On the Difficulty of Disentangling Race in Representations of Clinical Notes

Pre-trained Transformer-based language models are increasingly used for clinical predictive tasks [7, 13]. Prior work has shown that sensitive patient attributes such as race are implicitly encoded in learned representations of notes, which in turn implies that such models may exhibit biases and ultimately exacerbate disparities, if put into practice [1]. One possible approach to mitigate this risk is to *debias* [5] learned embeddings such that it is difficult or impossible to recover race from them. However, in some clinical tasks demographic information often *should* inform predictions [14]. Ideally one would therefore represent sensitive attributes in a transparent and controllable way that allows practitioners to decide when and to what extent predictions should rely on these attributes. A potential approach in this direction is to make the use of this information *explicit* by learning *disentangled representations* [15, 10, 4]. This would provide mechanisms to assess the change in model output (if any) as a function of race, and to this information only optionally.

With this motivation, we investigated the use of disentangled learning strategies for neural encoders over notes in patient EHR with the aim of isolating race information. Specifically, we evaluated: Gradient Reversal (GR) [6], Masked Transformers (Masked) [18], and Variational Autoencoders (VAE) [16]. We compared the impact of these methods on two pre-trained clinical language model encoders: Clinical BioBERT [2] and PubmedBERT[8]; and a BERT model pretrained from scratch on all clinical notes from MIMIC-III [11] following [17]. We conducted experiments with the phenotype classification benchmark from [9]. MIMIC-III is highly imbalanced both in terms of labels and demographic distribution (most minority groups are severely underrepresented). Therefore, we focused only on black and white patients and conditions with at least 20% of positive labels (8 conditions), because small sample sizes can lead to noisy interpretations of disparities [3].

We replicate the pipeline described in [9] and report AUROCs averaged over eight tasks, where for each task we average over five runs (with different random seeds). We also probe the representations to see if we can predict race (from the component of the representation which should not encode it after disentanglement). We sample 1000 patients and split them into train (80%), validation (10%), and test sets (10%). Race prediction model is a multi-layer feed forward neural network following [17]. We use notes of type Physician, Nursing, and Nursing/Other. Figure 1 depicts the impact of the disentanglement techniques on (a) phenotype, and, (b) race prediction. We observe the same trend across the different BERT models: Masking is unable to identify a subnetwork that captures clinical content without being predictive of race. GR and VAE result in representations that are less predictive of race to some extent but this results in a significant drop in downstream performance. This indicates that the methods not only fail to disentangle but that they also end up destroying relevant features [12].

In retrospect, this is somewhat unsurprising given prior results [1]: To the extent that clinical concepts in notes correlate with race, an encoder must either represent these (and be able to predict race) or not (and suffer in terms of predictive performance).

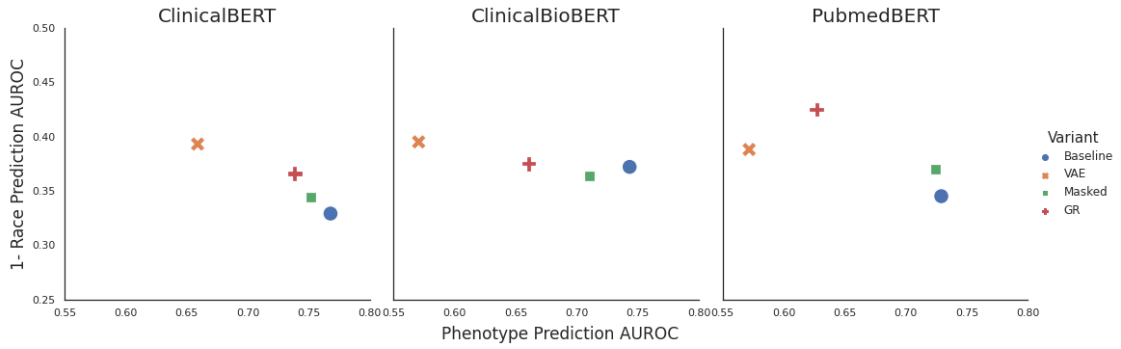

Figure 1: Impact of disentanglement techniques on race and phenotype prediction. **Baseline**: model without disentanglement; **VAE**: Variational Autoencoder, **Masked**: we create patient triplets (a, p, n), where a and p have the same diagnosis (based on ICD-9 codes) but different reported race while a and n have the same reported race but different diagnosis; **GR**: we model race as a token appended to each note and continue pretraining with a masked language modeling objective and two classifiers for race prediction: one with the race token and the other without; and perform gradient reversal on the latter.

References

[1] Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. Write it like you see it: Detectable differences in clinical notes by race lead to differential model recommendations. *arXiv preprint arXiv:2205.03931*, 2022.

[2] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[3] Silvio Amir, Jan-Willem van de Meent, and Byron C Wallace. On the impact of random seeds on the fairness of clinical classifiers. *arXiv preprint arXiv:2104.06338*, 2021.

[4] Pierre Colombo, Chloe Clavel, and Pablo Piantanida. A novel estimator of mutual information for learning to disentangle textual representations. *arXiv preprint arXiv:2105.02685*, 2021.

[5] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*, 2018.

[6] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[7] Sara Nouri Golmaei and Xiao Luo. Deepnote-gnn: predicting hospital readmission using clinical notes and patient network. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–9, 2021.

[8] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

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

[10] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. *arXiv preprint arXiv:1808.04339*, 2018.

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

[12] Abhinav Kumar, Chenhao Tan, and Amit Sharma. Probing classifiers are unreliable for concept removal and detection. *arXiv preprint arXiv:2207.04153*, 2022.

[13] Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pages 239–260. PMLR, 2021.

[14] Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan, and Nigam H Shah. Creating fair models of atherosclerotic cardiovascular disease risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 271–278, 2019.

[15] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1415–1424, 2017.

[16] Jake Vasilakes, Chrysoula Zerva, Makoto Miwa, and Sophia Ananiadou. Learning disentangled representations of negation and uncertainty. *arXiv preprint arXiv:2204.00511*, 2022.

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

[18] Xiongyi Zhang, Jan-Willem van de Meent, and Byron C Wallace. Disentangling representations of text by masking transformers. *arXiv preprint arXiv:2104.07155*, 2021.