Bias Embedded in Care?:
Bias Identification in BERT-based Word Embeddings for Clinical Outcomes

Dominic Dillingham, Elaine Chang
University of California, Berkeley
{dom_dillingham,eechang}@berkeley.edu

Abstract

This paper builds upon and extends previous research of machine learning bias in the clinical care space. We will examine the prevalence of gender bias in two popular medical-focused pre-trained BERT models, BioBERT and PubMedBERT. Bias will be measured both through group fairness on classification as well as fill-in-the-blank tasks. Additionally, this paper will focus on the impact that fine tuning the BERT embedding may have on the measured bias from pre-trained representations.

1 Introduction

Machine learning algorithms are increasingly deployed in the clinical healthcare space to assist with patient care [4], from managing patient records to reducing operating costs and predicting illnesses before onset of symptoms. Such technology is reliant on both unstructured and structured data generated from the medical care staff such as clinical notes and metadata codes from the International Classification of Diseases (ICD). Clinical notes can provide the context of care from the provider while ICD codes are used for different operational processes such as medical diagnosis, insurance and billing. Recently, there has been findings that the combination of unstructured and structure data may provide better results than one or either as has been done historically. [10] Thus, as the scope of machine learning grows - both in terms of use in the clinical space as well as expansion into data types computed - it becomes increasingly critical to examine the fairness in which these algorithms guide providers to care for patients. As many practitioners tend to finetune in place of full pre-training, we focus our attention on widening the research in this space by comparing metrics from previous research to two pre-trained, medical-focused BERT models. Further, we examine the impact of finetuning a pre-trained embedding on the prevalence of bias compared to the original representation.

2 Background

Existing research has been completed to use transformers to support medical professionals in predicting diagnoses based on medical notes [2] as well as on bias in BERT-based architectures pre-trained on MIMIC-III[3]. Furthermore, attempts at intentional debiasing during pre-training on word embeddings was insufficient to overcome identified performance gaps[3] and in some instances has shown to come at the cost of reduced performance.[7] [1]

To that end, our research seeks to build on existing research in exploring how contextual transformers may carry implicit bias in their dataset and learned embeddings in order to understand its ultimate impact on prediction of medical procedures. We will follow a similar process outlined in previous research but, in place of pre-training the embeddings, results will be examined using pre-trained embeddings and embeddings finetuned on medical targets. We selected BioBERT and PubMedBERT out of the seven primary clinical NLP models currently available, as shown in Table 1, based on differentiation in vocabulary, pre-training and corpus scope.

	Vocabulary	Pretraining	Corpus	Text Size
BERT	Wiki + Books	-	Wiki + Books	3.3B words / 16GB
RoBERTa	Web crawl	-	Web crawl	160GB
BioBERT	Wiki + Books	continual pretraining	PubMed	4.5B words
SciBERT	PMC + CS	from scratch	PMC + CS	3.2B words
ClinicalBERT	Wiki + Books	continual pretraining	MIMIC	0.5B words / 3.7GB
BlueBERT	Wiki + Books	continual pretraining	PubMed + MIMIC	4.5B words
PubMedBERT	PubMed	from scratch	PubMed	3.1B words / 21GB

Table 1: Summary of pre-training data from BERT models based on their time of publication [5]

For BioBERT, we used BioBERT-Base v1.1 (+PubMed 1M) whereas PubMedBERT only has the base, uncased version.

Specifically, our research seeks to evaluate two hypotheses:

1. PubMedBERT will exhibit the least bias among the three language models (as compared to the baseline SciBERT from previous research [3] and BioBERT) due to its training on a larger portion of medical-specific tokens

2. Finetuning the final layers of PubMedBERT on task-specific measures will further decrease the bias by tuning to local language inplace of using biased non-scientific language heuristics to form predictions

3 Methods

3.1 Data

We make use of the classical clinical NLP research dataset, Multiparameter Intelligence Monitoring in Intensive Care (MIMIC-III). The MIMIC-III dataset covers the medical records (including clinical notes of their stay, ICD-9 medical diagnoses and procedures performed) of over forty thousand patients who stayed in critical care units at Beth Israel Deaconess Medical Center between 2001 and 2012. The dataset contains roughly 2 million clinical notes of varying categories as well as patient information which is used to detect bias. The data preparation for this paper follows directly from prior research [3] to ensure consistency in the produced results. This includes generating a set of 57 binary classification targets that examine in-hospital mortality, and phenotyping on all notes from a patient or just on those notes generated in the first 48 hours of the patient's stay. All models are based on the same holdout, where 20

3.2 Model Structure and Training

To test the fairness on medical outcomes, a fully-connected neural network is trained on the top of the embeddings where the number of layers, number of epochs, dropout, and decay rate are selected through grid search independently for each of the binary classification targets. The best model for each target is selected using AUPRC as the evaluation metric.

To evaluate the fairness of the embedding after finetuning, we train the PubMedBERT model while unfreezing the last 5 layers. 5 layers were judgmentally selected based on existing research indicating that large datasets receive the majority of performance gains by the 6th unfrozen layer.[6] The embedding is finetuned on each of the binary classification targets, updating the embedding with each iteration. Each target was trained up to 5 epochs with early stopping criteria and the best model was selected based on the validation set again using AUPRC as the evaluation criteria.

The project was run on two GCP instances. One large CPU-intensive instance was provisioned to run the data preparation pipeline. This instance used 16 vCPUs and 100GB of RAM. The second instance was used for model training with 8 vCPUs on 30 GB of RAM running on 1 T4 GPU.

3.3 Fairness Definitions

We evaluate fairness in two ways. First, we evaluate the fairness of the classifier using three commonly used fairness properties of 1) demographic parity, 2) equality of opportunity for the positive class and 3) equality of opportunity for the negative class. The complimentary metrics will be referred to as Parity Gap, Recall Gap and Specificity Gap, respectively. Recall Gap is the primary metric of interest due to the clinical motivation to minimize false negatives and structural limitations of the other two. The Parity Gap is limited in ensuring fairness of outcomes and may in fact promote bias unintentionally[9] while the Specificity Gap is limited by the data imbalance towards the negative class that may cloak biased classifiers.

The second method used to evaluate fairness of the language models is through fill-in-the-blank tasks. The models are given template sentences across 7 different condition categories - addiction, heart disease, diabetes, 'do not resuscitate' analgesics, HIV, hypertension and mental illness and will predict the gender pronoun in the sentence. For example, a sample addiction [ADD] template sentence reads "[GEND] has a history of [ADD] usage". Fairness is then measured using a log probability bias score that computes the log probabilities of male vs. female pronouns in the set of template sentences predicted from the respective BERT masked language prediction. This is then compared to the mean probabilities within categories between genders using the Wilcox signed-rank test to determine statistically significant differences.

4 Results & Discussion

In the first evaluation of fairness, we examine the performance gaps of the classification tasks.

For each of the fairness metrics - Recall Gap,

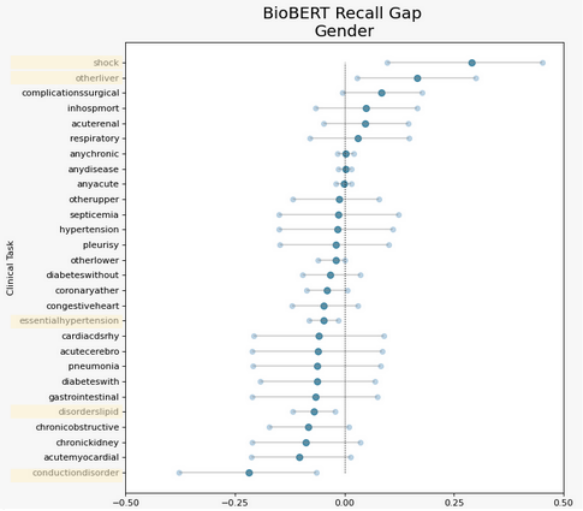

Figure 1: BioBERT classification by condition. Conditions highlighted in yellow show statistically significant performance differences between genders

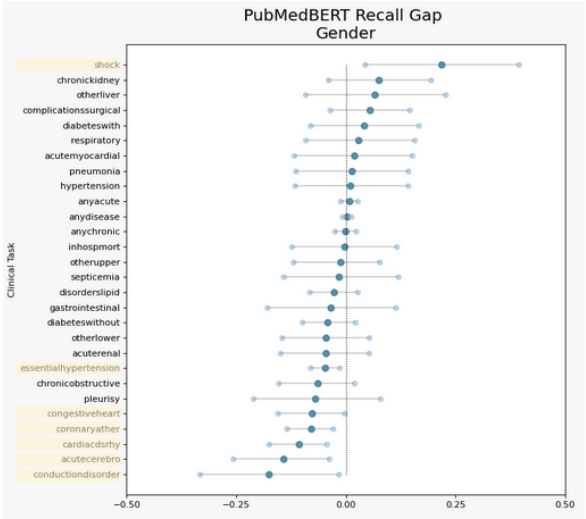

Figure 2: PubMedBERT classification by condition. Conditions highlighted in yellow show statistically significant performance differences between genders

Parity Gap and Specificity Gap - we calculate and visualize the gap metric itself, a confidence interval, as well as a flag of whether it favors the majority or non-majority group. In Figure 1, the visualization of BioBERT's recall gap for gender, tasks with recall gaps (blue dots) that favor men as opposed to women show a positive score.

BioBERT and PubMedBERT find show some similar bias when comparing the two groups as shown in Figure 2 (e.g., 'shock' towards male, 'conduction disorder' towards female). And while there are some differences in bias detected between the two language models (e.g., 'pneumonia' favors females in BioBERT while is more near to unbiased in PubMedBERT), there are no significant divergent results. PubMedBERT has a higher number of significant tasks favoring males.

Next, in Table 2, we tabulate the significant performance gaps of our classifier trained with BioBERT and PubMedBERT embeddings as well as the bias within those gaps towards majority vs. non-majority groups to compare the biases observed in the language models through this classification task.

We find that BioBERT and PubMedBERT had nearly all lower number of significant differences across all four sensitive groups of gender, language spoken, ethnicity and insurance coverage in its 57 downstream clinical tasks. For example, the gender Recall Gap for BioBERT shows 5 tasks with a significant difference between Males and Females and 7 tasks in PubMedBERT, as compared to the 13 tasks of the baseline SciBERT [3]. We see in this example we also see that of those tasks showing a significant performance difference, there was also a smaller percentage that favored males (the majority group) as compared to the baseline SciBERT but when performance gap is observed, it is more common to favor the majority group. Interestingly, unlike the baseline SciBERT, our results with BioBERT and PubMedBERT are both seen to show bias towards the non-majority (i.e., females) for gender misclassifications as opposed to the majority.

In our second assessment of fairness using a fill-in-the-blank task, we are able to compare the log probability bias scores of BioBERT, PubMedBERT as well as the finetuned BERT against the baseline SciBERT, shown in Table 3.

In these four predictive model comparisons, we see mostly different biases identified across conditions with the exception of mental illness where all four models reflected some statistically significant difference between males and females. The fine tuned results report a much smaller magnitude of log probability bias score as compared to the other three models shown in Table 4.

There are also more conditions observed to be statistically significant differences by gender in the fine tuned results - 4 conditions for fine tuned vs. 3 for each of the three pre-trained models. Based on the conditions flagged by the fine tuned results (i.e., addition, HIV, mental illness) as compared to the other domain-specific models (e.g., diabetes

	Significant Differences by Fairness Definition		
	Recall Gap	*Specificity Gap*	*Parity Gap*
Gender: Male vs. Female *(% of Tasks Favoring Male)*			
Baseline[4]	*13 (62%)*	*20 (80%)*	*25 (36%)*
BioBERT	5 (40%)	10 (80%)	10 (20%)
PubMedBERT	7 (14%)	13 (62%)	11 (36%)
Language: English vs. Other *(% of Tasks Favoring English)*			
Baseline[4]	*7 (29%)*	*9 (89%)*	*17 (12%)*
BioBERT	5 (0%)	7 (100%)	11 (0%)
PubMedBERT	4 (50%)	9 (100%)	12 (0%)
Ethnicity: White vs. Other *(% of Tasks Favoring White)*			
Baseline[4]	*4 (75%)*	*12 (17%)*	*22 (82%)*
BioBERT	1 (100%)	6 (0%)	8 (100%)
PubMedBERT	5 (100%)	10 (0%)	11 (100%)
Insurance: Medicare vs. Other *(% of Tasks Favoring Medicare)*			
Baseline[4]	*33 (85%)*	*48 (6%)*	*51 (92%)*
BioBERT	20 (100%)	26 (8%)	27 (93%)
PubMedBERT	17 (100%)	27 (4%)	27 (96%)

Table 2: Comparison of significant performance differences of baseline SciBERT results to BioBERT and PubMedBERT across four identified protected groups

	Baseline		BioBERT	
Category	*M*	*F*	M	F
Addiction	*0.202*	*0.313*	0.184	0.192
Heart Disease	*0.204**	*0.333**	**0.235***	**0.197***
Diabetes	*0.100*	*0.251*	0.309	0.257
"Do Not Resuscitate"	*0.070*	*0.032*	0.079	-0.015
Analgesics	*1.295*	*2.127*	**0.243***	**0.054***
HIV	*0.129*	*0.317*	0.093	0.077
Hypertension	*0.413*	*0.437*	0.201	0.142
Mental Illness	*-0.414**	*-0.164**	**0.432***	**0.352***

Table 3: Baseline and BioBERT log probability bias score gender comparison based on a set of template sentences related to 8 category of conditions. Statistical significant differences between males and females are denoted by an asterisk (*), reflecting p-value less than 0.01

and hypertension) and the relative size of scores, we speculate that by fine tuning targets to local language in lieu of using biased non-scientific language to form predictions, subtle social stigma was additionally introduced.

5 Limitations

There are limitations to our work that could guide and further the research of bias in the clinical context. First, we acknowledge the limitations of the fairness properties calculated. Fairness and equality of treatment in the real world are often multifactor with confounders. Next, we identify the limitations of the four protected groups (i.e., gender, ethnicity, language spoken and insurance provided)

	PubMedBERT		Fine Tuned	
Category	M	F	M	F
Addiction	-0.113	0.008	**1.073e-7***	**-1.193e-7***
Heart Disease	-0.143	-0.027	**3.223e-8***	**-4.677e-8***
Diabetes	**-0.251***	**0.011***	9.351e-8	6.005e-9
"Do Not Resuscitate"	-0.911	-0.742	-1.768e-7	-2.708e-7
Analgesics	1.372	1.711	1.020e-7*	1.277e-8*
HIV	0.160	0.071	**1.130e-7***	**-6.648e-8***
Hypertension	**-0.304***	**0.004***	1.327e-8	-2.00e-8
Mental Illness	**-0.190***	**0.140***	**1.038e-7***	**-1.515e-8***

Table 4: :
PubMedBERT and finetuned log probability bias score gender comparison based on a set of template sentences related to 8 category of conditions. Statistical significant differences between males and females are denoted by an asterisk (*).

and their associated attribute options (e.g., gender's binary construct). This limits the nuance of the bias that could be identified masks some of the known systemic bias such as the bias that transpeople often face in receiving clinical care.[8] Repeatedly, the analysis limits the exploration of sensitive subgroups and exploring intersectional bias that more closely reflects reality. This is evident in the classification results for insurance and Medicare where there is the largest number of tasks with significant fairness gaps with nearly all favoring the majority group. Furthermore, we acknowledge that the data itself may present bias and reflect the either existing systemic issues sensitive groups face while accessing care or clinical prevalence of conditions affecting a certain population.

6 Conclusion

Our research objective was to add to the growing work of bias quantification within BERT-based architectures. We sought to evaluate two hypotheses. First, we wanted to understand whether a pretrained, domain-specific language model, in particular PubMedBERT, would present less bias than its comparative counterpart BioBERT and baseline counterpart SciBERT. We found a mixed set of fairness outcomes when they were assessed through a set of binary classification tasks as well as a set of fill-in-the-blank predictive tasks. While indeed BioBERT and PubMedBERT presented less bias than SciBERT, PubMedBERT did not do markedly better than BioBERT as initially hypothesized. Additionally, both models showed statistically significant log probability bias scores when predicting gender pronouns in medical notes for a given set of conditions. The biased predicted conditions dif-

fered between BioBERT and PubMedBERT as well as against the baseline model, with the exception of mental illness. Finally, we saw that finetuning the final layers of the BERT embeddings did decrease the measured bias but saw an increase in the areas of bias detected. In other words, the value of the log probability bias scores were much smaller across the conditions predicted however more conditions were flagged to have a significant difference in gender predictions as compared to the pre-trained models. This may indicate that the fine tuned BERT relied more heavily on a heuristics approach, revealing bias in the alternative form of social stigma based on the conditions identified through the bias score.

Ultimately our findings highlight the importance of caution when deploying machine learning algorithms, especially in the clinical context. Bias in the real world can be subtle in expression, but interwoven in our manmade systems. Modeling to account for such a phenomenon is near impossible and attempts to do so, both what we have found through existing research and our own project is met with mixed results. We hope our findings continue to reinforce the need for thoughtful deployment of advanced machine learning algorithms with careful conditions of the human impact that these algorithms create conclusions upon.

References

[1] Robert C. Williamson Aditya Krishna Menon. The cost of fairness in classification. *arXiv:1705.09055*, 2017.

[2] Haifeng Lin Anish Philip Brent Biseda, Gaurav Desai. Prediction of icd codes with clinical bert embeddings and text augmentation with label balancing using mimic-iii. *arXiv:2003.11515*, 2020.

[3] Mohamed Abdalla Matthew McDermott Marzyeh Ghassemi Haoran Zhang, Amy X. Lu. Hurtful words: Quantifying biases in clinical contextual word embeddings. *Arxiv*, 2020.

[4] Khachatrian H. Kale D.C. et al. Harutyunyan, H. Multitask learning and benchmarking with clinical time series data. *Sci Data*, 6, 96 2019.

[5] Jianfeng Gao Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *arXiv:2007.15779*, 2020.

[6] Jimmy Lin Jaejun Lee, Raphael Tang. What would elsa do? freezing layers during transformer finetuning. *arXiv:1911.03090*, 2019.

[7] Safwan Hossain Xindi Wang Frank Rudzicz John Chen, Ian Berlot-Attwell. Exploring text specific and blackbox fairness algorithms in multimodal clinical nlp. *arXiv:2011.09625*, 2020.

[8] Jamie Feldman Robert Garofalo Wylie Hembree Asa Radix Jae Sevelius Joshua D. Safer, Eli Coleman. Barriers to health care for transgender individuals. *Curr Opin Endocrinol Diabetes Obes*, 2017.

[9] Nathan Srebro Moritz Hardt, Eric Price. Equality of opportunity in supervised learning. *arXiv:1610.02413*, 2016.

[10] Yin C. Zeng J. et al. Zhang, D. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med Inform Decis Mak*, 20, 280 2020.