What Do You See in this Patient? Behavioral Testing of Clinical NLP Models

Betty van Aken **Sebastian Herrmann** **Alexander Löser**
Berliner Hochschule für Technik (BHT)
{bvanaken, aloeser}@bht-berlin.de
sebastianhe93@gmail.com

Abstract

Decision support systems based on clinical notes have the potential to improve patient care by pointing doctors towards overseen risks. Predicting a patient's outcome is an essential part of such systems, for which the use of deep neural networks has shown promising results. However, the patterns learned by these networks are mostly opaque and previous work revealed both reproduction of systemic biases and unexpected behavior for out-of-distribution patients. For application in clinical practice it is crucial to be aware of such behavior. We thus introduce a testing framework that evaluates clinical models regarding certain changes in the input. The framework helps to understand learned patterns and their influence on model decisions. In this work, we apply it to analyse the change in behavior with regard to the patient characteristics *gender*, *age* and *ethnicity*. Our evaluation of three current clinical NLP models demonstrates the concrete effects of these characteristics on the models' decisions. They show that model behavior varies drastically even when fine-tuned on the same data with similar AUROC score. These results exemplify the need for a broader communication of model behavior in the clinical domain.

1 Introduction

Outcome prediction from clinical notes. The use of automatic systems in the medical domain is promising due to their potential exposure to large amounts of data from earlier patients. This data can include information that helps doctors make better decisions regarding diagnoses and treatments of a patient at hand. Outcome prediction models take patient information as input and then output probabilities for all considered outcomes (Choi et al., 2018; Khadanga et al., 2019). We focus this work on outcome models using natural language in the form of clinical notes as an input, since they are a common source of patient information and contain a multitude of possible variables.

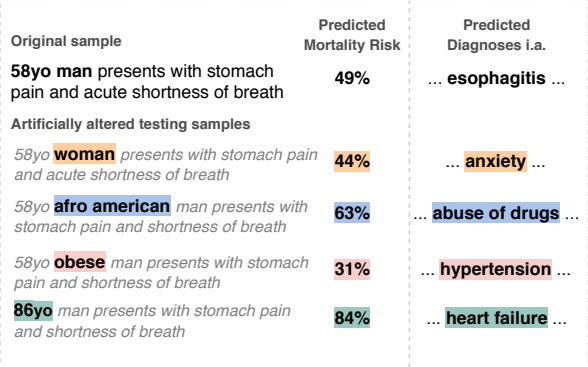

Figure 1: Minimal alterations to the patient description can have a large impact on outcome predictions of clinical NLP models. We introduce behavioral testing for the clinical domain to expose these impacts.

The problem of black box models for clinical predictions. Recent models show promising results on tasks such as mortality (Si and Roberts, 2019) and diagnosis prediction (Liu et al., 2018; Choi et al., 2018). However, since most of these models work as black boxes, it is unclear which features they consider important and how they interpret certain patient characteristics. From earlier work we know that highly parameterized models are prone to emphasize systemic biases in the data (Sun et al., 2019). Further, these models have high potential to disadvantage minority groups as their behavior towards out-of-distribution samples is often unpredictable. This behavior is especially dangerous in the clinical domain, since it can lead to underdiagnosis or inappropriate treatment (Straw, 2020). Thus, understanding models and allocative harms they might cause (Barocas et al., 2017) is an essential prerequisite for their application in clinical practice. We argue that more in-depth evaluations are needed to know whether models have learned medically meaningful patterns or not.

Behavioral testing for the clinical domain. As a step towards this goal, we introduce a novel test-

Proceedings of the 4th Clinical Natural Language Processing Workshop, pages 63 - 73

ing framework specifically for the clinical domain that enables us to examine the influence of certain patient characteristics on the model predictions. Our work is motivated by behavioral testing frameworks for general Natural Language Processing (NLP) tasks (Ribeiro et al., 2020) in which model behavior is observed under changing input data. Our framework incorporates a number of test cases and is further extendable to the needs of individual data sets and clinical tasks.

Influence of patient characteristics. As an initial case study we apply the framework to analyse the behavior of models trained on the widely used MIMIC-III database (Johnson et al., 2016). We analyse how sensitive these models are towards textual indicators of patient characteristics, such as *age*, *gender* and *ethnicity*, in English clinical notes. These characteristics are known to be affected by discrimination in health care (Stangl et al., 2019), on the other hand, they can represent important risk factors for certain diseases or conditions. That is why we consider it especially important to understand how these mentions affect model decisions.

Contributions. In summary, we present the following contributions in this work:
1) We introduce a behavioral testing framework specifically for clinical NLP models. We release the code for applying and extending the framework[1] to enable in-depth evaluations.
2) We present an analysis on the patient characteristics *gender*, *age* and *ethnicity* to understand the sensitivity of models towards textual cues regarding these groups and whether their predictions are medically plausible.
3) We show results of three state-of-the-art clinical NLP models and find that model behavior strongly varies depending on the applied pre-training. We further show that highly optimised models tend to overestimate the effect of certain patient characteristics leading to potentially harmful behavior.

2 Related Work

2.1 Clinical Outcome Prediction

Outcome prediction from clinical text has been studied regarding a variety of outcomes. The most prevalent being in-hospital mortality (Ghassemi et al., 2014; Jo et al., 2017; Suresh et al., 2018; Si and Roberts, 2019), diagnosis prediction (Tao et al., 2019; Liu et al., 2018, 2019a) and phenotyping (Liu et al., 2019b; Jain et al., 2019; Oleynik et al., 2019; Pfaff et al., 2020). In recent years, most approaches are based on deep neural networks due to their ability to outperform earlier methods in most settings. Most recently, Transformer-based models have been applied for prediction of patient outcomes with reported increases in performance (Huang et al., 2019; Zhang et al., 2020a; Tuzhilin, 2020; Zhao et al., 2021; van Aken et al., 2021; Rasmy et al., 2021). In this work we analyse three Transformer-based models due to their upcoming prevalence in the application of NLP in health care.

2.2 Behavioral Testing in NLP

Ribeiro et al. (2020) identify shortcomings of common model evaluation on held-out datasets, such as the occurrence of the same biases in both training and test set and the lack of broad testing scenarios in the held-out set. To mitigate these problems, they introduce CHECKLIST, a behavioral testing framework for general NLP abilities. In particular, they highlight that such frameworks evaluate input-output behavior without any knowledge of internal structures of a system (Beizer, 1995). Building upon CHECKLIST, Röttger et al. (2021) introduce a behavioral testing suite for the domain of hate speech detection to address the individual challenges of the task. Following their work, we create a behavioral testing framework for the domain of clinical outcome prediction, that comprise idiosyncratic data and respective challenges.

2.3 Analysing Clinical NLP Models

Zhang et al. (2020b) highlight the reproduction of systemic biases in clinical NLP models. They quantify such biases with the recall gap among patient groups and show that models trained on data from MIMIC-III inherit biases regarding gender, ethnicity, and insurance status–leading to higher recall values for majority groups. Log'e et al. (2021) further find disparities in pain treatment suggestions by language models for different races and genders. We take these findings as motivation to directly analyse the sensitivity of large pre-trained models with regard to patient characteristics. In contrast to earlier work and following Ribeiro et al. (2020), we want to eliminate the influence of existing data labels on our evaluation. Further, our approach simulates patient cases that are similar to real-life occurrences. It thus displays the actual impact of learned patterns on all analysed patient groups.

[1] URL: https://github.com/bvanaken/clinical-behavioral-testing

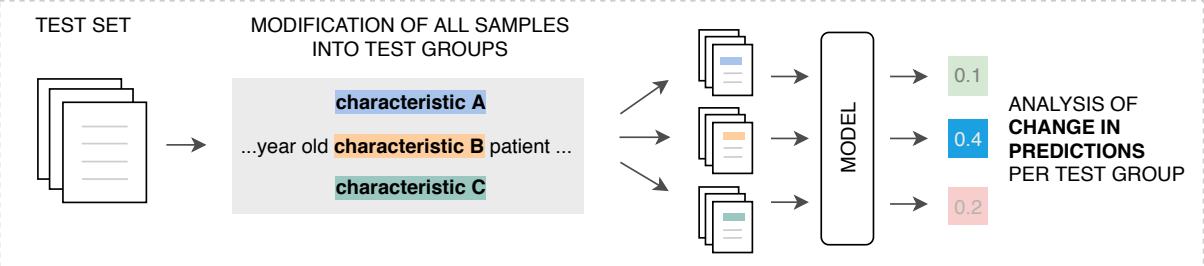

Figure 2: **Behavioral testing framework for the clinical domain**. Schematic overview of the introduced framework. From an existing test set we create test groups by altering specific tokens in the clinical note. We then analyse the change in predictions which reveals the impact of the mention on the clinical NLP model.

3 Behavioral Testing of Clinical NLP Models

Sample alterations. Our goal is to examine how clinical NLP models react to mentions of certain patient characteristics in text. Comparable to earlier approaches to behavioral testing we use sample alterations to artificially create different test groups. In our case, a test group is defined by one manifestation of a patient characteristic, such as *female* as the patient's gender. To ensure that we only measure the influence of this certain characteristic, we keep the rest of the patient case unchanged and apply the alterations to all samples in our test dataset. Depending on the original sample, the operations to create a certain test group thus include 1) changing a mention, 2) adding a mention or 3) keeping a mention unchanged (in case of a patient case that is already part of the test group at hand). This results in one newly created dataset per test group, all based on the same patient cases and only different in the patient characteristic under investigation.

Prediction analysis. After creating the test groups, we collect the models' predictions for all cases in each test group. Different from earlier approaches to behavioral testing we do not test whether predictions on the altered samples are true or false with regard to the ground truth. As van Aken et al. (2021) pointed out, clinical ground truth must be viewed critically, because the collected data does only show one possible pathway for a patient out of many. Further, existing biases in treatments and diagnoses are likely included in our testing data potentially leading to meaningless results. To prevent that, we instead focus on detecting how the model outputs change regardless of the original annotations. This way we can also evaluate very rare mentions (e.g. *transgender*) and observe their impact on the model predictions reliably. Figure 2 shows a schematic overview of the functioning of the framework.

Extensibility. In this study, we use the introduced framework to analyse model behavior with regard to patient characteristics as described in 4.2. However, it can also be used to test other model behavior like the ability to detect diagnoses when certain indicators are present in the text or the influence of stigmatizing language (cf. Goddu et al. (2018)). It is further possible to combine certain patient groups to test model behavior regarding intersectionality. While such analyses are beyond the scope of this paper, we include them in the published codebase as an example for further extensions.

4 Case Study: Patient Characteristics

4.1 Data

We conduct our analysis on data from the MIMIC-III database (Johnson et al., 2016). In particular we use the outcome prediction task setup by van Aken et al. (2021). The classification task includes 48,745 English admission notes annotated with the patients' clinical outcomes at discharge. We select the outcomes *diagnoses at discharge* and *in-hospital mortality* for this analysis, since they have the highest impact on patient care and present a high potential to disadvantage certain patient groups. We use three models (see 4.3) trained on the two *admission* to *discharge* tasks and conduct our analysis on the test set defined by the authors with 9,829 samples.

4.2 Considered Patient Characteristics

We choose three characteristics for the analysis in this work: *Age*, *gender* and *ethnicity*. While these characteristics differ in their importance as clinical risk factors, all of them are known to be subject to biases and stigmas in health care (Stangl et al.,

2019). Therefore, we want to test, whether the analysed models have learned medically plausible patterns or ones that might be harmful to certain patient groups. We deliberately also include groups that occur very rarely in the original dataset. We want to understand the impact of imbalanced input data especially on minority groups, since they are already disadvantaged by the health care system (Riley, 2012; Bulatao and Anderson, 2004).

When altering the samples in our test set, we utilize the fact that patients are described in a mostly consistent way in clinical notes. We collect all mention variations from the training set used to describe the different patient characteristics and alter the samples accordingly in an automated setup. Details regarding all applied variations can be found in the public repository linked in 1.

Age. The age of a patient is a significant risk factor for a number of clinical outcomes. Our test includes all ages between 18 and 89 and the [** Age over 90**] de-idenfitication label from the MIMIC-III database. By analysing the model behavior on changing age mentions we can get insights on how the models interpret numbers, which is considered challenging for current NLP models (Wallace et al., 2019).

Gender. A patient's gender is both a risk factor for certain diseases and also subject to unintended biases in healthcare. We test the model's behavior regarding gender by altering the gender mention and by changing all pronouns in the clinical note. In addition to *female* and *male*, we also consider *transgender* as a gender test group in our study. This group is extremely rare in clinical datasets like MIMIC-III, but since approximately 1.4 million people in the U.S. identify as transgender (Flores et al., 2016), it is important to understand how model predictions change when the characteristic is present in a clinical note.

Ethnicity. The ethnicity of a patient is only occasionally mentioned in clinical notes and its role in medical decision-making is controversial, since it can lead to disadvantages in patient care (Anderson et al., 2001; Snipes et al., 2011). Earlier studies have also shown that ethnicity in clinical notes is often incorrectly assigned (Moscou et al., 2003). We want to know how clinical NLP models interpret the mention of ethnicity in a clinical note and whether their behavior can cause unfair treatment. We choose *White*, *African American*, *Hispanic* and *Asian* as ethnicity groups for our evaluation, as they are the most frequent ethnicities in MIMIC-III.

	PubMedBERT	CORe	BioBERT
Diagnoses	**83.75**	83.54	82.81
Mortality	**84.28**	84.04	82.55

Table 1: Performance of three state-of-the-art models on the tasks *diagnoses* (multi-label) and *mortality prediction* (binary task) in % AUROC. PubMedBERT outperforms the other models in both tasks by a small margin.

4.3 Clinical NLP Models

In this study, we apply the introduced testing framework to three existing clinical models which are fine-tuned on the tasks of diagnosis and mortality prediction. We use public pre-trained model checkpoints and fine-tune all models on the same training data with the same hyperparameter setup[2]. The models are based on the BERT architecture (Devlin et al., 2019) as it presents the current state-of-the-art in predicting patient outcomes. Their performance on the two tasks is shown in Table 1. We deliberately choose three models based on the same architecture to investigate the impact of pre-training data while keeping architectural considerations aside. In general the proposed testing framework is model agnostic and works with any type of text-based outcome prediction model.

BioBERT. Lee et al. (2020) introduced BioBERT which is based on a pre-trained BERT Base (Devlin et al., 2019) checkpoint. They applied another language model fine-tuning step using biomedical articles from PubMed abstracts and full-text articles. BioBERT has shown improved performance on both medical and clinical downstream tasks.

CORe. Clinical Outcome Representations (CORe) by van Aken et al. (2021) are based on BioBERT and extended with a pre-training step that focuses on the prediction of patient outcomes. The pre-training data includes clinical notes, Wikipedia articles and case studies from PubMed. The tokenization is similar to the BioBERT model.

PubMedBERT. Gu et al. (2020) recently introduced PubMedBERT based on similar data as BioBERT. They use PubMed articles and abstracts but instead of extending a BERT Base model, they

[2]Batch size: 20; learning rate: 5e-05; dropout: 0.1; warmup steps: 1000; early stopping patience: 20.

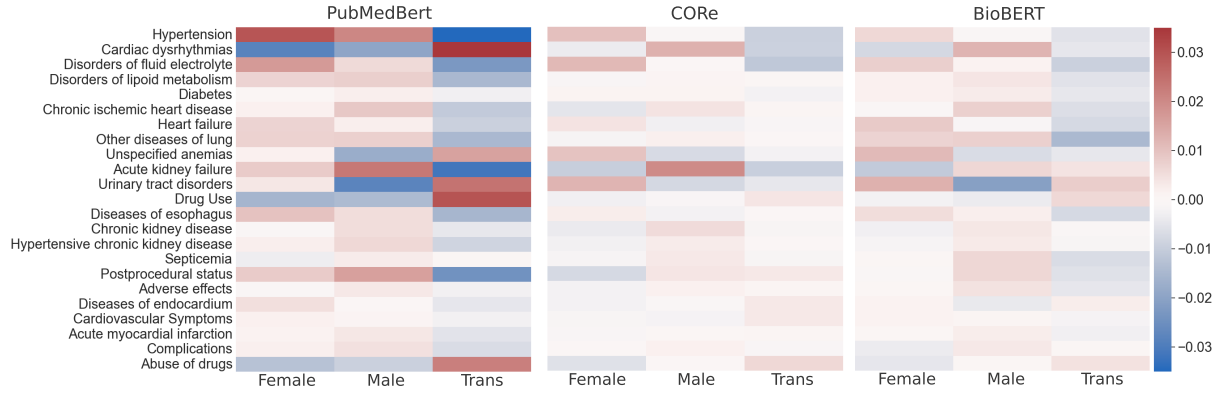

Figure 3: Influence of **gender** on predicted diagnoses. Blue: Predicted probability for diagnosis is below-average; red: predicted probability above-average. PubMedBERT shows highest sensitivity to gender mention and regards many diagnoses less likely if *transgender* is mentioned in the text. Graph shows deviation of probabilities on 24 most common diagnoses in test set.

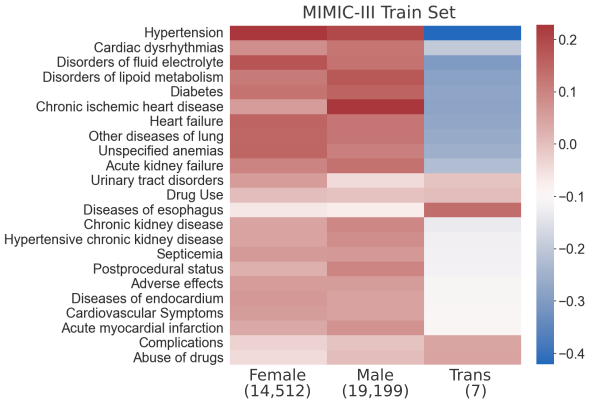

Figure 4: Original distribution of diagnoses per **gender** in MIMIC-III. Cell colors: Deviation from average probability. Numbers in parenthesis: Occurrences in the training set. Most diagnoses occur less often in transgender patients due to their very low sample count.

train PubMedBERT from scratch. The tokenization is adjusted to the medical domain accordingly. The model reaches state-of-the-art results on multiple medical NLP tasks and outperforms the other analysed models on the outcome prediction tasks.

5 Results

We present the results on all test cases by averaging the probabilities that a model assigns to each test sample. We then compare the averaged probabilities across test cases to identify which characteristics have a large impact on the model's prediction over the whole test set. The values per diagnosis in the heatmaps shown in Figure 3, 4, 7 and 8 are defined using the following formula:

$$c_i = p_i - \frac{\sum_j^N p_j}{N} \quad (1)$$

where c_i is the value assigned to test group i, p is the (predicted) probability for a given diagnosis and N is the number of all test groups except i.

We choose this illustration based on the concept of partial dependence plots (Friedman, 2001) to highlight both positive and negative influence of a characteristic on model behavior. Since all test groups are based on the same patients and only differ regarding the characteristic at hand, even small differences in the averaged predictions can point towards general patterns that the model learned to associate with a characteristic.

5.1 Influence of Gender

Transgender mention leads to lower mortality and diagnoses predictions. Table 2 shows the mortality predictions of the three analysed models with regard to the gender assigned in the text. While the predicted mortality risk for female and male patients lies within a small range, all models predict the mortality risk of patients that are described as transgender as lower than non-transgender patients. This is probably due to the relative young age of most transgender patients

	PubMedBERT	CORe	BioBERT
Female	**0.335**	0.239	0.119
Male	0.333	**0.245**	**0.121**
Transgender	0.326	0.229	0.117

Table 2: Influence of **gender** on mortality predictions. PubMedBERT assigns highest risk to female, the other models to male patients. Notably, all models decrease their mortality prediction for transgender patients.

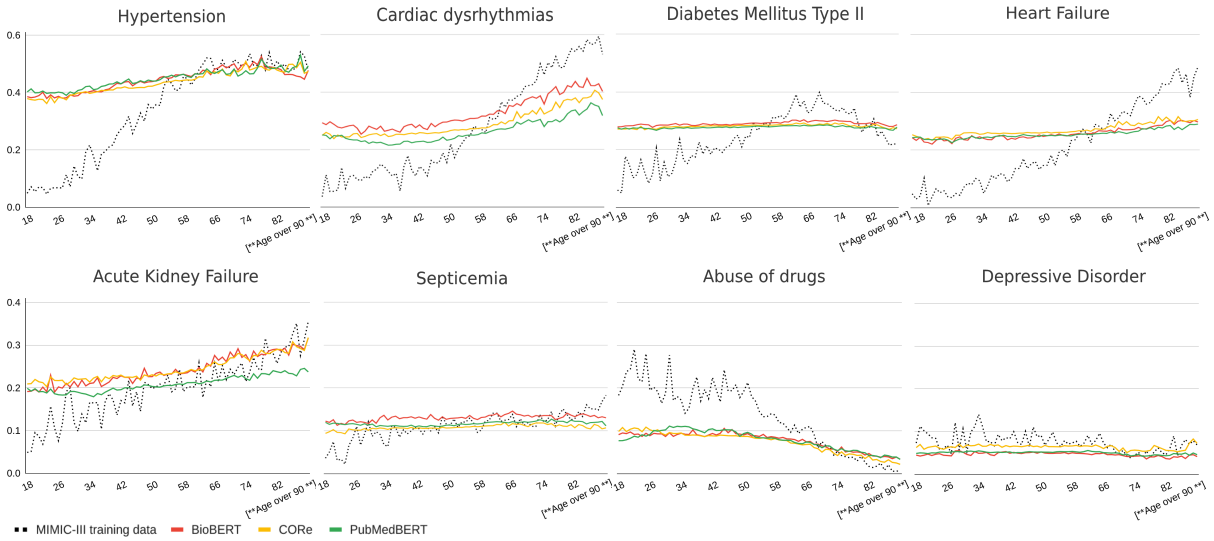

Figure 5: Influence of **age** on diagnosis predictions. The x-axis is the simulated age and the y-axis is the predicted probability of a diagnosis. All models follow similar patterns with some diagnosis risks increasing with age and some decreasing. The original training distributions (black dotted line) are mostly followed but attenuated.

in the MIMIC-III training data, but can be harmful to older patients identifying as transgender at inference time.

Sensitivity to gender mention varies per model. Figure 3 shows the change in model prediction for each diagnosis with regard to the gender mention. The cells of the heatmap are the deviations from the average score of the other test cases. Thus, a red cell indicates that the model assigns a higher probability to a diagnosis for this gender group. We see that PubMedBERT is highly sensitive to the change of the patient gender, especially regarding transgender patients. Except from few diagnoses such as *Cardiac dysrhythmias* and *Drug Use / Abuse*, the model predicts a lower probability to diseases if the patient letter contains the transgender mention. The CORe and BioBERT models are less sensitive in this regard. The most salient deviation of the BioBERT model is a drop in probability of *Urinary tract disorders* for male patients, which is medically plausible due to anatomic differences (Tan and Chlebicki, 2016).

Patterns in MIMIC-III training data are partially inherited. In Figure 4 we show the original distribution of diagnoses per gender in the training data. Note that the deviations are about 10 times larger than the ones produced by the model predictions in Figure 3. This indicates that the models take gender as a decision factor, but only among others. Due to the very rare occurrence of transgender mentions (only seven cases in the training data), most diagnoses are underrepresented for this group. This is partially reflected by the model predictions, especially by PubMedBERT, as described above. Other salient patterns such as the prevalence of *Chronic ischemic heart disease* in male patients are only reproduced faintly by the models.

5.2 Influence of Age

Mortality risk is differently influenced by age. Figure 6 shows the averaged predicted mortality per age for all models and the actual distribution from the training data (dotted line). We see that

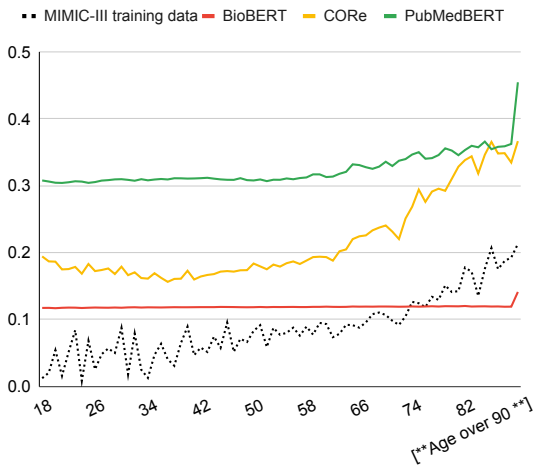

Figure 6: Influence of **age** on mortality predictions. X-axis: Simulated age; y-axis: predicted mortality risk. The three models are differently calibrated and only CORe is highly influenced by age.

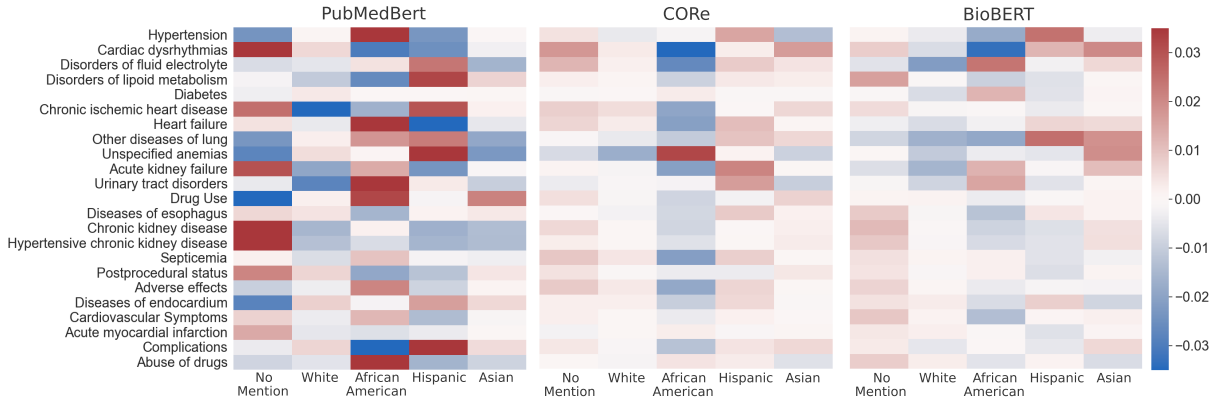

Figure 7: Influence of **ethnicity** on diagnosis predictions. Blue: Predicted probability for diagnosis is below-average; red: predicted probability above-average. PubMedBERT's predictions are highly influenced by ethnicity mentions, while CORe and BioBERT show smaller deviations, but also disparities on specific groups.

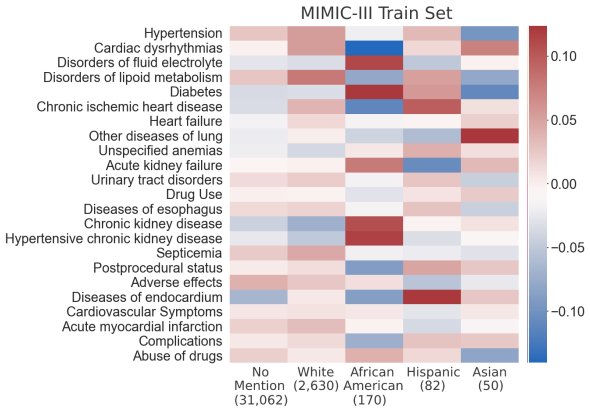

Figure 8: Original distribution of diagnoses per **ethnicity** in MIMIC-III. Cell colors: Deviation from average probability. Numbers in parenthesis: Occurrences in the training set. Both the distribution of samples and the occurrences of diagnoses are highly unbalanced in the training set.

BioBERT does not take age into account when predicting mortality risk except for patients over 90. PubMedBERT assigns a higher mortality risk to all age groups with a small increase for patients over 60 and an even steeper increase for patients over 90. CORe follows the training data the most while also inheriting peaks and troughs in the data.

Models are equally affected by age when predicting diagnoses. We exemplify the impact of age on diagnosis prediction on eight outcome diagnoses in Figure 5. The dotted lines show the distribution of the diagnosis within an age group in the training data. The change of predictions regarding age are similar throughout the analysed models with only small variations such as for *Cardiac dysrhythmias*. Some diagnoses are regarded more probable in older patients (e.g. *Acute Kidney Failure*) and others in younger patients (e.g. *Abuse of drugs*). The distributions per age group in the training data are more extreme, but follow the same tendencies as predicted by the models.

Peaks indicate lack of number understanding. From earlier studies we know that BERT-based models have difficulties dealing with numbers in text (Wallace et al., 2019). The peaks that we observe in some predictions support this finding. For instance, the models assign a higher risk of *Cardiac dysrhythmias* to patients aged 73 than to patients aged 74, because they do not capture that these are consecutive ages. Therefore, the influence of age on the predictions might solely be based on the individual age tokens observed in the training data.

5.3 Influence of Ethnicity

Mention of any ethnicity decreases prediction of mortality risk. Table 3 shows the mortality predictions when different ethnicities are mentioned and when there is no mention. We observe that

	PubMedBERT	CORe	BioBERT
No mention	**0.333**	**0.243**	**0.120**
White	0.329	0.235	0.119
African Amer.	0.329	0.239	0.116
Hispanic	0.331	0.237	0.118
Asian	0.330	0.238	0.118

Table 3: Influence of **ethnicity** on mortality predictions. The mention of an ethnicity decreases the predicted mortality risk. White and African American patients are assigned with the lowest mortality risk (gray-shaded).

the mention of any of the ethnicities leads to a decrease in mortality risk prediction in all models, with White and African American patients receiving the lowest probabilities.

Diagnoses predicted by PubMedBERT are highly sensitive to ethnicity mentions. Figure 7 depicts the influence of ethnicity mentions on the three models. Notably, the predictions of PubMedBERT are strongly influenced by ethnicity mentions. Multiple diagnoses such as *Chronic kidney disease* are more often predicted when there is no mention of ethnicity, while diagnoses like *Hypertension* and *Abuse of drugs* are regarded more likely in African American patients and *Unspecified anemias* in Hispanic patients. While the original training data in Figure 8 shows the same strong variance among ethnicities, this is not inherited the same way in the CORe and BioBERT models. However, we can also observe deviations regarding ethnicity in these models.

African American patients are assigned lower risk of diagnoses by CORe and BioBERT. The heatmaps showing predictions of CORe and BioBERT reveal a potentially harmful pattern in which the mention of *African American* in a clinical note decreases the predictions for a large number of diagnoses. This pattern is found more prominently in the CORe model, but also in BioBERT. Putting these models into clinical application could result in fewer diagnostic tests to be ordered by physicians and therefore lead to disadvantages in the treatment of African American patients. This is particularly critical as it would reinforce existing biases in health care (Nelson, 2002).

6 Discussion

Model behaviors show large variance. The results described in 5 reveal large differences in the influence of patient characteristics throughout models. The analysis shows that there is no overall *best* model, but each model has learned both useful patterns (e.g. age as a medical plausible risk factor) and potentially dangerous ones (e.g. decreases in diagnosis risks for minority groups). The large variance is surprising since the models have a shared architecture and are fine-tuned on the same data–they only differ in their pre-training. And while the reported AUROC scores for the models (Table 1) are close to each other, the variance in learned behavior show that we should consider in-depth analyses a crucial part of model evaluation in the clinical domain. This is especially important since harmful patterns in clinical NLP models are often fine-grained and difficult to detect.

Model scoring can obfuscate critical behavior. The analysis has shown that PubMedBERT which outperforms the other models in both mortality and diagnosis prediction by AUROC show larger sensitivity to mentions of gender and ethnicity in the text. Many of them–like lower diagnosis risk assignment to African American patients–might lead to undertreatment. This is alerting since it particularly affects minority groups which are already disadvantaged by the health care system. It also shows that instead of measuring clinical models regarding rather abstract scores, looking at their potential impact to patients should be further emphasized. To communicate model behavior to medical professionals one possible direction could be to use behavioral analysis results as a part of clinical model cards as proposed by Mitchell et al. (2019).

Limitations of the proposed framework. Unlike other behavioral testing setups (see 2.2), results of our framework cannot be easily categorized into *correct* and *false* behavior. While increased risk allocations can be beneficial to a patient group due to doctors running additional tests, they can also lead to mistreatment or other diagnoses being overlooked. Same holds for the influence of rare mentions, such as *transgender*: One could argue that based on only seven occurrences in the training set the characteristic should have less impact on model decisions overall. However, some features e.g. regarding rare diseases should be recognized as important even if very infrequent. Since our models often lack such judgement, the decision about which patient characteristic to consider a risk factor and their impact on outcome predictions is still best made by medical professionals. Nevertheless, decision support systems can be beneficial if their behavior is transparently communicated. With this framework we want to take a step towards improving this communication.

7 Conclusion

In this work, we introduced a behavioral testing framework for the clinical domain to understand the effects of textual variations on model predictions. We apply this framework to three current clinical NLP models to examine the impact of cer-

tain patient characteristics. Our results show that the models–even with very close AUROC scores–have learned very different behavioral patterns, some of them with high potential to disadvantage minority groups. With this work, we want to emphasize the importance of model evaluation beyond common metrics especially in sensitive areas like health care. We recommend to use the results of these evaluations for discussions with medical professionals. Being aware of specific model behavior and incorporating this knowledge into clinical decision making is a crucial step towards safe deployment of such models. For future work we consider iterative model fine-tuning with medical professionals in the loop a promising direction to teach models which patterns to stick to and which ones to discard.

Acknowledgments

We would like to thank Dr. med. Simon Ronicke for the valuable input. Our work is funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) under grant agreement 01MD19003B (PLASS) and 01MK2008MD (Servicemeister).

References

Matthew Anderson, Susan Moscou, Celestine Fulchon, and Daniel Neuspiel. 2001. The role of race in the clinical presentation. *Family medicine*, 33:430–4.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *SIGCIS, Philadelphia, PA*.

Boris Beizer. 1995. *Black-box testing: techniques for functional testing of software and systems*. John Wiley & Sons, Inc.

Rodolfo A Bulatao and Norman B Anderson. 2004. Understanding racial and ethnic differences in health in late life: A research agenda. *National Academies Press (US)*.

Edward Choi, Cao Xiao, Walter F. Stewart, and Jimeng Sun. 2018. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4552–4562.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

A.R. Flores, J.L. Herman, G.J. Gates, and T.N.T. Brown. 2016. How many adults identify as transgender in the united states? *Los Angeles, CA: The Williams Institute*.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. 2014. Unfolding physiological state: mortality modelling in intensive care units. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 75–84. ACM.

Anna Goddu, Katie O'Conor, Sophie Lanzkron, Mustapha Saheed, Somnath Saha, Carlton Haywood, and Mary Catherine Beach. 2018. Do words matter? stigmatizing language and the transmission of bias in the medical record. *Journal of General Internal Medicine*, 33.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Sarthak Jain, Ramin Mohammadi, and Byron C. Wallace. 2019. An analysis of attention over clinical notes for predictive tasks. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 15–21, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yohan Jo, Lisa Lee, and Shruti Palaskar. 2017. Combining lstm and latent topic modeling for mortality prediction. *arXiv preprint arXiv:1709.02842*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*, 3(1):1–9.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. 2019. Using clinical notes with time series data for ICU management. In *Proceedings of the 2019 Conference on Empirical Methods*

in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6432–6437, Hong Kong, China. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Dianbo Liu, Dmitriy Dligach, and Timothy Miller. 2019a. Two-stage federated phenotyping and patient representation learning. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 283–291, Florence, Italy. Association for Computational Linguistics.

Dianbo Liu, Dmitriy Dligach, and Timothy Miller. 2019b. Two-stage federated phenotyping and patient representation learning. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 283–291, Florence, Italy. Association for Computational Linguistics.

Jingshu Liu, Zachariah Zhang, and Narges Razavian. 2018. Deep EHR: chronic disease prediction using medical notes. In *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2018, 17-18 August 2018, Palo Alto, California*, volume 85 of *Proceedings of Machine Learning Research*, pages 440–464. PMLR.

C'ecile Log'e, Emily L. Ross, David Yaw Amoah Dadey, Saahil Jain, Adriel Saporta, Andrew Y. Ng, and Pranav Rajpurkar. 2021. Q-pain: A question answering dataset to measure social bias in pain management. *ArXiv*, abs/2108.01764.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Susan Moscou, Matthew R Anderson, Judith B Kaplan, and Lisa Valencia. 2003. Validity of racial/ethnic classifications in medical records data: an exploratory study. *American journal of public health*, 93(7):1084–1086.

Alan Nelson. 2002. Unequal treatment: confronting racial and ethnic disparities in health care. *Journal of the national medical association*, 94(8):666.

Michel Oleynik, Amila Kugic, Zdenko Kasáč, and Markus Kreuzthaler. 2019. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. *Journal of the American Medical Informatics Association*, 26(11):1247–1254.

Emily R Pfaff, Miles Crosskey, Kenneth Morton, and Ashok Krishnamurthy. 2020. Clinical annotation research kit (clark): Computable phenotyping using machine learning. *JMIR medical informatics*, 8(1):e16042.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Wayne J Riley. 2012. Health disparities: gaps in access, quality and affordability of medical care. *Transactions of the American Clinical and Climatological Association*, 123:167.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Yuqi Si and Kirk Roberts. 2019. Deep patient representation of clinical notes via multi-task learning for mortality prediction. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2019:779–788.

Shedra Snipes, Sherrill Sellers, Adebola Tafawa, Lisa Cooper, Julie Fields, and Vence Bonham. 2011. Is race medically relevant? a qualitative study of physicians' attitudes about the role of race in treatment decision-making. *BMC health services research*, 11:183.

Anne L Stangl, Valerie A Earnshaw, Carmen H Logie, Wim van Brakel, Leickness C Simbayi, Iman Barré, and John F Dovidio. 2019. The health stigma and discrimination framework: a global, crosscutting framework to inform research, intervention development, and policy on health-related stigmas. *BMC medicine*, 17(1):1–13.

Isabel Straw. 2020. The automation of bias in medical artificial intelligence (AI): decoding the past to create a better future. *Artif. Intell. Medicine*, 110:101965.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Harini Suresh, Jen J. Gong, and John V. Guttag. 2018. Learning tasks for multitask learning: Heterogenous patient populations in the ICU. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 802–810. ACM.

Chee Tan and Maciej Chlebicki. 2016. Urinary tract infections in adults. *Singapore Medical Journal*, 57:485–490.

Yifeng Tao, Bruno Godefroy, Guillaume Genthial, and Christopher Potts. 2019. Effective feature representation for clinical text concept extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 1–14, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alexander Tuzhilin. 2020. Predicting clinical diagnosis from patients electronic health records using bert-based neural networks. In *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25-28, 2020, Proceedings*, volume 12299, page 111. Springer Nature.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online. Association for Computational Linguistics.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Dongyu Zhang, Jidapa Thadajarassiri, Cansu Sen, and Elke Rundensteiner. 2020a. Time-aware transformer-based network for clinical notes series prediction. In *Machine Learning for Healthcare Conference*, pages 566–588. PMLR.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020b. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120.

Yun Zhao, Qinghang Hong, Xinlu Zhang, Yu Deng, Yuqing Wang, and Linda Petzold. 2021. Bertsurv: Bert-based survival models for predicting outcomes of trauma patients. *arXiv preprint arXiv:2103.10928*.