used by the base model was accuracy, then accuracy would be used to report the performance difference, as different combinations of demographic features are used to train the model. To calculate this difference, we took:

$$WOD(metric)_g - WD(metric)_g$$

where g stands for a specific subgroup or intersectional subgroup in question, and metric stands for the specific metric the developer wants to test for.

The rationale behind using the original metric is to facilitate the generalizability of the framework. This method allows developers to understand how the different combinations of demographic information affect the model's fairness outcome in terms of the originally intended metric, in addition to the pre-specified fairness metric. For example, the mortality risk model reported its results using AUROC and AUPR. Even though we report the TPR disparity of the model in Chapter 8, AUROC and AUPR disparity are also reported to draw a full picture of the performance and disparity of the models for different demographic groups of patients.

6.2.2 Worst Case Analysis

The above analysis looks at the bias by comparing models that were trained with any demographic information to the one that did not. Although this method provides a great insight into what additional bias looks like by comparing it to a model without it, it does not show the bias of a given model on its own. To study that, we used the worst-case disparity.

As shown by Ghosh et al. [17], the worst-case min-max ratio is used to cover all potential subgroups within a data set by examining the most unfavorable outcome. Similarly, we use worst-case disparity min-max ratio, referred to as worst-case disparity in the rest of the paper, to show the highest level of inequity / bias that results within a group as a result of using a model. To calculate the worst-case disparity for a given group,

$$\frac{min\{P(\hat{Y}=1|A \in sg_i, Y=1)\forall_i \in N\}}{max\{P(\hat{Y}=1|A \in sg_i, Y=1)\forall_i \in N\}}$$

where $\hat{Y}$ is a binary predictor, A is a member of some given group sg and N is the number of subgroups within a given group.

The work by Ghosh et al. [17] uses this formula to investigate the worst-case disparity in intersectional subgroups. However, for our study, we will utilize it to analyze general groups, as the number of patients in each intersectional subgroup might be low, potentially inflating the results. Providing protection to smaller numbers of people in a subgroup is among the recommendations for future work discussed in Chapter 10. In this study, the farther away the ratio is from one, the higher the worst-case disparity among a given group. Therefore, the higher the ratio is, the lower the disparity, and the lower the ratio, the higher the disparity.

This study enables us to observe the initial worst-case bias in the model trained without demographic data, as well as each of the other models separately. While it is an effective tool for assessing disparity, it should be used in conjunction with trade-off analysis to ensure the accuracy of the results. This is because even if the performance of a particular model decreases, as long as it decreases for all subgroups in the study, the worst-case disparity will decrease (the ratio will increase). For instance, if a model has a TPR of [50,80,90] for three subgroups A, B, and C, and another model has a TPR of [30,30,30], then the worst-case disparity of the first model will be .56 while the second one will have a ratio of 1, even though the second model has significantly lower performance across all three groups. Therefore, examining it in isolation might cause us to overlook this aspect. If, when viewed alongside the trade-off analysis, it is found that the models under consideration are enhancing the benefits of the subgroups being studied, then the worst-case analysis becomes useful to ensure that no groups are severely disadvantaged and left behind.

The computation of the worst-case analysis is solely based on the True Positive Rate (TPR), rather than incorporating both the TPR and the original metric as done in trade-off analysis. This approach is adopted because the worst-case disparity focuses on the disparity in benefits, utilizing the Equal Opportunity fairness metric as outlined by [17]; [20]. The equal opportunity metric seeks to ensure that the positive rates among different groups are equal, provided that the individuals in each group meet the qualifications. The formula for this metric is as follows:

$$P(\hat{Y} = 1 | A \in sg_i, Y = 1) = P(\hat{Y} = 1 | A \in sg_j, Y = 1) \forall_{i,j} \in N, i \neq j$$

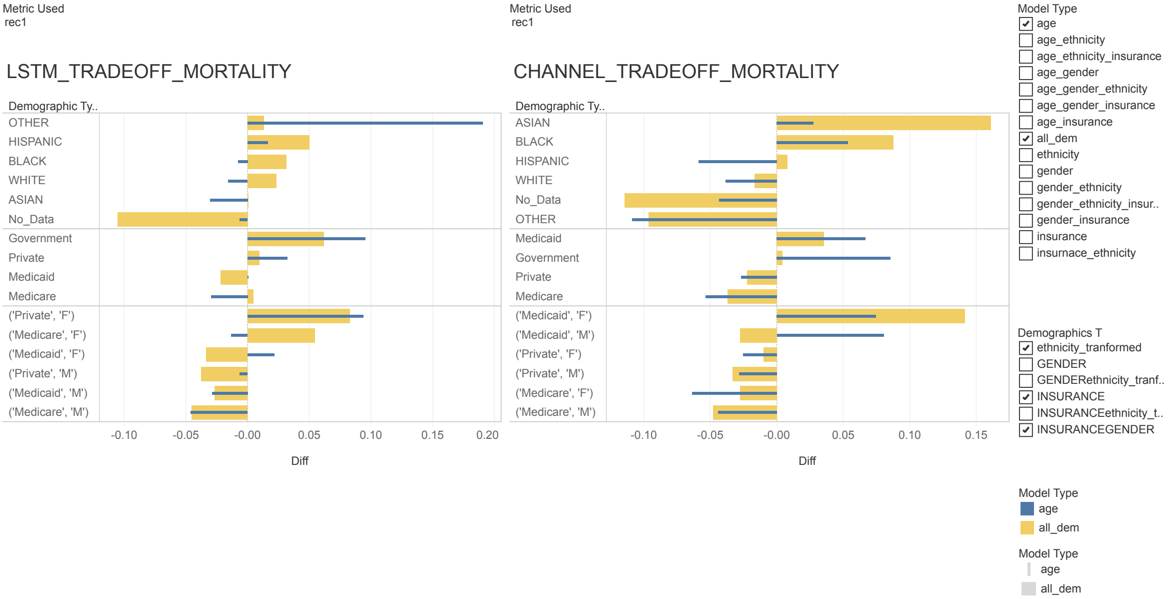
where $\hat{Y}$ is a binary predictor and A is a member of some given group sg.

Chapter 7: DASHBOARD INTRODUCTION FOR TRADE-OFF AND BIAS ANALYSIS

To better understand the trade-off and bias analysis patterns, we built an interactive Tableau dashboard for visualization of the results. This dashboard showcases the comparison of performance-bias trade-offs across various models and architectures, simplifying the comparison process. Although the decision on which demographic information to utilize may not always be straightforward, such a dashboard streamlines the examination of patterns and aids in decision-making.

For the two base models used to test the framework in Chapter 6, we created two side-by-side plots, one for each application, making it a total of four plots per application. We refer to each side-by-side plot as a dashboard. The first dashboard is for trade-off analysis, and each plot pertains to the type of architecture used to train the model. A sample trade-off analysis dashboard is shown in Fig. 7.1 and thoroughly explained in the subsequent sections. In contrast, the second dashboard shows the worst-case analysis outcome, with one plot for each architecture. A sample dashboard is illustrated in Fig. 7.2 and detailed explanations are provided in the upcoming sections.

The dashboards built for analyzing the two test case models are provided in Chapter 7.3.



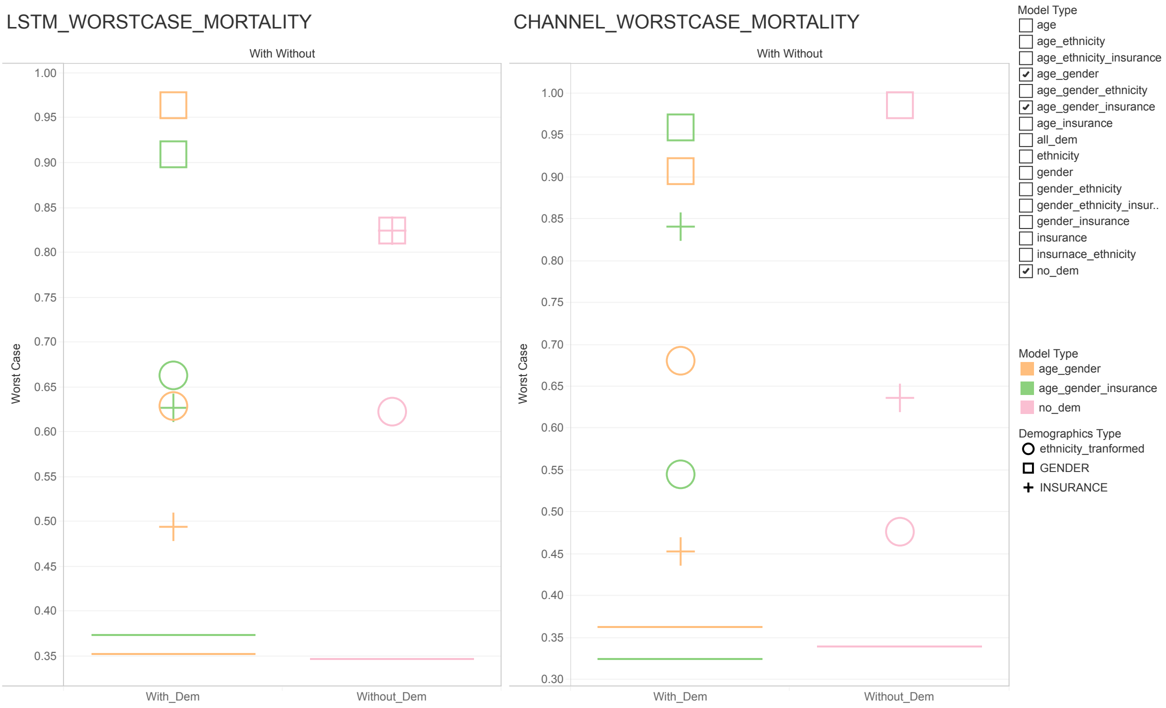
Model Type
☑ age
☐ age_ethnicity
☐ age_ethnicity_insurance
☐ age_gender
☐ age_gender_ethnicity
☐ age_gender_insurance
☐ age_insurance
☑ all_dem
☐ ethnicity
☐ gender
☐ gender_ethnicity
☐ gender_ethnicity_insur..
☐ gender_insurance
☐ insurance
☐ insurnace_ethnicity

Demographics T
☑ ethnicity_tranformed
☐ GENDER
☐ GENDERethnicity_tranf..
☑ INSURANCE
☐ INSURANCEethnicity_t..
☑ INSURANCEGENDER

Figure 7.1: Screeshot of trade-off analysis dashboard

7.1 Trade-off Analysis Dashboard

The purpose of this dashboard is to compare the results of the trade-off analysis of different models to effectively assess the additional bias introduced by each model. This facilitates the identification of models with the least amount of additional bias introduced. Fig. 7.3 is used to explain the various elements of the dashboard, focusing on one plot of the dashboard shown in Fig. 7.1. Although the explanation pertains to one plot of the two, the same interpretation applies to the other one as well. Fig. 7.3 shows the trade-off analysis for mortality risk using the channel LSTM model. In this figure, A denotes the model architecture used to train the model and the application for which the model is trained, while B, model type, denotes the type of demographic information used to train the model. The



Model Type
☐ age
☐ age_ethnicity
☐ age_ethnicity_insurance
☑ age_gender
☐ age_gender_ethnicity
☑ age_gender_insurance
☐ age_insurance
☐ all_dem
☐ ethnicity
☐ gender
☐ gender_ethnicity
☐ gender_ethnicity_insur..
☐ gender_insurance
☐ insurance
☐ insurnace_ethnicity
☑ no_dem

Figure 7.2: Screenshot of worst-case disparity analysis dashboard

different types of demographic information used as features are connected with an underscore. For instance, if the model type is group1_group2, then group1 and group2 information were used as features to train this model. The only exception to this is ‘all_dem’, which means all demographic information was used for model training. For the two test cases in Chapter 3, it means all age, gender, ethnicity, and insurance features were used to train the models. C denotes the type of demographic group of patients being studied. ‘ethnicity_transformed’ refers to the ethnicity of patients, but it is named as is because multiple ethnicities had to be consolidated to give out the five subgroups, Asian, Black, Hispanic, White, No_data and Other, used for this study. Two back-to-back ethnicities show the study of an intersectional group of patients. For instance, ’INSURANCEGENDER’ is the intersectional subgroup of insurance and gender; some subgroups are female patients with private insurance and female patients with medicare insurance.

In sections B and C, users can compare the various types of demographic features used to train the models, as well as the different demographic groups intended for study. This can be done by selecting the specific features and groups they wish to analyze by toggling the check boxes associated with each. The different demographic features used to train the models selected from section B can be seen color-coordinated in section D, whereas the different demographic groups studied in section C add or remove additional elements in the y-axis to show the corresponding outputs. For studying TPR disparity, users can select ”Rec1” from the dropdown menu depicted as E in fig. 7.3. Alternatively, for the original metric, all necessary metrics required to report the results are available in the drop-down menu.

As mentioned above, the Y-axis shows the subgroup of the demographic information being studied, which is arranged per demographic category so that subgroups from the same category appear together. When a user clicks on a toggle for a group to be added or removed from section C, the resulting group is accordingly included or excluded from the Y-axis. The X-axis shows the trade-off disparity for each subgroup being studied.

For example, in Figure 7.3, the dashboard pertains to a mortality risk prediction application using a channel wise LSTM model architecture and the analysis focuses on TPR disparity. The blue bar represents a model trained with age as an additional input feature, while the yellow bar represents a model trained with all demographic features as input. The demographic groups under study include ethnicity, insurance, and an intersectional group between insurance and gender.

7.2 Worst Case Disparity Analysis Dashboard

This dashboard illustrates the outcome of the worst-case analysis. Fig. 7.4 is a zoomed-in version of Fig. 7.2 used to showcase the different elements of the plot. Similar to the previous section, while only one plot is utilized to explain the dashboard, the same interpretation applies to both plots on the worst-case analysis dashboard.

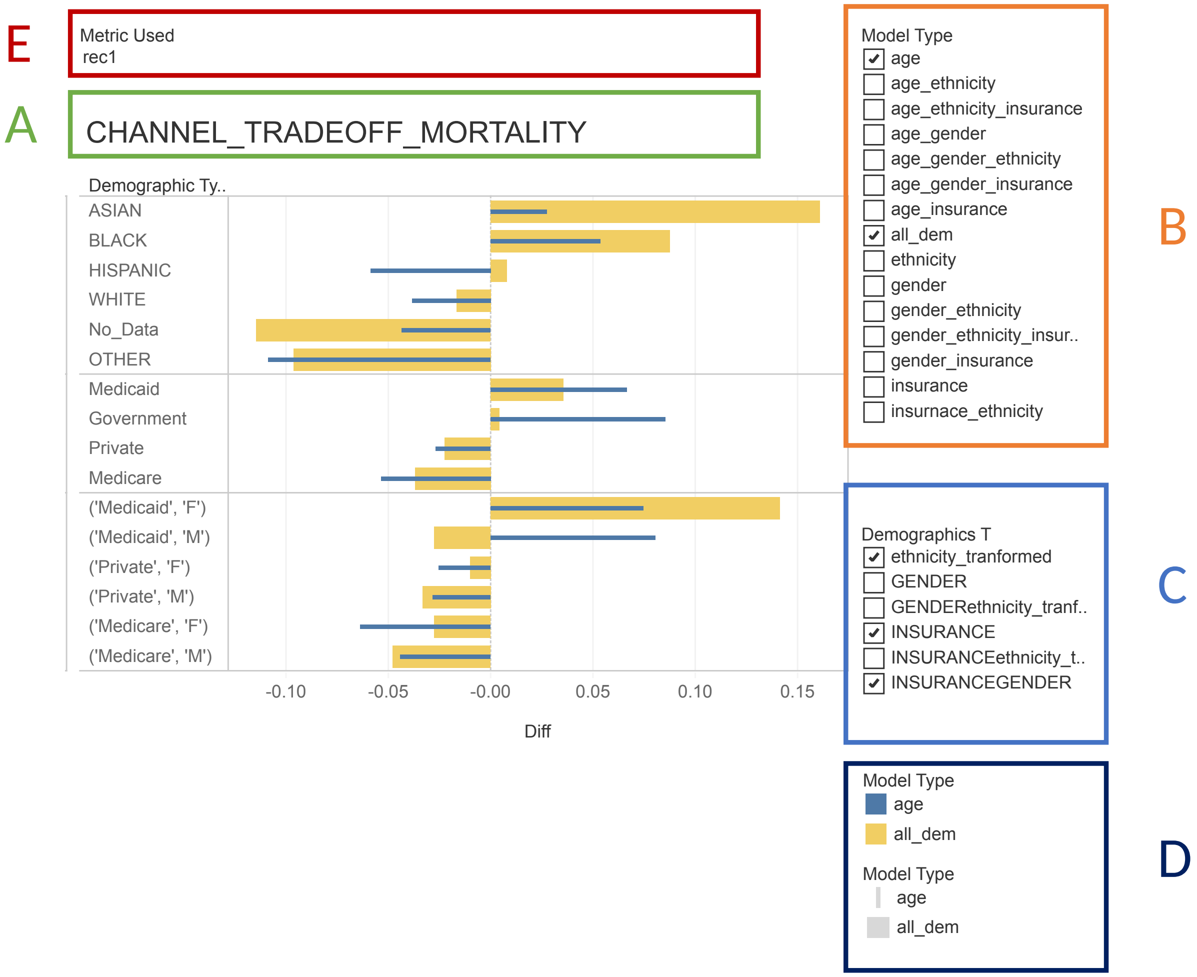
Section F of the plot displays the model architecture used to train the model and the application for which the model is trained. Section G allows users to examine models trained with different demographic information, with color-coordinated outcomes, as shown in section H. A detailed explanation of section G is provided in the preceding section, where the section is presented as section B. The only additional element is the type of model 'no_dem', which refers to the model trained without any demographic information. As mentioned in the previous chapter, the worst-case analysis is solely conducted for ethnicity, gender, and insurance groups, excluding intersectional groups to prevent result inflation. The different types of demographics studied are presented in section I.

The Y-axis represents the worst-case disparity ratio of the models, while the X-axis displays the outcomes of different models. The ”Without Dem” column shows the output of the model trained without demographic features, labeled as “No Dem” in section G. The “With Dem” column shows every model trained with demographic features, enabling side-by-side comparison of models with and without demographic features, as well as comparison of different models with demographic features on a single axis. The color coded horizontal line seen on the plot corresponds to the TPR of the model as a whole. For example, in Fig. 7.4, the outcomes of models ‘age_gender’, ‘age_gender_insurance’, and ‘no_demographic’ are compared.

7.3 Dashboard Availability

The dashboards described in the previous sections are hyperlinked below.

- Readmission risk trade-off analysis
- Readmission risk worst-case disparity analysis
- IHM risk trade-off analysis
- IHM risk worst-case disparity analysis



Metric Used
rec1

CHANNEL_TRADEOFF_MORTALITY

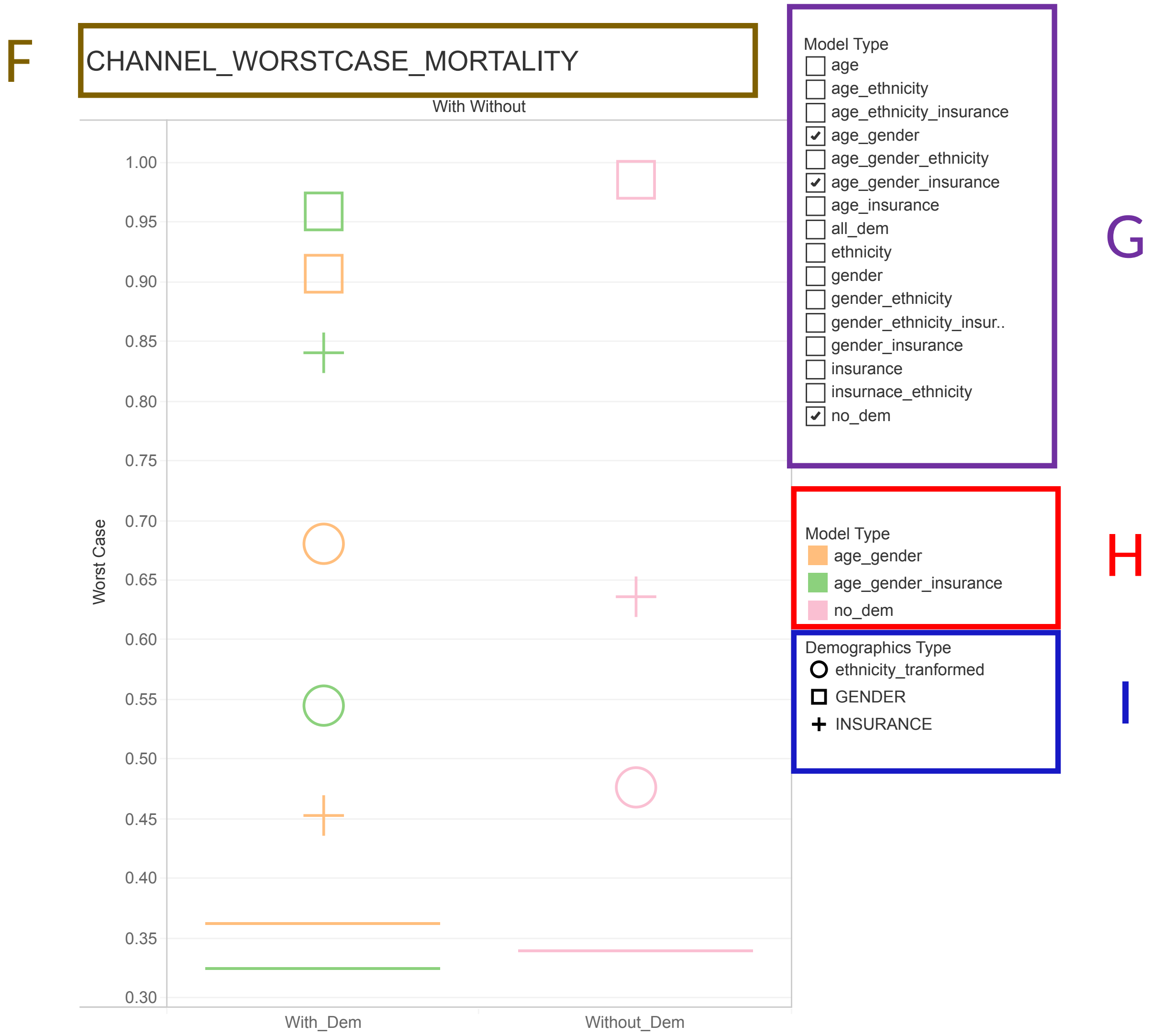
Model Type
☑ age
□ age_ethnicity
□ age_ethnicity_insurance
□ age_gender
□ age_gender_ethnicity
□ age_gender_insurance
□ age_insurance
☑ all_dem
□ ethnicity
□ gender
□ gender_ethnicity
□ gender_ethnicity_insur..
□ gender_insurance
□ insurance
□ insurnace_ethnicity

Demographics T
☑ ethnicity_tranformed
□ GENDER
□ GENDERethnicity_tranf..
☑ INSURANCE
□ INSURANCEethnicity_t..
☑ INSURANCEGENDER

Model Type
🟦 age
🟨 all_dem

Model Type
▏ age
⬜ all_dem

Figure 7.3: Screenshot of mortality application trade-off analysis dashboard where A: model architecture + analysis type+ application type, B: model type, C: demographic category being studied, D: bar color for the model selected in section B, E: metric type.



CHANNEL_WORSTCASE_MORTALITY

Model Type
- ☐ age
- ☐ age_ethnicity
- ☐ age_ethnicity_insurance
- ☑ age_gender
- ☐ age_gender_ethnicity
- ☑ age_gender_insurance
- ☐ age_insurance
- ☐ all_dem
- ☐ ethnicity
- ☐ gender
- ☐ gender_ethnicity
- ☐ gender_ethnicity_insur..
- ☐ gender_insurance
- ☐ insurance
- ☐ insurnace_ethnicity
- ☑ no_dem

Figure 7.4: Screenshot of mortality application worst-case disparity analysis dashboard where F: model architecture + analysis type+ application type, G: model type, H: bar color for the model selected in section G, I: demographic group being studied

Chapter 8: SAMPLE INTERPRETATIONS OF KEY RESULTS FROM TRADE-OFF AND BIAS ANALYSIS

The framework introduced in Chapter 6 was tested using the two base models outlined in Chapter 3. As detailed in Chapter 4, the model training utilized four demographic information groups: gender, ethnicity, insurance, and age. Consequently, 16 models were trained for each architecture and application. Given that there are two applications, each with two model architectures, a total of 64 models were trained and analyzed. The bias and trade-off analysis were conducted for three out of the four mentioned groups: gender, ethnicity, and insurance.

The two types of analysis, tradeoff analysis, and worst-case disparity analysis, assess different kinds of biases. It is essential to consider the findings of both analyses when deciding which demographic information to incorporate. The trade-off analysis dashboard primarily shows two types of biases. First, when the result is negative for some groups being studied, and positive for others, it shows that there is varying benefit across these groups. This is observed when the bar plot is to the right of the center for some and to the left for others, with different magnitudes. Second, when there is a noticeable gap between the magnitude of results across different subgroups studied, that also indicates varying benefits from the use of a given demographic information. In contrast, for the worst-case disparity dashboards, the further the ratio is from one, the higher the disparity within the group being studied.

Returning to the question of when to utilize demographic information for model training, ideally, it would be when there is a collective increase in benefit across all subgroups being studied along with a lowered worst-case disparity within each group studied. However, as shown by the results of the IHM model and the readmission risk model, this ideal scenario may not always happen. In that case, trade-off tolerance and the specific application determine the acceptable trade-off for a given increase in performance.

When interpreting the dashboards to decide which demographic information to use for model training, it is important to understand that the decision is not always clear-cut. It is also crucial to note that not all demographic information will consistently benefit all subgroups being studied. In the following sections, we present various models that yield intriguing results and patterns. Given the large number of models (64 in total), documenting the analysis of each one is impractical. Instead, we present select results that we found particularly interesting as they provide insight into the decision-making process. As stated above, decision-making is not a linear process and depends on the trade-off tolerance of each application and developer. However, these interpretations illustrate the thought process involved in making these decisions. The results in the following subsections are presented for TPR disparity and are reported across different architectures. The same analysis can be applied to understand the varying outcomes across diverse model architectures and outcome reporting metrics.

8.1 Readmission Risk - LSTM Architecture

Figures 8.1 and 8.2 present key findings of the readmission risk model using the LSTM architecture. As detailed in Chapter 5 and Figure 8.1, incorporating all demographic features increased benefits for most, but not all, groups. In contrast, alternative models such as the 'insurance_ethnicity' model resulted in greater benefits for larger number of groups. Furthermore, this model showed a smaller decrease in benefits for the three groups it did not benefit: Asian males, white patients with Medicaid insurance, and Hispanic patients with Medicare. The benefits for these three groups decreased by 1%, 1%, and 2%, respectively, a decrease much smaller than that observed in the 'all_demographic' model for the groups this model did not benefit.

The worst-case disparity plot (Figure 8.2) demonstrates that the 'all_demographic' and 'insurance_ethnicity' models reduce the worst-case disparity across all three groups compared to the `WOD` model. A comparison between the 'all demographic' and 'insurance ethnicity' models shows that the 'insurance_ethnicity' model reduces the disparity better for gender and insurance groups. In contrast, the 'all_demographic' model is more effective in reducing worst-case disparity for ethnicity.

Upon reviewing the trade-off analysis and worst case disparity plots for this architecture, it can be observed that there is a better alternative to using either no demographic information or all demographic information. This alternative improves the benefit for nearly all groups, and for the groups it did not benefit, the maximum benefit reduction was 2%. Furthermore, it effectively mitigates the worst disparity compared to both the no demographic information and all demographic information models. Additionally, this model also exhibits a higher True Positive Rate (TPR) than both the 'no_demographic' and 'all_demographic' models.

This plot also prompts an intriguing question about the relationship between ethnicity, insurance, and patient readmission risk and why insurance and patient ethnicity are better predictors in this model. It calls for further investigation into potential systemic bias in the healthcare system, which is further discussed in Chapter 10.

8.2 Readmission Risk - LSTM_CNN Architecture

The models 'gender insurance', 'insurance', and 'age_gender_insurance' are particularly noteworthy, as shown in Figures 8.3 and 8.4. The maximum decrease in benefit is 15% for the 'insurance' model across government insured male patients, 22% for the 'gender_insurance' model across government insured white patients and 7% for 'age_gender_insurance' model across government insured white patients, all compared to the WOD model.

In this situation, where no model presents a clear benefit, it becomes essential to consider the trade-offs. For example, the 'gender_insurance' model has fewer groups experiencing reduced benefits, with some of these reductions $\geq$ 10%. In contrast, the 'age_gender_insurance' model has a greater number of groups with decreased benefits, although the extent of this reduction is typically 5% or less.

Upon examining the worst-case disparity plots, it can be seen that the True Positive Rate (TPR) for 'insurance' and 'gender_insurance' models is higher than 'age_gender_insurance', which is nearly equivalent to WOD model. However, the worst-case disparity within the insurance category is greater for 'insurance' and 'gender_insurance' models compared to the WOD model. Meanwhile the 'age_gender_ethnicity' model shows roughly similar levels of disparity within the insurance category to the WOD model. The worst-case disparity of all three models across ethnicity and gender is close to the WOD model. Therefore, in this context, the decision on which demographic to use would depend on the developer's priorities, such as the maximum number of groups with decreased benefit, the maximum benefit reduction, or the lowest worst-case disparity introduced.

8.3 Mortality Risk - LSTM Architecture

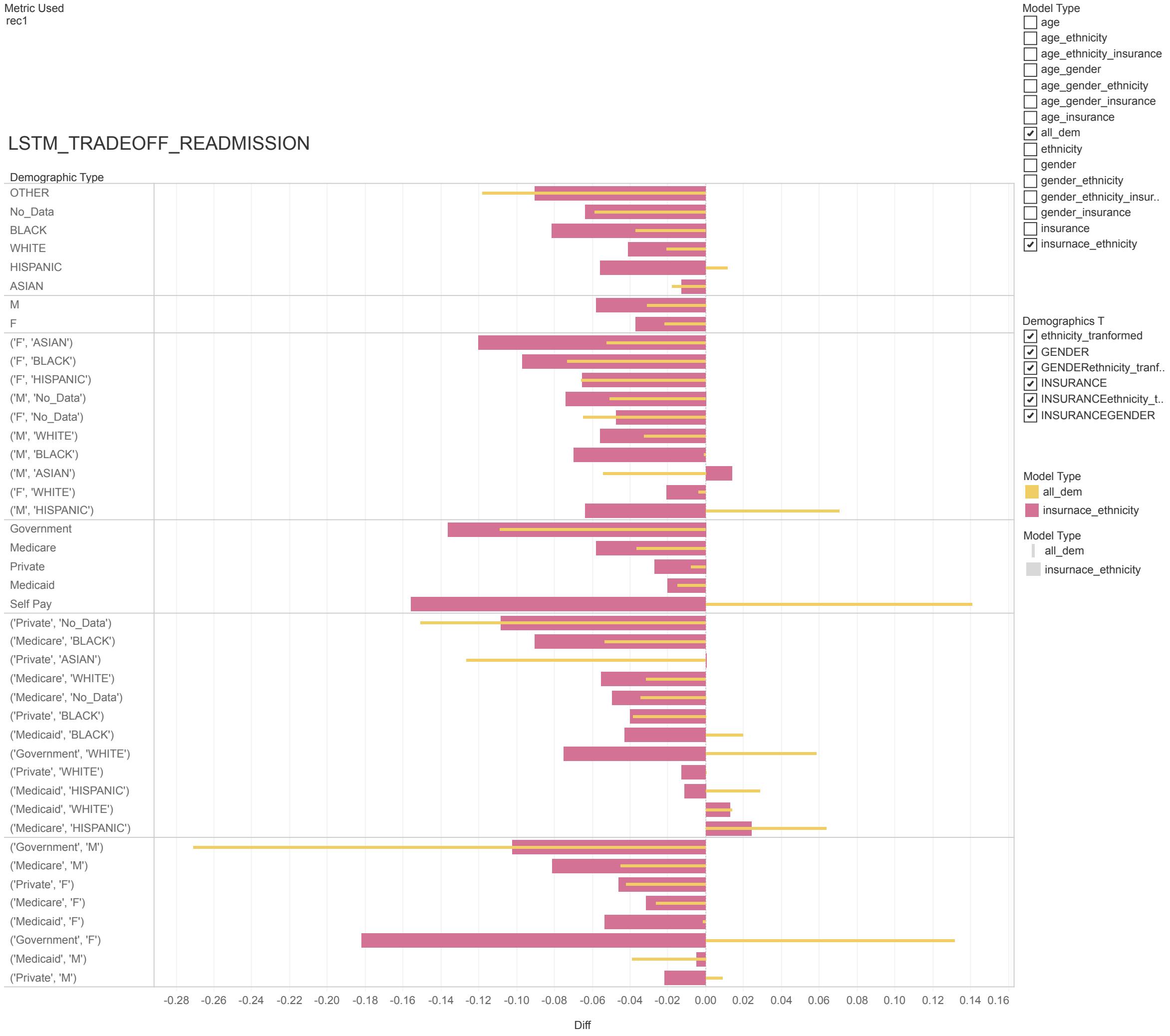
The models of 'gender_ethnicity' and 'gender_insurance' are particularly intriguing. When examining the worst-case disparity plot (referenced as fig 8.6), it can be seen that the 'gender_insurance' model reduces disparity more effectively than the 'gender_ethnicity' model across ethnicity, and the difference between the two models across gender and insurance groups is not substantial. However, a closer look at the trade-off analysis plot (fig 8.5) for these two models reveals a different pattern: the 'gender_ethnicity' model tends to disadvantage most of the groups studied, while the 'gender_insurance' model tends to advantage them. This insight cannot be learned from the worst-case disparity analysis plot alone. Therefore, as previously stated, it is crucial to consider both plots together when analyzing and deciding which demographic information to utilize.

8.4 Mortality Risk - Channel LSTM Architecture

The most notable result of this architecture is the consistent pattern of additional bias displayed by most of the models. Despite being trained with various combinations of demographic information, these models consistently alter advantages for the same demographic groups. As shown in Figure 8.7, all models reduce the benefits for Asian and black patients to varying degrees. It is also observable that all models reduce the benefits for patients with Medicaid insurance, black female patients, and Medicaid insured female patients.

This unanimous bias is a unique outcome, unseen in other applications and architectures, especially when compared to the mortality application model with LSTM architecture. Compared to this model, since the only variable changed is the training model, further investigation is required into why this model architecture exhibited this kind of performance bias. This pattern could indicate the existence of algorithmic or systemic bias, which calls for a comprehensive analysis. Figure 8.7 is designed to illustrate this consistent bias pattern. However, due to the selection of numerous models, understanding each one might be challenging.

As illustrated in Figure 8.7, this is an instance where no single model consistently errs towards benefit or disadvantage. All these models provide advantages to some groups and not to others. Therefore, despite potential differences in tolerance for the selection of demographic information, this might be a case where refraining from using any will prevent the introduction of additional bias into the model.

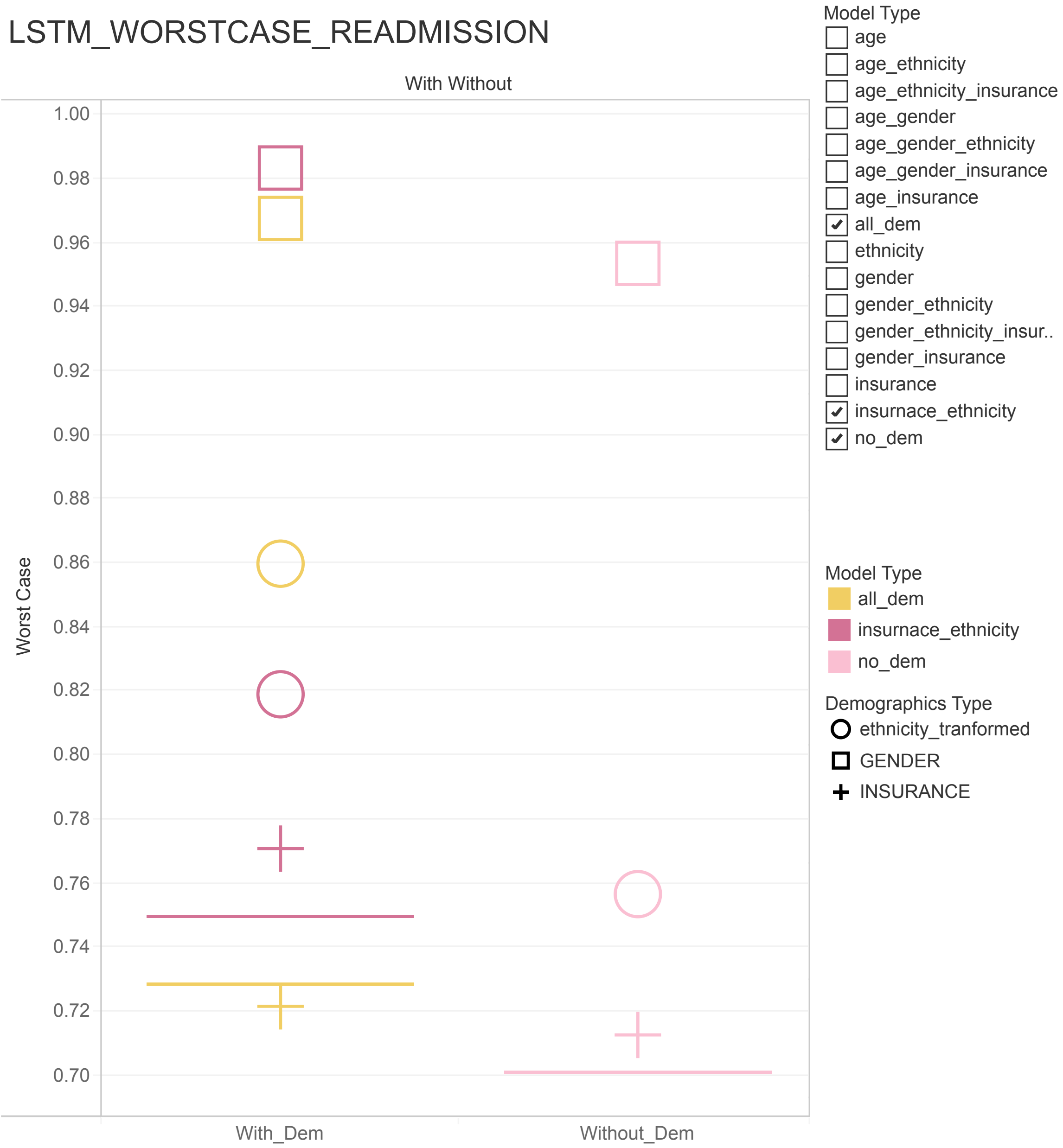


Metric Used
rec1

Model Type
☐ age
☐ age_ethnicity
☐ age_ethnicity_insurance
☐ age_gender
☐ age_gender_ethnicity
☐ age_gender_insurance
☐ age_insurance
☑ all_dem
☐ ethnicity
☐ gender
☐ gender_ethnicity
☐ gender_ethnicity_insur..
☐ gender_insurance
☐ insurance
☑ insurnace_ethnicity

Demographics T
☑ ethnicity_tranformed
☑ GENDER
☑ GENDERethnicity_tranf..
☑ INSURANCE
☑ INSURANCEethnicity_t..
☑ INSURANCEGENDER

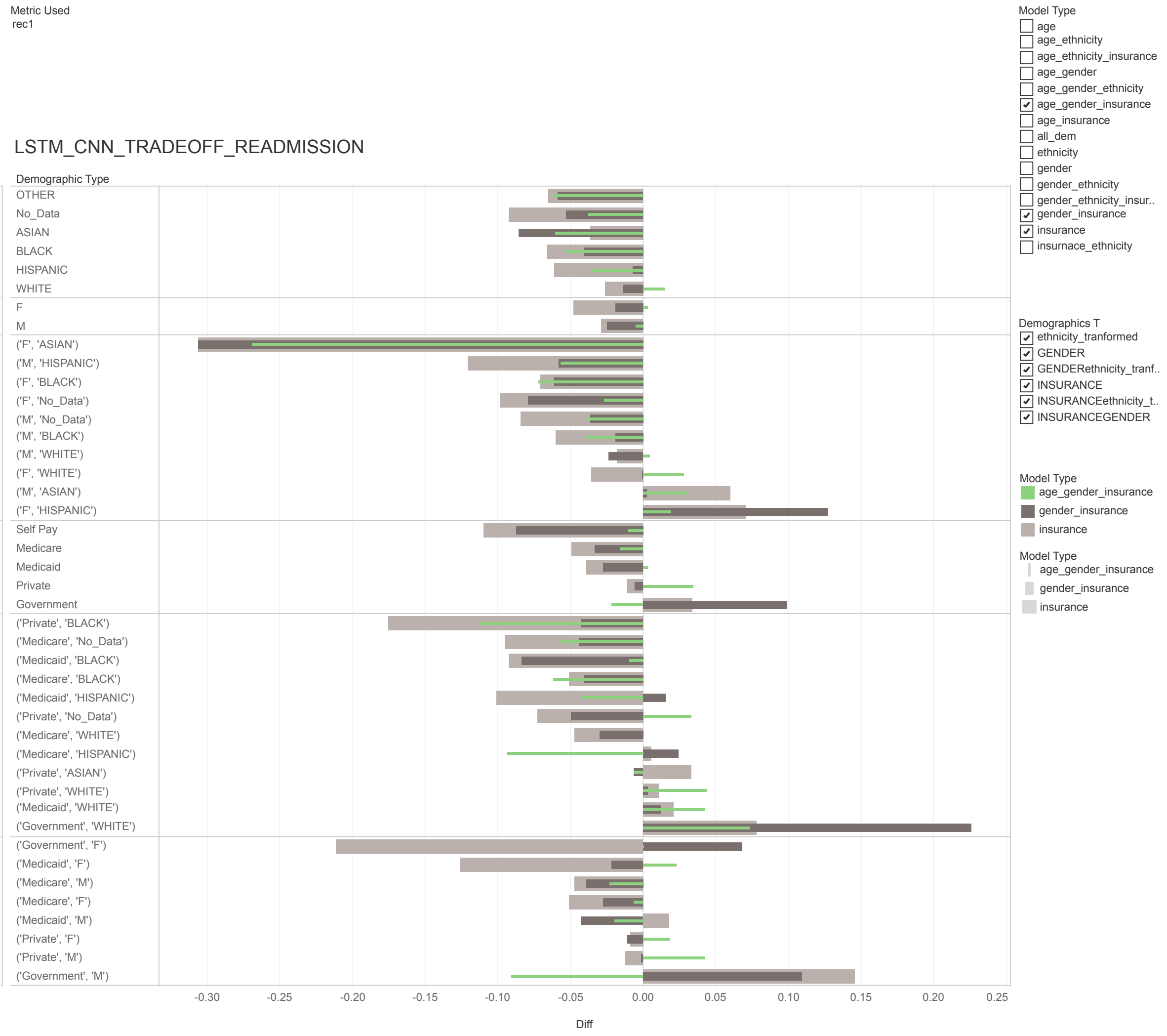
Figure 8.1: Screenshot of readmission risk application trade-off analysis dashboard showing model type ‘all_dem’ and ‘insurance_ethnicity’



Model Type

- □ age
- □ age_ethnicity
- □ age_ethnicity_insurance
- □ age_gender
- □ age_gender_ethnicity
- □ age_gender_insurance
- □ age_insurance
- ☑ all_dem
- □ ethnicity
- □ gender
- □ gender_ethnicity
- □ gender_ethnicity_insur..
- □ gender_insurance
- □ insurance
- ☑ insurnace_ethnicity
- ☑ no_dem

Figure 8.2: Screenshot of readmission risk application worst-case disparity analysis dashboard showing model type 'all_dem' and 'insurance_ethnicity'

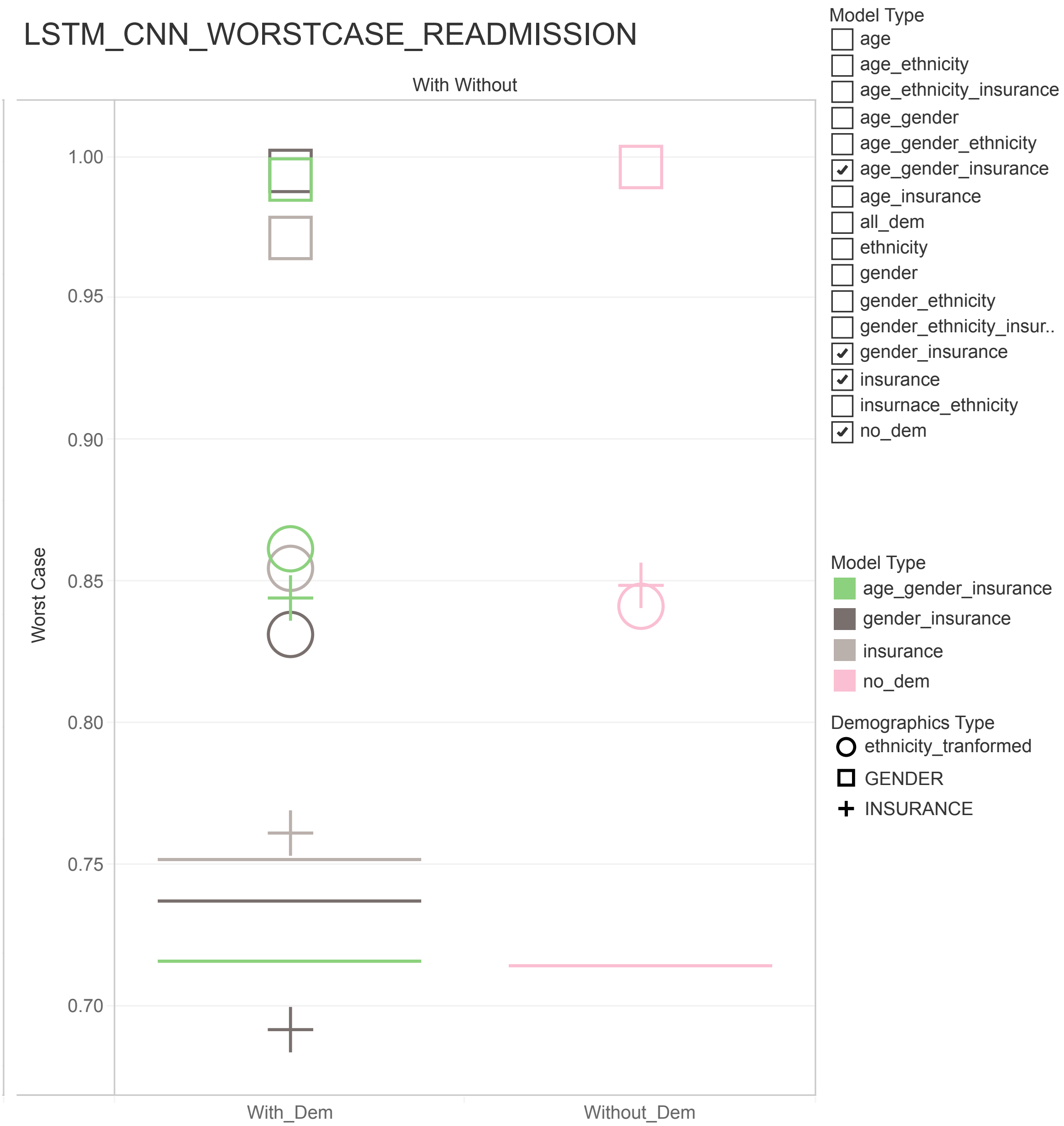


Metric Used
rec1

Model Type
□ age
□ age_ethnicity
□ age_ethnicity_insurance
□ age_gender
□ age_gender_ethnicity
☑ age_gender_insurance
□ age_insurance
□ all_dem
□ ethnicity
□ gender
□ gender_ethnicity
□ gender_ethnicity_insur..
☑ gender_insurance
☑ insurance
□ insurnace_ethnicity

Demographics T
☑ ethnicity_tranformed
☑ GENDER
☑ GENDERethnicity_tranf..
☑ INSURANCE
☑ INSURANCEethnicity_t..
☑ INSURANCEGENDER

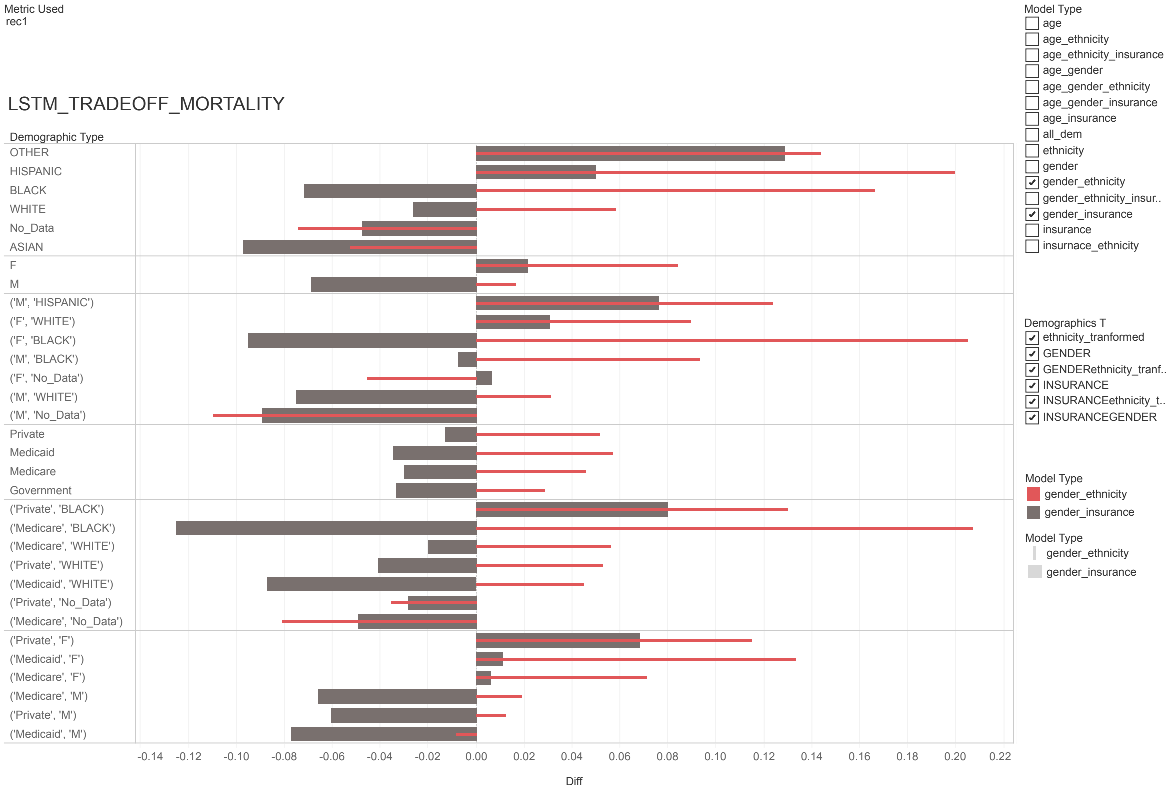
Figure 8.3: Screenshot of trade-off analysis plot for readmission risk application using LSTM_CNN architecture



Model Type

☐ age
☐ age_ethnicity
☐ age_ethnicity_insurance
☐ age_gender
☐ age_gender_ethnicity
☑ age_gender_insurance
☐ age_insurance
☐ all_dem
☐ ethnicity
☐ gender
☐ gender_ethnicity
☐ gender_ethnicity_insur..
☑ gender_insurance
☑ insurance
☐ insurnace_ethnicity
☑ no_dem

Figure 8.4: Screenshot of worst-case disparity analysis plot for readmission risk application using LSTM_CNN architecture



Metric Used
rec1

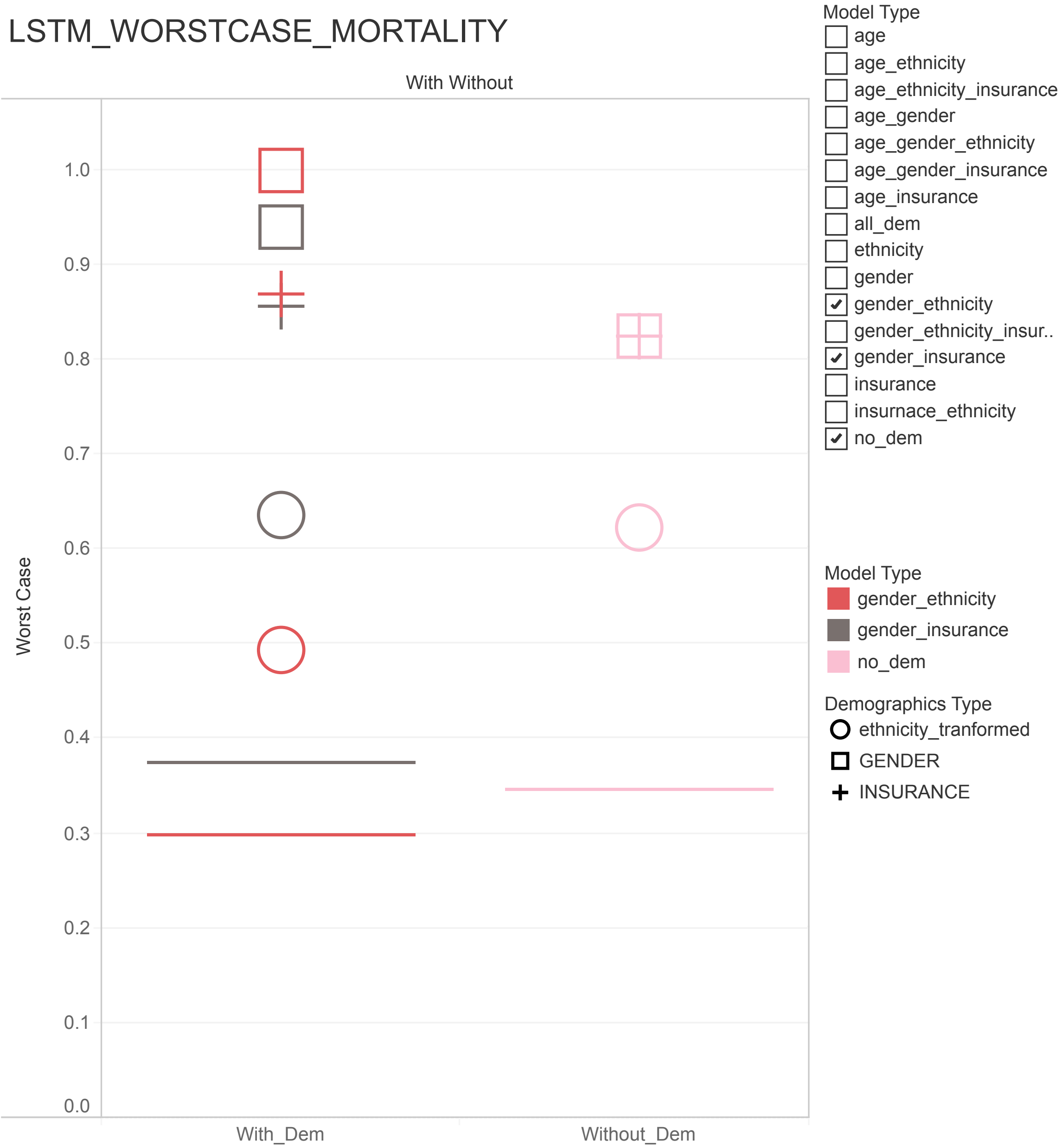
Model Type
☐ age
☐ age_ethnicity
☐ age_ethnicity_insurance
☐ age_gender
☐ age_gender_ethnicity
☐ age_gender_insurance
☐ age_insurance
☐ all_dem
☐ ethnicity
☐ gender
☑ gender_ethnicity
☐ gender_ethnicity_insur..
☑ gender_insurance
☐ insurance
☐ insurnace_ethnicity

Demographics T
☑ ethnicity_tranformed
☑ GENDER
☑ GENDERethnicity_tranf..
☑ INSURANCE
☑ INSURANCEethnicity_t..
☑ INSURANCEGENDER

Model Type
🟥 gender_ethnicity
▨ gender_insurance

Model Type
gender_ethnicity
gender_insurance

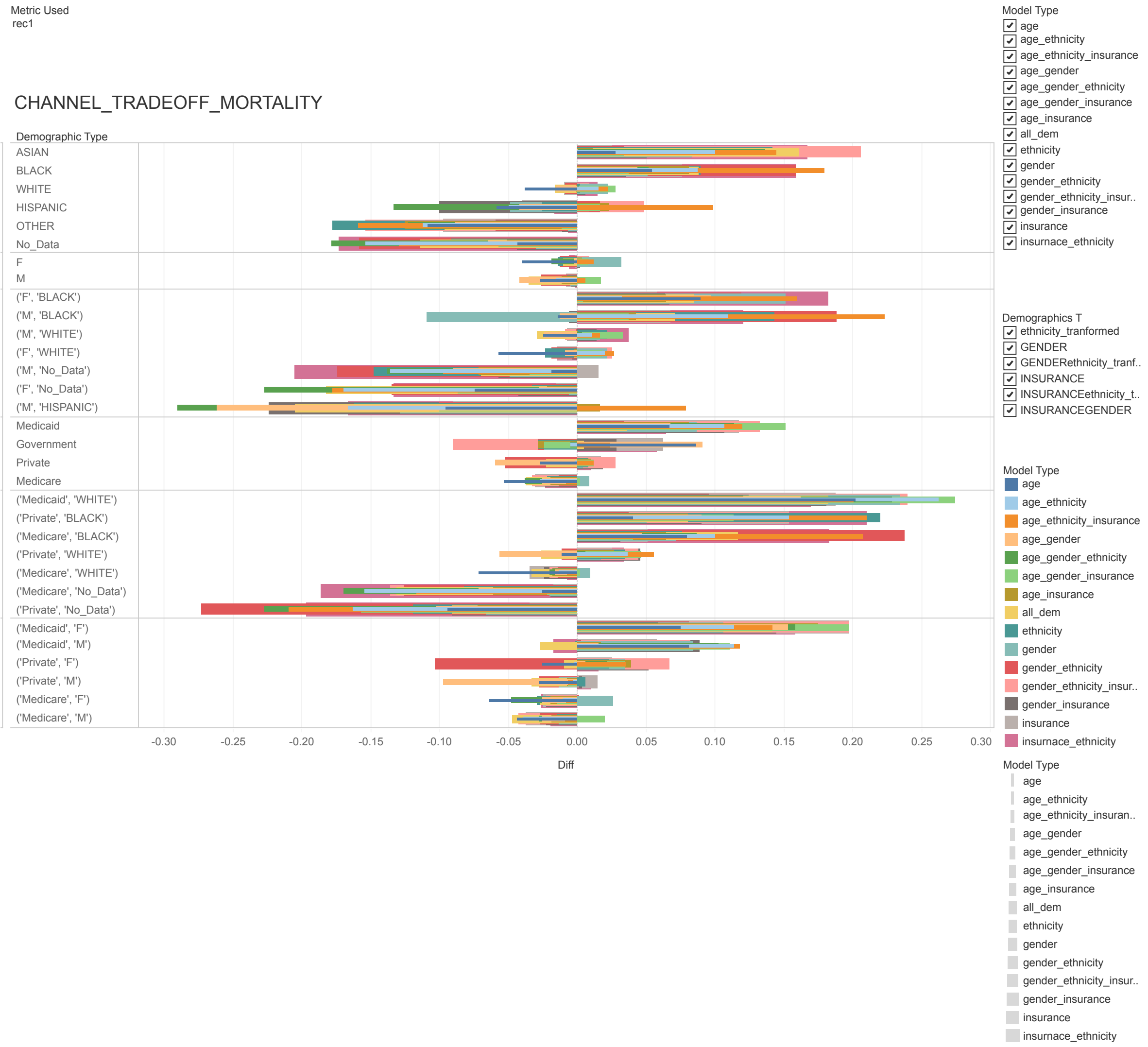
Figure 8.5: Screenshot of trade-off analysis plot for mortality risk application using LSTM architecture



Model Type

☐ age
☐ age_ethnicity
☐ age_ethnicity_insurance
☐ age_gender
☐ age_gender_ethnicity
☐ age_gender_insurance
☐ age_insurance
☐ all_dem
☐ ethnicity
☐ gender
☑ gender_ethnicity
☐ gender_ethnicity_insur..
☑ gender_insurance
☐ insurance
☐ insurnace_ethnicity
☑ no_dem

Figure 8.6: Screenshot of worst-case disparity analysis plot for mortality risk application using LSTM architecture



Metric Used
rec1

Model Type
☑ age
☑ age_ethnicity
☑ age_ethnicity_insurance
☑ age_gender
☑ age_gender_ethnicity
☑ age_gender_insurance
☑ age_insurance
☑ all_dem
☑ ethnicity
☑ gender
☑ gender_ethnicity
☑ gender_ethnicity_insur..
☑ gender_insurance
☑ insurance
☑ insurnace_ethnicity

Demographics T
☑ ethnicity_tranformed
☑ GENDER
☑ GENDERethnicity_tranf..
☑ INSURANCE
☑ INSURANCEethnicity_t..
☑ INSURANCEGENDER

Figure 8.7: Screenshot of trade-off analysis plot for mortality risk application using channel-wise LSTM architecture

Chapter 9: MAIN TAKEAWAY

The dashboards discussed in the previous chapter demonstrate the versatility of this framework. These dashboards can be used to highlight various aspects, such as the starting bias in a model trained without demographic data, and to compare the trade-offs between different models. The choice of model depends on the priority of the developers and the requirements of the application. However, this framework equips users with the necessary tools to review and make an informed decision. Furthermore, it also identifies potential systemic biases that may manifest in the output.

This section discusses the three main takeaways from the experiments conducted in the preceding sections, which involved training 64 models.

Higher performance doesn't always correlate with fair

One of the most significant insights gained from the above experiments is the understanding that better performance does not necessarily equate to a fairer model. This underscores the significance of performing trade-off and bias analysis, even after achieving improved model performance through adjustments to features or parameters. In some cases, although high-performing models may demonstrate a higher disparity ratio (low disparity) across certain groups, this consistency may not extend to all groups under study. Hence, it is crucial to conduct the analysis across all relevant groups.

For instance, in Fig. 9.1, the worst-case analysis outcome of the readmission risk model with LSTM architecture is showcased. It can be seen that the model trained with insurance features has a higher True Positive Rate (TPR) compared to the model trained with additional ethnicity feature. However, it also has the lowest proportion of worst-case disparity across all demographic groups studied compared to the model trained with ethnicity. This highlights the complexity of balancing performance and fairness considerations in model development.

Model architecture and application matter

Another significant insight gained from these experiments is that a demographic group utilized as input for a particular application, resulting in a fairer output compared to alternatives, may not yield the same outcome when assessed in other healthcare applications or even different architectures within the same application. While it may appear tempting to extrapolate the results of one analysis to another application or architecture, caution must be taken to avoid wrong conclusions.

While this trend is observable across many more dashboards, Fig. 9.2 is an illustrative example. This figure compares models trained with age and age_insurance for the readmission risk application. Age decreases the disparity across ethnicity and gender compared to age_insurance for the LSTM architecture but seems to exhibit the opposite trend for the LSTM_CNN architecture. Conversely, age_insurance appears to better reduce the disparity across insurance compared to the age model for LSTM but demonstrates the opposite trend for the LSTM_CNN model.

WOD model is not bias-free

The two most significant takeaways regarding the trade-off between using no, some, or all demographic information are:

First, not using demographic information for model training does not equate to bias-free outcomes. This is evident from the four worst-case analysis plots shown in fig. 9.4 and 9.3, each illustrating the worst-case bias introduced by each model individually. Fig. 9.4 and 9.3 show that omitting demographic information from model training ("Fairness through Blindness") does not guarantee bias-free outcomes. None of the worst-case disparity plots show a perfect 1.0 worst-case disparity ratio, demonstrating that models that do not incorporate demographic data are not free of bias. This insight is vital for both ML and healthcare professionals. For healthcare professionals, knowing this helps to understand that the models deployed in the healthcare sector today may not be fair to everyone. Therefore, human intervention is cruicial when using these systems. For ML professionals, these outcomes show that training models with or without demographic information might result in bias to some groups. Therefore, it is important to analyze and document the bias before the deployment of any healthcare models. Second, using demographic information doesn't always guarantee better-performing and fairer models. The outcomes of some of the models, such as one in 8.7, demonstrated that there are instances where a model without demographic data is preferable, as all alternatives introduce more bias than the WOD model.

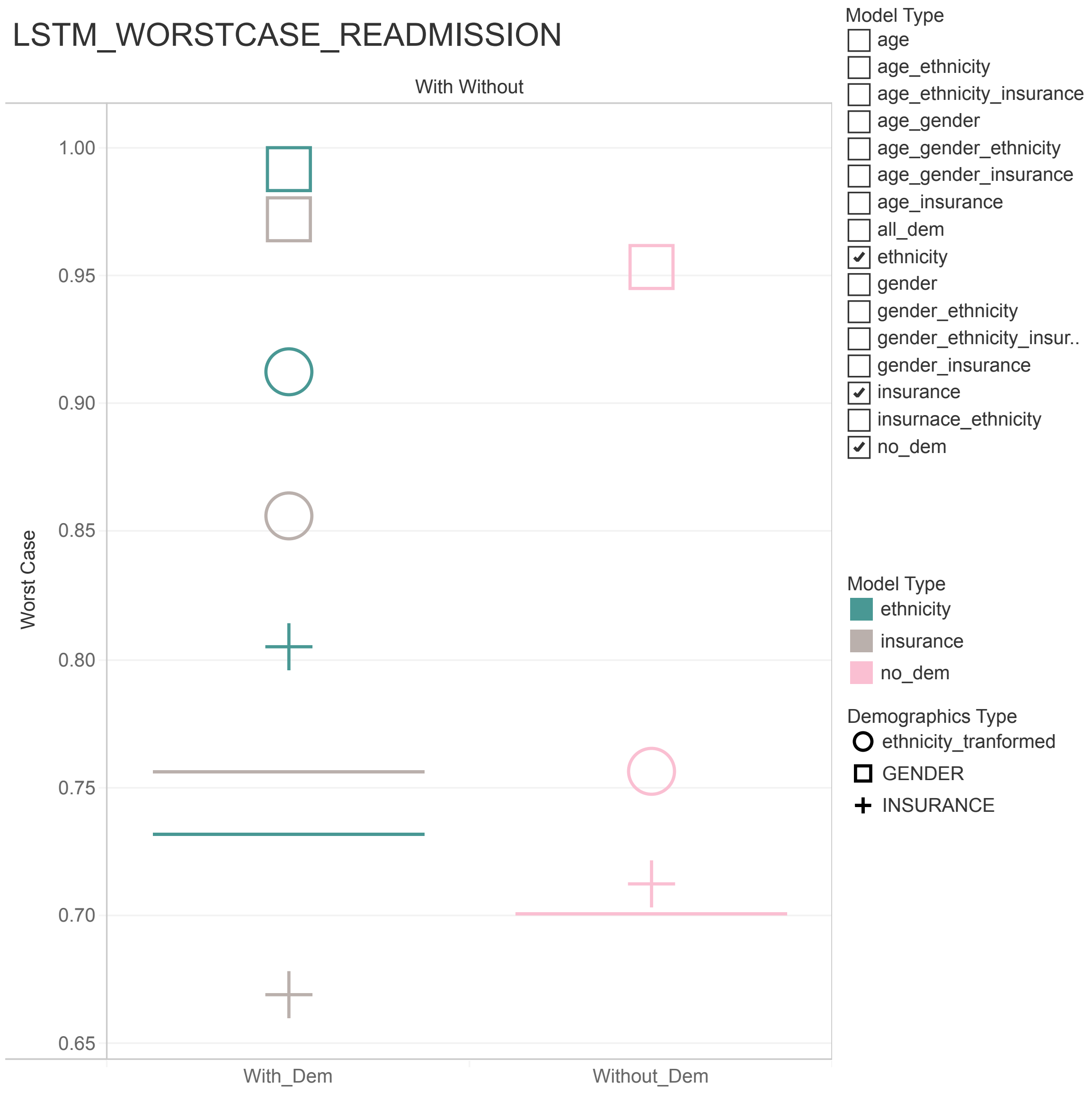
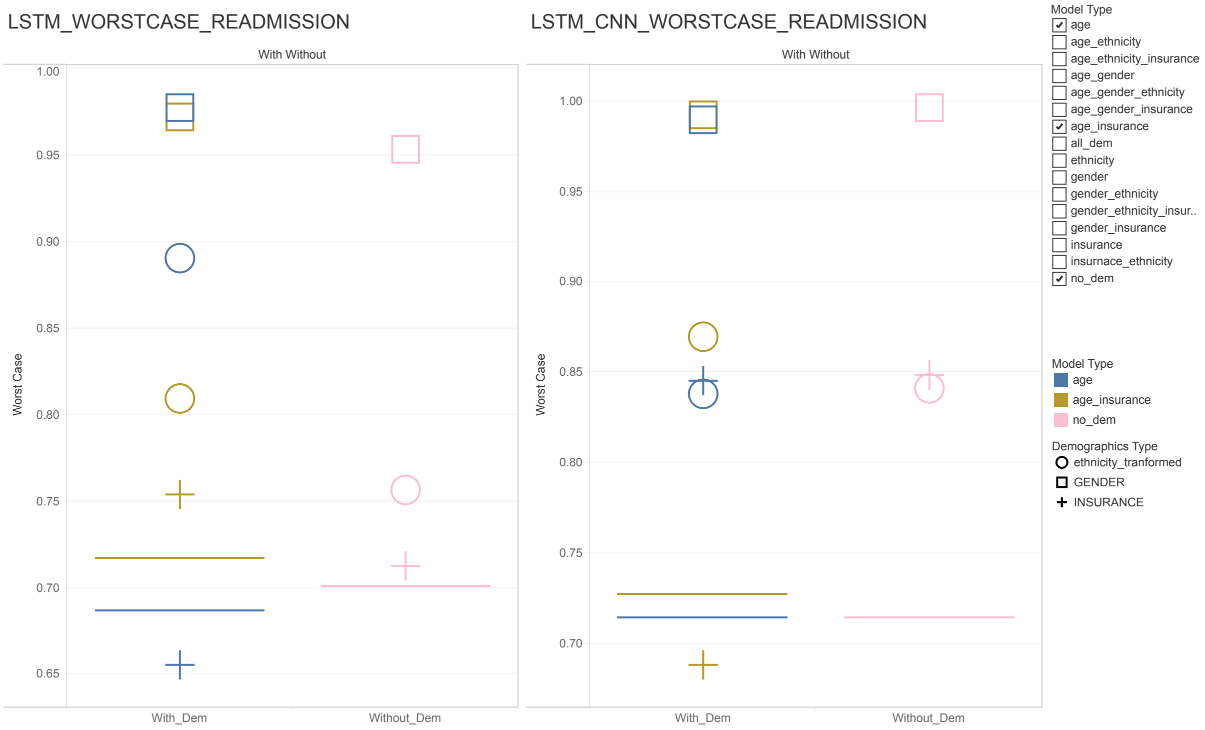


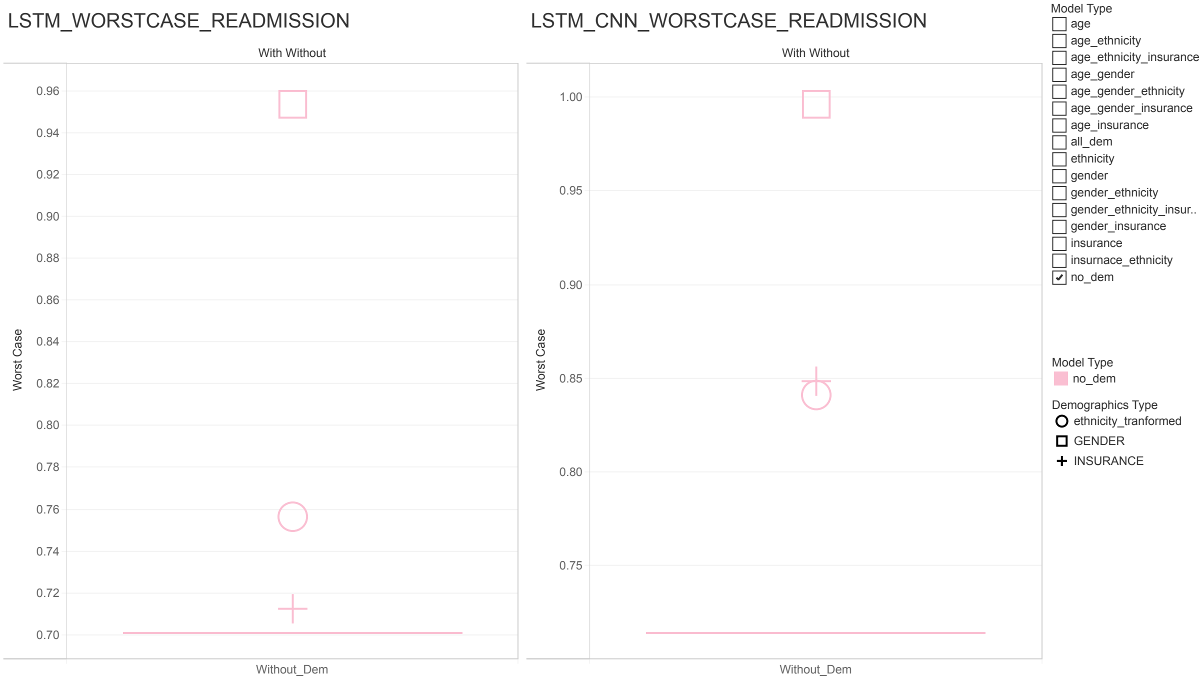
Figure 9.1: Main Takeaway - screenshot of readmission risk application: LSTM Architecture worst-case disparity analysis



Model Type

☑ age
☐ age_ethnicity
☐ age_ethnicity_insurance
☐ age_gender
☐ age_gender_ethnicity
☐ age_gender_insurance
☑ age_insurance
☐ all_dem
☐ ethnicity
☐ gender
☐ gender_ethnicity
☐ gender_ethnicity_insur..
☐ gender_insurance
☐ insurance
☐ insurnace_ethnicity
☑ no_dem

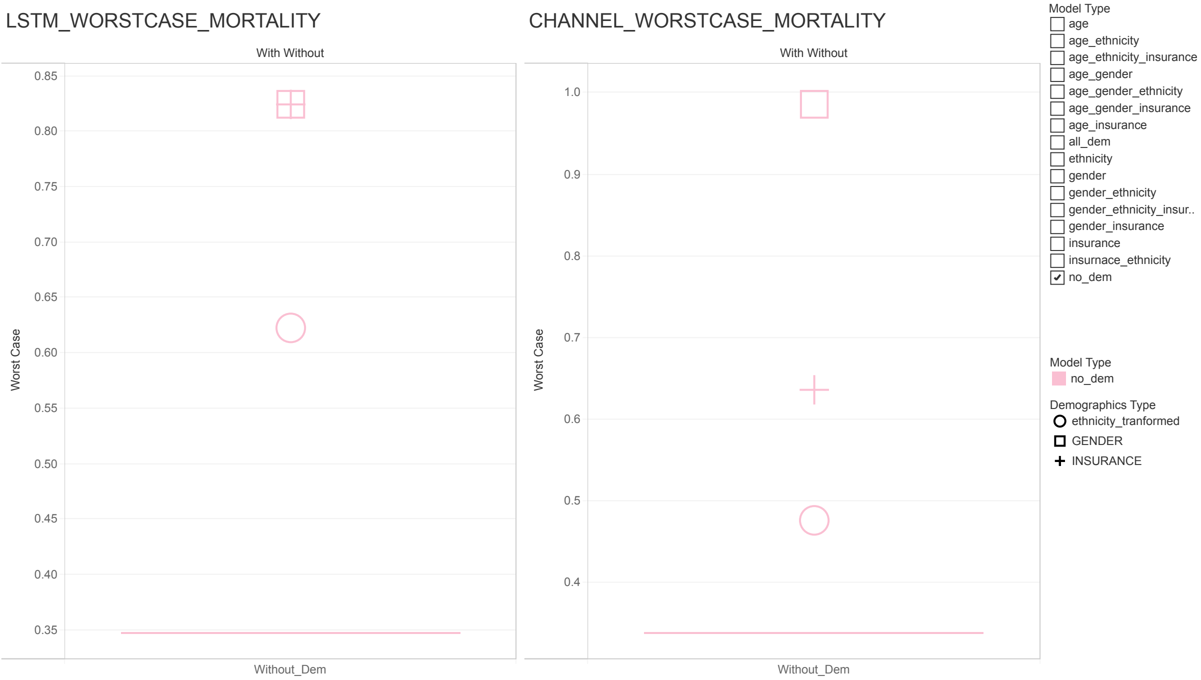
Figure 9.2: Main Takeaway - screenshot of readmission risk application: worst-case disparity analysis



Model Type

☐ age
☐ age_ethnicity
☐ age_ethnicity_insurance
☐ age_gender
☐ age_gender_ethnicity
☐ age_gender_insurance
☐ age_insurance
☐ all_dem
☐ ethnicity
☐ gender
☐ gender_ethnicity
☐ gender_ethnicity_insur..
☐ gender_insurance
☐ insurance
☐ insurnace_ethnicity
☑ no_dem

Figure 9.3: Main Takeaway - screenshot of readmission risk application WOD Models: worst-case disparity analysis



Model Type

- □ age
- □ age_ethnicity
- □ age_ethnicity_insurance
- □ age_gender
- □ age_gender_ethnicity
- □ age_gender_insurance
- □ age_insurance
- □ all_dem
- □ ethnicity
- □ gender
- □ gender_ethnicity
- □ gender_ethnicity_insur..
- □ gender_insurance
- □ insurance
- □ insurnace_ethnicity
- ☑ no_dem

Figure 9.4: Main Takeaway - screenshot of mortality risk application WOD Models: worst-case disparity analysis

Chapter 10: LIMITATIONS AND FUTURE WORK

10.1 Limitations

Time Limitation

Although this analysis is highly valuable for seeing the nuances of relationships between demographic information, different architectures, and applications, its implementation may require significant time and resources. Running multiple models simultaneously numerous times can be time-consuming, especially if the number of demographic features considered is high or if the model architecture is complex. In addition, a significant amount of time is also required to go through the results of each model and understand the results of each before making decisions. Nonetheless, we think that the time invested in this analysis is critical before deploying ML models for healthcare decisions and is comparable to other testing protocols before deploying more traditional healthcare solutions.

Small Groups Have Outsized Effect

There are two key challenges here. First, when examining subgroups and different intersectional subgroups, if the number of patients in a subgroup is small, the reported results may be disproportionately inflated or deflated compared to other groups. This can lead to misleading conclusions. Therefore, it is crucial to consider this factor carefully. Second, while implementing methods to address this issue, we must also ensure the protection of patients within these small groups. Historically, individuals in these marginalized intersectional groups have faced bias. Disregarding their data due to potential inflation undermines the purpose of these dashboards. Therefore, we need an approach that accounts for a smaller number while protecting patients in this category [18].

10.2 Future Work

In this work, we developed a framework to assess the trade-offs associated with using demographic information for training healthcare models. However, the use of demographic information remains contentious. Although our framework can be utilized for a transparent and intentional use of demographic information, it does not dive into why less biased demographic features yield better results. For example, a study has shown that patients' quality of care varies based on their insurance status, with privately insured patients having lower mortality risks[36]. Additionally, research has also revealed gender-based disparities in the quality of care. Female patients receive better treatment for certain conditions, while male patients experience better outcomes for others.[6]. So when these features reduce the bias in model training, it is crucial to understand why such information improves model performance and reduce bias. Further studies are necessary to separate intrinsic historical and systemic bias from that attributed to ML models and the use of demographic data.

A crucial next step for this work would involve incorporating a mechanism to explain how demographic information utilized by the models is used to generate the observed results. This step would not only assist developers in determining which demographic features to include in model training but also shed light on existing biases within the healthcare sector on a case-by-case basis. Furthermore, it would also reveal implicit or explicit correlations between these features and various healthcare applications, thereby opening opportunities to address existing issues in healthcare [24].

Furthermore, as highlighted in the above limitations, it is essential to develop a mechanism that protects subgroups with a small patient count while ensuring that the reported results are not unnecessarily inflated or deflated due to the small patient count in the subgroup under study. Only then will it be possible to protect all patients and ensure that they receive fair treatment through these models.

Lastly, at present, the results can be visually inspected and interpreted. While this is generally acceptable as the results are primarily based on demographic information that introduces obvious biases, in scenarios where the results of two or more models are a close call, a streamlined process for interpreting results should be established. This would ensure the repeatability of the work by different programmers and promote consistency in selecting demographic features.

Bibliography

[1] Selin Akgun and Christine Greenhow. Artificial intelligence in education: Addressing ethical challenges in k-12 settings. *AI and Ethics*, 2(3):431–440, 2022.

[2] Shahriar Akter, Yogesh K Dwivedi, Shahriar Sajib, Kumar Biswas, Ruwan J Bandara, and Katina Michael. Algorithmic bias in machine learning-based marketing models. *Journal of Business Research*, 144:201–216, 2022.

[3] Amine Amyar, Romain Modzelewski, Hua Li, and Su Ruan. Multi-task deep learning based ct imaging analysis for covid-19 pneumonia: Classification and segmentation. *Computers in biology and medicine*, 126:104037, 2020.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. propublica, 23 may, 2016.

[5] Machine Bias. There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica.- URL: https://www. propublica. org/article/machine-bias-risk-assessments-in-criminalsentencing (: 25.11. 2022)*, 2016.

[6] Chloe E Bird, Marc N Elliott, John L Adams, Eric C Schneider, David J Klein, Jacob W Dembosky, Sarah Gaillot, Allen M Fremont, and Amelia M Haviland. How do gender differences in quality of care vary across medicare advantage plans? *Journal of General Internal Medicine*, 33:1752–1759, 2018.

[7] Luisa N Borrell, Jennifer R Elhawary, Elena Fuentes-Afflick, Jonathan Witonsky, Nirav Bhakta, Alan HB Wu, Kirsten Bibbins-Domingo, José R Rodríguez-Santana, Michael A Lenoir, James R Gavin III, et al. Race and genetic ancestry in medicine—a time for reckoning with racism, 2021.

[8] Esteban González Burchard, Elad Ziv, Natasha Coyle, Scarlett Lin Gomez, Hua Tang, Andrew J Karter, Joanna L Mountain, Eliseo J Pérez-Stable, Dean Sheppard, and Neil Risch. The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine*, 348(12):1170–1175, 2003.

[9] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

[10] Zhisheng Chen. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications*, 10(1):1–12, 2023.

[11] Youngduck Choi, Chill Yi-I Chiu, and David Sontag. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41, 2016.

[12] Richard S Cooper. Race and genomics. *The New England journal of medicine*, 348(12):1166, 2003.

[13] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022.

[14] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

[15] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Proxy non-discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120*, 2017.

[16] Kadija Ferryman and Mikaela Pitcan. Fairness in precision medicine. 2018.

[17] Avijit Ghosh, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pages 22–34. PMLR, 2021.

[18] Usman Gohar and Lu Cheng. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. *arXiv preprint arXiv:2305.06969*, 2023.

[19] Richard Ha, Peter Chang, Jenika Karcich, Simukayi Mutasa, Eduardo Pascual Van Sant, Michael Z Liu, and Sachin Jambawalikar. Convolutional neural network based breast cancer risk stratification using a mammographic dataset. *Academic radiology*, 26(4):544–549, 2019.

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

[21] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

[22] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[23] Yu-Wei Lin, Yuqian Zhou, Faraz Faghri, Michael J Shaw, and Roy H Campbell. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PloS one*, 14(7):e0218942, 2019.

[24] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[25] Colleen K McIlvennan, Zubin J Eapen, and Larry A Allen. Hospital readmissions reduction program. *Circulation*, 131(20):1796–1803, 2015.

[26] Xueyan Mei, Hao-Chih Lee, Kai-yue Diao, Mingqian Huang, Bin Lin, Chenyu Liu, Zongyu Xie, Yixuan Ma, Philip M Robson, Michael Chung, et al. Artificial intelligence–enabled rapid diagnosis of patients with covid-19. *Nature medicine*, 26(8):1224–1228, 2020.

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

[28] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

[29] E Röösli, S Bozkurt, and T Hernandez-Boussard. Peeking into a black box, the fairness and generalizability of a mimic-iii benchmarking model. sci data. 2022; 9 (1): 24. *An evaluation finding various forms of unfair performance in an open access benchmarking model, and calling for greater thoroughness and transparency in reporting of artificial intelligence-based tools. Article*, 2022.

[30] Alexander Rusanov, Nicole G Weiskopf, Shuang Wang, and Chunhua Weng. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC medical informatics and decision making*, 14(1):1–9, 2014.

[31] Barret Rush, Leo Anthony Celi, and David J Stone. Applying machine learning to continuously monitored physiological data. *Journal of clinical monitoring and computing*, 33(5):887–893, 2019.

[32] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In

BIOCOMPUTING 2021: proceedings of the Pacific symposium, pages 232–243. World Scientific, 2020.

[33] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

[34] Thanveer Shaik, Xiaohui Tao, Niall Higgins, Lin Li, Raj Gururajan, Xujuan Zhou, and U Rajendra Acharya. Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2):e1485, 2023.

[35] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):12495, 2019.

[36] Christine S Spencer, Darrell J Gaskin, and Eric T Roberts. The quality of care delivered to patients within the same hospital varies by insurance type. *Health Affairs*, 32(10):1731–1739, 2013.

[37] Dilber Uzun Ozsahin, Cemre Ozgocmen, Ozlem Balcioglu, Ilker Ozsahin, and Berna Uzun. Diagnostic ai and cardiac diseases. *Diagnostics*, 12(12):2901, 2022.

[38] Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.