Exploring Text Specific and Blackbox Fairness Algorithms in Multimodal Clinical NLP

John Chen[1,2], Ian Berlot-Attwell [1,2], Safwan Hossain [1,2], Xindi Wang [2,3], Frank Rudzicz [1,2,4]
[1] University of Toronto, [2] Vector Institute, [3] University of Western Ontario, [4] St. Michael's Hospital
{johnc, ianberlot, hossa120, frank }@cs.toronto.edu
xwang842@uwo.ca

Abstract

Clinical machine learning is increasingly multimodal, collected in both structured tabular formats and unstructured forms such as free text. We propose a novel task of exploring *fairness* on a multimodal clinical dataset, adopting *equalized odds* for the downstream medical prediction tasks. To this end, we investigate a modality-agnostic fairness algorithm - equalized odds post processing - and compare it to a text-specific fairness algorithm: debiased clinical word embeddings. Despite the fact that debiased word embeddings do not explicitly address equalized odds of protected groups, we show that a text-specific approach to fairness may simultaneously achieve a good balance of performance *and* classical notions of fairness. We hope that our paper inspires future contributions at the critical intersection of clinical NLP and fairness. The full source code is available here: https://github.com/johntiger1/multimodal_fairness

1 Introduction

Natural language processing is increasingly leveraged in sensitive domains like healthcare. For such critical tasks, the need to prevent discrimination and bias is imperative. Indeed, ensuring equality of health outcomes across different groups has long been a guiding principle of modern health care systems (**?**). Moreover, medical data presents a unique opportunity to work with different *modalities*, specifically *text* (e.g., patient narratives, admission notes, and discharge summaries) and numerical or categorical data (often denoted *tabular* data, e.g., clinical measurements such as blood pressure, weight, or demographic information like ethnicity). Multi-modal data is not only reflective of many real-world settings, but machine learning models which leverage both structured and unstructured data often achieve greater performance than their individual constituents (**?**). While prior work studied fairness in the text and tabular modalities in isolation, there is little work on applying notions of algorithmic fairness in the broader multimodal setting (**??**).

Our work brings a novel perspective towards studying fairness algorithms for models which operate on *both* text and tabular data, in this case applied to the MIMIC-III clinical dataset (MIMIC-III) (**?**). We evaluate two fairness algorithms: equalized-odds through post-processing, which is agnostic to the underlying classifier, and word embedding debiasing which is a text-specific technique. We show that ensembling classifiers trained on structured and unstructured data, along with the aforementioned fairness algorithms, can both improve performance and mitigate unfairness relative to their constituent components. We also achieve strong results on several MIMIC-III clinical benchmark prediction tasks using a dual modality ensemble; these results may be of broader interest in clinical machine learning (**??**).

2 Background

2.1 Combining Text and Tabular Data in Clinical Machine Learning

Prior work has shown that combining unstructured text with vital sign time series data improves performance on clinical prediction tasks. **?** showed that augmenting an SVM with text information in addition to vital signs data improved retrospective sepsis detection. **?** showed that using a text-based risk score improves performance on prediction of death after surgery for a pediatric dataset. Closest to our work, **?** introduced a joint-modality neural network which outperforms single-modality neural networks on several benchmark prediction tasks for MIMIC-III.

2.2 Classical fairness metrics

Many algorithmic fairness notions fall into one of two broad categories: individual fairness enforcing fairness across individual samples, and group fairness seeking fairness across protected groups (e.g. race or gender). We focus on a popular group-level fairness metric: *Equalized Odds* (EO) (**?**). Instead of arguing that average classification probability should be equal across all groups (also known as *Demographic Parity*) – which may be unfair if the underlying group-specific base rates are unequal – EO allows for classification probabilities to differ across groups only through the underlying ground truth. Formally, a binary classifier $\widehat{Y}$ satisfies EO for a set of groups $\mathcal{S}$ if, for ground truth Y and group membership A:

$$\Pr(\hat{Y} = 1 \,|\, Y = y, A = a) = \Pr(\hat{Y} = 1 \,|\, Y = y, A = a') \\ \forall y \in \{0, 1\}, \forall a, a' \in \mathcal{S}$$

In short, the true positive (TP) and true negative (TN) rates should be equal across groups.

2.3 Equalized Odds Post Processing

? proposed a model-agnostic post-processing algorithm that minimizes this group specific error discrepancy while considering performance. Briefly, the post-processing algorithm determines group-specific random thresholds based on the intersection of group-specific ROC curves. The multimodality of our underlying data and the importance of privacy concerns in the clinical setting make post-processing especially attractive as it allows fairness to be achieved agnostic to the inner workings of the base classifier.

2.4 Debiasing word embeddings

Pretrained word embeddings encode the societal biases of the underlying text on which they are trained, including gender roles and racial stereotypes (**???**). Recent work has attempted to mitigate this bias in context-free embeddings while preserving the utility of the embeddings. **?** analyzed gender subspaces by comparing distances between word vectors with pairs of gender-specific words to remove bias from gender-neutral words. **?** extended this work to the multi-class setting, enabling debiasing in race and religion. Concurrent to their work, (**?**) propose iterative null space projection as a technique to hide information about protected attributes by casting it into the null space of the classifier. Following the recent popularity of BERT and ELMo, **?** consider extending debiasing to sentence-level, contextualized representations.

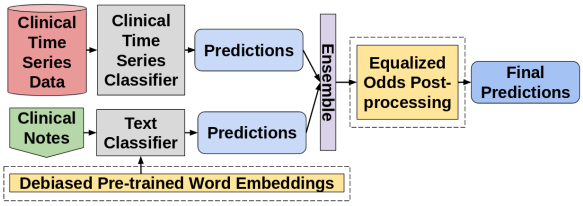

Figure 1: Experimental setup and ensemble architecture. Fairness approaches are indicated in dotted boxes.

3 Experimental Setup

3.1 Clinical Prediction Tasks

MIMIC-III contains deidentified health data associated with 60,000 intensive care unit (ICU) admissions (**?**). It contains both unstructured textual data (in the form of clinical notes) and structured data (in the form of clinical time series data and demographic, insurance, and other related meta-data). We focus on two benchmark binary prediction tasks for ICU stays previously proposed by **?**: in-hospital mortality prediction (IHM), which aims to predict mortality based on the first 48 hours of a patient's ICU stay, and phenotyping, which aims to retrospectively predict the acute-care conditions that impacted the patient. Following **?** we extend the prediction tasks to leverage clinical text linked to their ICU stay. For both tasks the classes are higly imbalanced: in the IHM task only 13.1% of training examples are positive, and the relative imbalance of the labels in the phenotyping class can be seen in Figure 2. To account for the label imbalance we evaluate performance using AUC ROC and AUC PRC. More details can be found in Appendix A.

3.2 Fairness Definition

Next, we consider how we can extend a definition of fairness to this multimodal task. Following work by **?** in the single-modality setting, we examine True Positive and True Negative rates on our clinical prediction task between different protected groups. Attempting to equalize these rates corresponds to satisfying *Equalized Odds*. EO satisfies many desiderata within clinical settings, and has been used in previous clinical fairness work (**???**). While EO does not explicitly incorporate the *multimodality* of our data, it accurately emphasizes the importance of the *downstream* clinical prediction task on the protected groups. Nonetheless, we

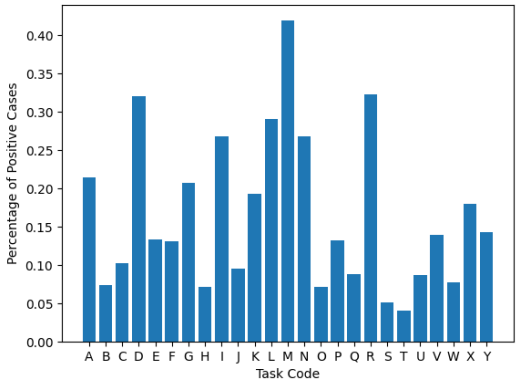

Figure 2: Percentage of positive train cases for each of the 25 phenotyping tasks. The critical care conditions corresponding to the task codes can be found in Table 3 of the Appendix

Sensitive Group	Train Count	Test Count	% of Test
F	7940	1415	44.0 %
M	9708	1778	56.0 %
ASIAN	408	60	1.9 %
BLACK	1658	285	8.9 %
HISPANIC	521	107	3.3 %
OTHER	2655	459	14.4 %
WHITE	12406	2282	71.5 %
Government	356	74	2.3 %
Medicaid	1362	205	6.4 %
Medicare	9857	1757	55.0 %
Private	4946	932	29.2 %
Self Pay	133	33	1.0 %
UNKNOWN	994	192	6.1 %

Table 1: Distribution of sensitive-attributes over train and test data for the In-Hospital Mortality task

acknowledge that EO alone is insufficient for practical deployment; naïve application can result in unacceptable performance losses and thus consultations with physicians and stakeholders must be held (**?**).

3.3 Classification Models

We provide brief descriptions below with details available in Appendix B.

- **Structured Data Model:** Following **?**, we use a channel-wise bidirectional Long Short Term Memory network (bi-LSTM).
- **Unstructured Textual Data:** We use a CNN encoder to extract the semantic features from clinical notes. Importantly, we experiment with training word embeddings from scratch and utilizing pre-trained BioWordVec embeddings (**?**).
- **Ensemble:** We perform logistic regression on the output binary classification probabilities from the previous models.

4 Fairness Setup

4.1 Sensitive groups

Recall that EO explicitly ensures fairness with respect to sensitive groups while debiasing implicitly depends upon it. Leveraging the demographic data in MIMIC-III, we consider ethnicity (divided into Asian, Black, Hispanic, White and other), biological sex (divided into male and female), and insurance type (divided into government, medicare, medicaid, self-pay, private, and unknown). With the exception of biological sex, the sensitive groups are highly imbalanced (see Table 1). Note that insurance-type has been shown to be a proxy for socioeconomic status (SES) (**?**).

4.2 Equalized Odds Post-Processing

We apply our equalized-odds post processing algorithm on the predictions of the trained single-modality classifiers (physiological signal LSTM model as well as text-only CNN model) as well as the trained ensemble classifier. Note that we apply EO postprocessing only once for each experiment: either on the outputs of the single-modality model, or on the ensemble predictions. The fairness approaches are mutually exclusive: we do not consider applying EO postprocessing together with debiased word embeddings. We consider using both soft prediction scores (interpretable as probabilities) as well as thresholded hard predictions as input to the post-processing algorithm. These choices impact the fairness performance trade-off as discussed further in Section 5.

4.3 Socially Debiased Clinical Word Embeddings

While clinically pre-trained word embeddings may improve downstream task performance, they are not immune from societal bias (**?**). We socially debias these clinical word embeddings following **?**. We manually select sets of social-specific words (see Appendix C) to identify the fairness-relevant social bias subspace. Formally, having identified the basis vectors $\{b_1, b_2, ..., b_n\}$ of the social bias

subspace $\mathcal{B}$, we can find the projection w_B of a word embedding w:

$$w_{\mathcal{B}} = \sum_{i=1}^{n} \langle w, b_i \rangle b_i$$

Next we apply hard debiasing, which will remove bias from existing word embeddings by subtracting w_B, their component in this fairness subspace. This yields w', our socially debiased word embedding:

$$w' = \frac{w - w_B}{\|w - w_B\|}$$

We consider debiasing with respect to race and gender. The race debiased embeddings are re-used for insurance tasks as empiric research has indicated that the use of proxy groups in fairness can be effective (**?**) and SES is strongly related to race (**?**).

5 Results and Analysis

	IHM		Phenotyping	
	AUC PRC	AUC ROC	Macro AUCROC	Overall AUCROC
Harutyunyan et. al (2019) – No Text	0.515	0.862	0.776	0.825
Khadanga et. al (2019) – Ensemble	0.525	0.865	–	–
Ours – Text Only	0.472	0.815	0.766	0.829
Ours – Text Only + BioWordVec	0.489	0.841	0.771	0.837
Ours – Ensemble	**0.582**	0.880	0.822	0.861
Ours – Ensemble + BioWordVec	**0.582**	**0.886**	**0.829**	**0.870**

Table 2: Leveraging clinical pretrained word embeddings improves performance compared to training word embeddings from scratch in the text-only model. Ensembling the text-only model with the clinical time series classifier improves performance further.

5.1 Ensembling clinical word embeddings with structured data improves performance

Empirically, we observe superior performance to prior literature on a suite of clinical prediction tasks in Table 2; more tasks are evaluated in Appendix Table A. Full hyperparameter settings and code for reproducibility can be found here [1]. The ensemble model outperforms both constituent classifiers (AUC plot on Figure 3). This holds even when fairness/debiasing techniques are applied, emphasizing the overall effectiveness of leveraging multi-modal data. However, the ensemble's improvements in performance do not directly translate to improvements in fairness; see the True Positive (TP) graph in Figure 3, where the maximum TP gap remains consistent under the ensemble.

5.2 Debiased word embeddings and the fairness performance trade-off

Improving fairness usually comes at the cost of reduced performance (**?**). Indeed, across all tasks, fairness groups and classifiers, we observe the group-specific disparities of TP and TN rates generally diminish when equalized odds post-processing is used (see Appendix F for additional results). However, this post-processing also leads to a degradation in the AUC. Note that we apply EO-post processing on hard (thresholded) predictions of the classifiers. If instead *soft* prediction scores are used as inputs to the post-processing step, both the performance degradation and the fairness improvement are softened (**?**).

Generally, word embedding debiasing (WED) also helps reduce TP/TN discrepancies, although not to the same extent as EO postprocessing. Remarkably, in certain tasks, WED also yields a performance improvement, even compared to the fairness-free, unconstrained ensemble classifier. In particular, for the AUC graph in Figure 3, leveraging debiased word embeddings improves the performance of the ensemble; at the same time, the TP and TN group discrepancy ranges are improved. However, we stress that this outcome was not consistently observed and further investigation is warranted.

We emphasize that EO and WED serve different purposes with different motivations. While EO explicitly seeks to minimize the TP/TN range between sensitive groups (reflected in its performance on the first two plots in Figure 3), WED seeks to neutralize text-specific bias in the word-embeddings. Despite the difference in goals, and despite operating only on the text-modality of the dataset, WED is still able to reduce the group-specific TP/TN range; recent work on *proxy fairness* in text has shown that indirect correlation between bias in text and protected attributes may be useful in achieving parity (**?**).

Although WED demonstrate some good proper-

[1] https://github.com/johntiger1/multimodal_fairness/clinicalnlp

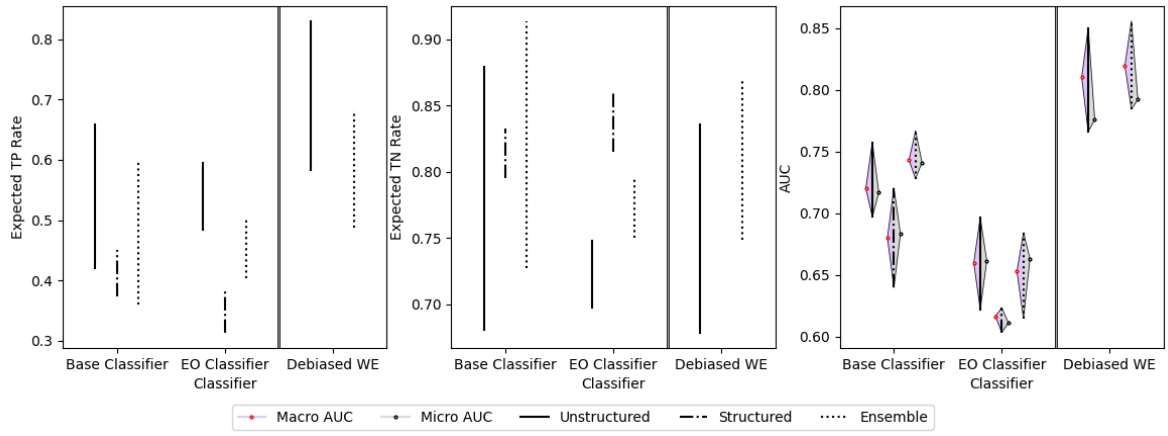

Figure 3: Plots of TP Rate, TN Rate, and AUC on phenotyping task M for groups defined by sensitive attribute of race. Each vertical black line represents a classifier (line style indicating modality); the length of the line represents the range of scores over fairness groups. In the TP/TN graphs, a shorter line represents better fairness; there is less discrepancy between the maximum and minimum group-specific TP/TN rates. In the AUC graph (far right), the higher the vertical position of the line, the better the performance. EO is effective at reducing the spread in TP/TN rates for the ensemble classifier (first two graphs) at the cost of performance (far right) graph. Meanwhile, debiased word embeddings both improves fairness, reducing the length of the line in the first two graphs, while achieving superior performance in AUC graph

ties with respect to both fairness and performance for our specific dataset and task, we caution that they represent only one approach to fairness in NLP (**?**). Indeed, WED suffers from shortcomings related to intersectional fairness (**?**), and we encourage further discussion into concretely defining fair, real-world NLP tasks and developing novel algorithms.

Our results highlight the important role practitioners and stakeholders play in algorithmic fairness on clinical applications. The trade-off between performance and fairness, whether between the soft and hard labels used for EO, or between EO and debiased word embeddings, must be balanced based on numerous real world factors.

6 Discussion

In this paper, we propose a novel multimodal fairness task for the MIMIC-III dataset, based on equalized odds. We provide two baselines: a classifier-agnostic fairness algorithm (equalized odds post-processing) and a text-specific fairness algorithm (debiased word embeddings). We observe that both methods generally follow the fairness performance tradeoff seen in single-modality tasks. EO is more effective at reducing the disparities in group-specific error rates while word-embedding debiasing has better performance. Future work can consider more generalized notions of fairness such as preferences-based frameworks, or extend text-specific fairness to contextualized word embeddings (**??**). Further analysis of the fairness performance tradeoff, especially in multimodal settings, will facilitate equitable decision making in the clinical domain.

7 Acknowledgements

We would like to acknowledge Vector Institute for office and compute resources. We would also like to thank Matt Gardner for his help with answering questions when using AllenNLP (**?**). John Chen and Safwan Hossain are funded by an Ontario Graduate Scholarship and a Vector Institute Research Grant. Ian Berlot-Attwell is funded by a Canada Graduate Scholarships-Master's, and a Vector Institute Research Grant. Frank Rudzicz is supported by a CIFAR Chair in AI.

A Details on Clinical Prediction Tasks

A.1 Defining the Multimodal MIMIC-III Benchmark Prediction Tasks

Existing work by (**?**) previously defined four benchmark clinical prediction tasks on ICU stays information from the large MIMIC-III database. They produce a derived dataset, focusing on 17 time-series clinical features, without text. The goal is predict the task specific outcome (mortality, phenotyping, decompensation, length-of-stay) for the given ICU stay. We utilize their derived dataset directly, which provides training and test examples for all four tasks, but join the derived dataset back with the original to obtain linked clinical text. We make the key choice that we drop examples without *relevant* (i.e. no causal leakage) extracted clinical notes, as in (**?**). Thus, we concretely define the Combined Modality MIMIC-III Benchmark Prediction Task as extending the benchmark clinical prediction task by (**?**) to include linked clinical text. If there are no notes associated with an example, then we remove this instance from the task. Note that we also drop ICU stays which only have unusable notes due to causal leakage; for instance death reports for mortality prediction.

A.2 Note extraction

To extract relevant notes, we build a mapping from the derived dataset provided by (**?**) and the MIMIC-III database. For each training and test instance in each task, we find the clinical notes in the MIMIC-III database. For the IHM task, if we do not find any notes within the first 48 hours of their stay, we drop the patient, since there is no *relevant* textual information. Note that this is consistent with the original task formulation by (**?**) of in-hospital mortality prediction using at most the first 48 hours of clinical data. Furthermore, this follows (**?**).

For the phenotyping task, which is not covered by (**?**), we relax this time condition. In the original formulation of the task, phenotyping is a *retrospective* multilabel multiclass classification task, meaning that all vital signs data associated with the ICU stay is provided and can be used by the model. Therefore, we only drop the patient if there are no notes for the entire ICU stay.

A.3 Preprocessing

We use the same preprocessing as in (**?**), finding it to be mildly beneficial for performance.

A.4 Cohort statistics

In the medical literature, *cohort selection* is the process of selecting the population of patients for inclusion in a study. These patients will then provide the training instances for the clinical prediction task. We report the cohort statistics for our binary clinical prediction multimodal tasks.

A.4.1 In-Hospital Mortality

Sensitive Group	Train Count	Test Count	% of Test
F	7940	1415	44.0 %
M	9708	1778	56.0 %
ASIAN	408	60	1.9 %
BLACK	1658	285	8.9 %
HISPANIC	521	107	3.3 %
OTHER	2655	459	14.4 %
WHITE	12406	2282	71.5 %
Government	356	74	2.3 %
Medicaid	1362	205	6.4 %
Medicare	9857	1757	55.0 %
Private	4946	932	29.2 %
Self Pay	133	33	1.0 %
UNKNOWN	994	192	6.1 %

A.4.2 Phenotyping

Sensitive Group	Train Count	Test Count	% of Test
F	15638	2750	44%
M	19803	3504	56%
ASIAN	826	133	2.1 %
BLACK	3378	575	9.1 %
HISPANIC	1158	206	3.3 %
OTHER	5004	854	13.7 %
WHITE	25075	4486	71.7 %
Government	845	150	2.4 %
Medicaid	2850	433	6.9 %
Medicare	18702	3298	52.7 %
Private	10784	1923	30.7 %
Self Pay	380	73	1.2 %
UNKNOWN	1880	377	6.0 %

A.5 Task Statistics

A.5.1 In-Hospital Mortality

Label	Train Set Count	Test Set Count
0	15337	2829
1	2311	364

A.5.2 Phenotyping

Plots of the prevalance of the 25 critical care conditions can be found in Figures 4 and 2 for the test and train sets respectively, a legend that doubles

as the full list of phenotyping tasks is available in Table 3.

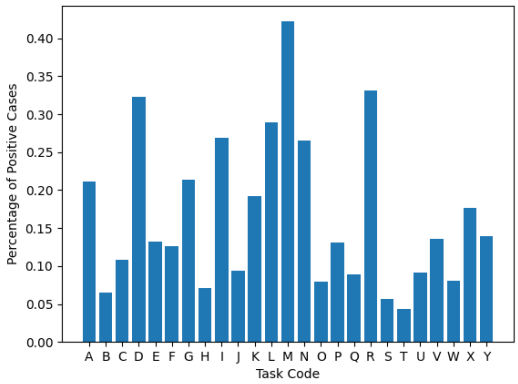

Figure 4: Percentage of positive test cases for each of the 25 phenotyping tasks

B Model Details

B.1 Structured Data Model

We use the baseline developed by **?**. The structured data model takes as input a time-series of 17 clinical variables, which are extracted features for the benchmark tasks introduced in the same paper. The model is a channel-wise LSTM where each clinical variable is transcoded by a bidirectional LSTM, concatenated with the other transcoded sequences and passed to a final LSTM for prediction.

B.2 Unstructured Data Model

We implement a simple CNN-based encoder (**??**) to process the clinical notes and produce a task-specific prediction. We experiment with various settings including model architecture, word embedding dimension, preprocessing, varying the maximum number of tokens, L2 regularization and batch size. Below, we report the final hyperparameters and settings used to generate all plots and reported throughout.

Our CNNEncoder is built using the AllenNLP framework (**?**). We use 1D kernel (n-gram) filter sizes of 2, 3 and 5, learning 5 filters for each filter size. Convolution is done on word embedding representations of the input, across n-gram windows of the sequence, and are pooled before being combined. The CNNEncoder produces a single fixed size vector, and we use a simple linear layer on top to perform the classification.

For all multimodal tasks, we limit the maximum number of tokens input to 1536, taking the most re-

Code	Task
A	Acute and unspecified renal failure
B	Acute cerebrovascular disease
C	Acute myocardial infarction
D	Cardiac dysrhythmias
E	Chronic kidney disease
F	Chronic obstructive pulmonary disease and bronchiectasis
G	Complications of surgical procedures or medical care
H	Conduction disorders
I	Congestive heart failure
J	Coronary atherosclerosis and other heart disease
K	Diabetes mellitus with complications
L	Diabetes mellitus without complication
M	Disorders of lipid metabolism
N	Essential hypertension
O	Fluid and electrolyte disorders
P	Gastrointestinal hemorrhage
Q	Hypertension with complications and secondary hypertension
R	nonhypertensive
S	Other liver diseases
T	Other lower respiratory disease
U	Other upper respiratory disease
V	Pleurisy
W	Pneumonia (except that caused by tuberculosis or sexually transmitted disease)
X	pneumothorax
Y	pulmonary collapse

Table 3: List of critical care conditions in the phenotyping task, and their corresponding alphabetic codes.

cent notes first (taking care to avoid causal leakage as described in 3.1), and apply preprocessing as in (**?**). For the decompensation task, we subsample the number of training instances due to engineering and efficiency reasons. From 2 million possible training instances, we sample 50 000 examples, with weighting to balance the number of positive and negatively training instances in a 50/50 split.

We train for up to 50 epochs, using Adam optimizer with learning rate set to 0.001. When we use pretrained word embeddings (either debiased or not), we do not finetune or update them. We do not use any L2 regularization or dropout, instead employing early stopping with patience of 5 epochs, using validation loss as the stopping criterion. We use batch size 256. Training is completed on 1 NVIDIA Titan Xp with 12 GB of memory.

B.3 Ensemble Model

We use scikit-learn (**?**) with the default setting of L2 regularization with $C = 1$

C Sets of social-specific Words

C.1 Sets of Gender-specific Words

- {"he", "she"}
- {"his", "hers"}
- {"son", "daughter"}
- {"father", "mother"}
- {"male", "female"}
- {"boy", "girl"}
- {"uncle", "aunt"}

C.2 Sets of Racial-specific Words

- {"black", "caucasian", "asian", "hispanics"}
- {"african", "caucasian", "asian", "hispanics"}
- {"black", "white", "asian", "hispanics"}
- {"africa", "america", "asia", "hispanics"}
- {"africa", "america", "china", "hispanics"}
- {"africa", "europe", "asia", "hispanics"}
- {"black", "caucasian", "asian", "latino"}
- {"african", "caucasian", "asian", "latino"}
- {"black", "white", "asian", "latino"}
- {"africa", "america", "asia", "latino"}
- {"africa", "america", "china", "latino"}
- {"africa", "europe", "asia", "latino"}
- {"black", "caucasian", "asian", "spanish"}
- {"african", "caucasian", "asian", "spanish"}
- {"black", "white", "asian", "spanish"}
- {"africa", "america", "asia", "spanish"}
- {"africa", "america", "china", "spanish"}
- {"africa", "europe", "asia", "spanish"}

D Hard Debiasing

Hard debiasing is a debiasing algorithm which involves two steps: neutralize and equalize. Neutralization ensures that all the social-neural words in the social subspace do not contain bias (e.g. doctors and nurses). Equalization forces that social-specific words are equidistant to all words in each equality set (e.g. the bias components in man and woman are in opposite directions but with same magnitude) (**??**). Following **?**, hard debiasing is formulated as follows: given a bias social subspace $\mathcal{B}$ spanned by the vectors $\{b_1, b_2, ..., b_n\}$, the embedding of a word in this subspace is:

$$w_{\mathcal{B}} = \sum_{i=1}^{n} \langle w, b_i \rangle b_i$$

To neutralize, each word $w \in N$, where N is the set of social-neural words, remove the bias components from the word and the re-embedded word $\overrightarrow{w}$ is obtained as:

$$\overrightarrow{w} = \frac{w - w_{\mathcal{B}}}{\parallel w - w_{\mathcal{B}} \parallel}$$

To equalize, for an equality set E, let μ be the mean embeddings of the equlity set E, which is defined as:

$$\mu = \frac{w}{E} \sum_{w \in E}$$

For each word $w \in E$, the equalization is defined as:

$$\hat{w} = (\mu - \mu_{\mathcal{B}}) + \sqrt{1 - \parallel \mu - \mu_{\mathcal{B}} \parallel^2} \frac{w - w_{\mathcal{B}}}{\parallel w - w_{\mathcal{B}} \parallel}$$

When doing racial debiasing, we divide ethnicity into groups: White, Black, Asian, and Hispanics. We do not contain the "other" group as it hard to define social-specific sets and analogies for "other".

E Phenotyping Task

In Figure 3 we plot performance and fairness for the phenotyping task, specifically the detection of

disorders of lipid metabolism. This task was selected as it is the phenotyping task with the most balanced labels with 16855 negative instances and 12239 positive instances in the training data. Thus, it should be more amenable to EO postprocessing. As expected we see that EO postprocessing succeeds in reducing the TP/TN ranges at the cost of AUC. We also again see that ensembling improves performance both before and after postprocessing. For this task specifically we observe that using debiased word embeddings improves AUC compared to the non-debiased word embeddings.

F Full Results

Our experiment universe consisted of the cross product between choice of protected attribute (gender, ethnicity, insurance status), task (phenotyping, in-hospital mortality prediction, decompensation), hard vs soft EO postprocessing and word embedding vs debiased word embedding.

F.1 Fairness/Performance on the In-Hospital Mortality Task

We provide a more detailed set of graphs for an in-hospital mortality prediction task, where we used hard EO postprocessing on protected groups defined by insurance status. We illustrate the TP/TN/AUC metrics for each protected group in Figure 5.

In this task configuration, as well as the task configuration in Figure 3 EO postprocessing is applied to hard classification of the three classifiers in the Base Classifier column, to produce the EO Classifier column. The Debiased Word Embedding (WE) column contains an unstructured classifier using word embeddings debiased for 4 ethnicities, and an ensemble created by merging the aforementioned classifier with the structured base classifier. We utilize debiasing on ethnicity type as a proxy for insurance status, as mentioned in the Discussion.

Note that EO post-processing sometimes *worsens* the TP/TN spread, as in the TP graph for the structured classifier. We therefore qualify our EO results by noting the limitations of our real-world dataset, which include significant group and label imbalance and non-binary group labels, all of which impact the results of EO post-processing (see Appendix A.4).

Finally, on this task configuration, we observe that debiased word embeddings are not a panacea. We note that WED has slightly worsened the TP gap, and does not offer a clear cut performance improvement as on the phenotyping task M. Therefore, further research is needed to explore when and why debiased word embeddings may simultaneously improve fairness and performance. Ultimately, domain expertise and focus on the downstream impact on the patient experience will be critical for leveraging any of these fair machine learning models in clinical applications.

F.2 Full table of results

The performance for all model and tasks tried can be found in Table 4. Note that debiased word embeddings can improve the performance (micro and macro AUC), even compared to an unconstrained classifier using clinically relevant BioWordVecembeddings.

	IHM		Phenotyping		Decompen.	
	AUC PRC	AUC ROC	Macro AUC ROC	Micro AUC ROC	AUC PRC	AUC ROC
Harutyunyan et. al (2019) – No Text	0.515	0.862	0.776	0.825	0.344	0.911
Khadanga et. al (2019) – Ensemble	0.525	0.865	–	–	0.345	0.907
Ours – Text Only	0.472	0.815	0.766	0.829	0.235	0.867
Ours – Text Only + BioWordVec	0.489	0.841	0.771	0.837	0.225	0.879
Ours – Text Only + BioWordVec + Debiasing	0.392	0.790	0.831	0.874	0.265	0.331
Ours – Ensemble	**0.582**	0.880	0.822	0.861	0.399	0.917
Ours – Ensemble + BioWordVec	**0.582**	**0.886**	0.829	0.870	**0.404**	**0.920**
Ours – Ensemble + BioWordVec + Debiasing	0.539	0.870	**0.854**	**0.888**	0.405	0.922

Table 4: Leveraging clinical pretrained word embeddings improves performance compared to training word embeddings from scratch in the text-only model. Ensembling the text-only model with the clinical time series classifier improves performance further. As with **?**, our results are not directly comparable with **?** since we ignore patients without any clinical notes.

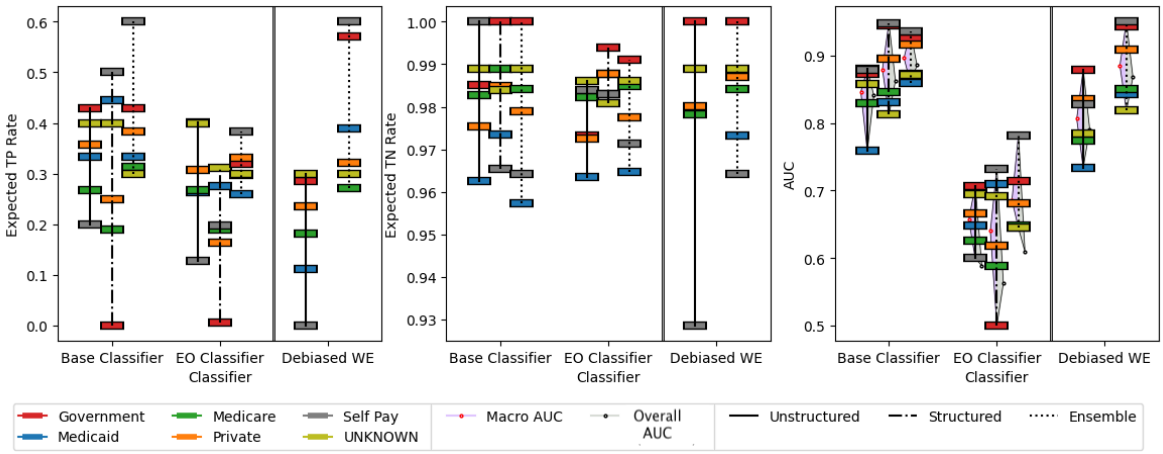

Figure 5: Plot of Fairness and Performance on the in-hospital mortality task. Note that debiased word embeddings slightly worsen the TP gap in this task (left most graph), while improving the TN gap (middle graph). EO reduces both gaps, at a major cost in performance (right most graph).