FAIR PREDICTION OF RISK FOR ICU TREATMENTS FOR HYPOTENSION IN MINORITY SUBPOPULATIONS

Peniel N. Argaw* **& Esther Brown***
John A. Paulson School of Engineering and Applied Sciences
Harvard University
Cambridge, MA, USA
{peniel,estherbrown}@g.harvard.edu

Isaac S. Kohane
Department of Biomedical Informatics
Harvard University
Boston, MA, USA
isaac_kohane@hms.harvard.edu

ABSTRACT

Despite the large volumes of datasets available today, the majority of machine learning models are not able to equally represent all members of a population. In the clinical setting, issues around datasets and representation are exacerbated. For minority subpopulations, less representation in the data leads to biased classifiers in risk prediction models. Therefore, it is important to build models that increase the fairness of medical risk predictions for minority subpopulations. This work builds upon past work to propose risk prediction models for ICU treatments on hypotension that are fair for minority subgroups.

1 INTRODUCTION

A key goal of prediction models in a medical setting is to provide reliable risk predictions that correspond to observed risks of the patients, also known as calibration. Subpopulations that are well represented in the training data tend to have good calibration. However, minority subpopulations are generally not well represented in the dataset and as a result, are vulnerable to incorrect predictions compared to that of the majority subpopulation (Rajkomar et al., 2018). Likewise in low-resource settings, datasets with low dimensionality are prone to negative effects from prediction models used on high dimensional data (Boutilier et al., 2021). To improve fair prediction of risk for underrepresented groups, past work has looked at transfer learning and oversampling to address biased classifiers (Gao & Cui, 2020; Yin et al., 2019). Other work has shown the positive effects of re-calibration in minority populations when predicting risk in a medical setting (Barda et al., 2021). The main contribution of this work is to provide a process for improving biased classifiers using a multi-calibration technique. This technique improves model calibration scores across all subpopulations, specifically looking at the risk for hypotension in the ICU.

2 METHODOLOGY

This work initially evaluates fairness in a risk model applied to the MIMIC-III dataset (Johnson et al., 2016). MIMIC-III is a publicly available medical dataset with records of patients who stayed in the critical care units of the Beth Israel Deaconness Medical Center in Boston, MA between 2001 and 2012. We evaluated the risk for hypotension treatments using sixteen common labs and vitals. Then, we predicted for the likelihood of having a treatment (vasopressor, fluid bolus therapy or both) versus no treatment. A set of target minority populations were defined by splitting the population with a decision tree. The decision tree was trained on six protected demographic variables where the labels were the binary treatment or no treatment labels. The two models implemented for the risk prediction were Logistic regression (LR) and RNN. For the LR model, the inverse of regularization strength was set to 0.001, the class weights were balanced and included an L2 normalization penalty term. The RNN model was created using a lowered dimensional embedding that computes vector representations that are dependent on patients' past vitals in their ICU trajectory.

Many calibration techniques do not guarantee fair predictions for all subgroups Dwork et al. (2012), so we used the multi-calibration algorithm created by Hebert-Johnson et al. (2018) and adapted by Barda et al. (2021), in order to recalibrate the predictors across populations. The algorithm chooses a random subgroup from the set of subgroups, evaluates the maximum residual loss, checks if the loss exceeds the given threshold (α) and if so adjusts the predicted outputs by the magnitude of the violation on the deciles of the subgroup. Finally, the algorithm outputs a post-processed set of

*equal contribution

predictions for the training and testing populations which guarantees calibration for all subgroups above the given threshold.

Model performance was assessed through different discrimination and calibration metrics. Discrimination refers to a model's ability to separate classes and has been shown to be most useful in diagnostic settings. Calibration, on the other hand, assesses the agreement between true and predicted risk scores (Steyerberg et al., 2010). Discrimination was evaluated with the area under the curve (AUC). Calibration was evaluated with calibration-in-the-large (CITL) which provides an odds ratio of the mean difference between the true and predicted risk, and calibration slope (CS) which evaluates the average strength (coefficient of the logits) of the model. In a perfectly calibrated model, CITL and CS are both 1. Additionally, we asses statistical significance with p-values.

3 RESULTS

The results of the decision tree identified 15 subgroups with a tree depth of 6. Overall the model was able to predict the treatment outcomes with a precision of 0.79 for no treatment and 0.30 for treatment (overall accuracy of 0.677). We then tested the performance of each of the 15 groups in our trained LR. For the remainder of the results, we evaluated the effects of multi-calibration on the best and worst performing subgroups (best performing group had an AUC of 0.747, and the worst performing group had an AUC of 0.586). After applying multi-calibration, the LR model showed much greater improvement in the CITL and CS values. The same was seen in the subgroups as well (Table 1). Likewise the RNN model showed improvement in the CITL and CS values. AUC, however, showed little to no changes. In general, the RNN model performs better than the LR model across the different metrics because of the time dependency of the LSTM, which can learn additional representations of the patients over their trajectory in the ICU.

Table 1: Discrimination and calibration results from LR and RNN on varying training and testing scenarios (80/20 split respectively): trained and tested on the full population (full), trained on the full population, tested on the best performing subgroup (best), and trained on the full population, tested on the worst performing subgroup (worst).

Model	**CITL**	**CS**	**AUC**	p**-value**
Before multi-calibration				
LR (full)	0.315	0.788	0.695	<0.01
LR (best)	0.797	1.307	0.747	<0.01
LR (worst)	0.174	0.276	0.586	<0.01
RNN (full)	0.813	0.943	0.901	<0.01
RNN (best)	0.924	0.958	0.908	<0.01
RNN (worst)	0.724	0.905	0.896	<0.01
After multi-calibration				
LR (full)	0.923	0.848	0.695	<0.01
LR (best)	1.143	1.320	0.747	<0.01
LR (worst)	0.239	0.276	0.586	<0.01
RNN (full)	0.960	0.962	0.906	<0.01
RNN (best)	0.936	0.968	0.909	<0.01
RNN (worst)	1.101	0.925	0.902	<0.01

4 CONCLUSION

The application of the multi-calibration algorithm showed improvements in performance across all subgroups by assuring each subgroup's residuals were above a specified tolerance hyperparameter. Moreover, this algorithm can be applied to models that are sensitive to class imbalance and models that are time dependent, improving overall calibration. Though discrimination metrics like AUC provide helpful information regarding the model's ability to separate classes, it ignores possible biases in minority populations leading to false positives and negatives. Multi-calibration can lead to more fair risk scores, regardless of a patient's demographics or missing data, without negatively affecting discrimination. Future work will examine the effect of multi-calibration in other datasets, beyond that of previous work (Barda et al. (2021)), as well as show the performance on additional calibration and discrimination metrics. Subsequently, this work can be extended to low resource settings where lower dimensional data may be multi-calibrated to ameliorate predicted risk scores.

REFERENCES

Noam Barda, Gal Yona, Guy N Rothblum, Philip Greenland, Morton Leibowitz, Ran Balicer, Eitan Bachmat, and Noa Dagan. Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association*, 28:549–558, 3 2021. ISSN 1527-974X. doi: 10.1093/jamia/ocaa283.

Justin J Boutilier, Timothy C Y Chan, Manish Ranjan, and Sarang Deo. Risk stratification for early detection of diabetes and hypertension in resource-limited settings: Machine learning analysis. *Journal of Medical Internet Research*, 23:e20123, 1 2021. ISSN 1438-8871. doi: 10.2196/20123.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Yan Gao and Yan Cui. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nature Communications*, 11, 12 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18918-3.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. volume 80, pp. 1939–1948. PMLR, 4 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Alistair E.W. 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*, 3:160035, 12 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169:866–872, 12 2018. ISSN 0003-4819. doi: 10.7326/M18-1990. URL https://www.acpjournals.org/doi/abs/10.7326/M18-1990. doi: 10.7326/M18-1990.

Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. Assessing the performance of prediction models. *Epidemiology*, 21:128–138, 1 2010. ISSN 1044-3983. doi: 10.1097/EDE.0b013e3181c30fb2.

Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. pp. 5697–5706. IEEE, 6 2019. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.00585.