CPSC 464: Fairness in Predicting the Risk of Stillbirth and Preterm Pregnancy

Ivy Fan, Cove Geary, Bradley Yam
Professor: Nisheeth K. Vishnoi

May 3, 2021

Abstract

This paper reproduces results in machine learning fetal-health risk prediction and evaluates four models for fairness across racial categories. This paper demonstrates that state of the art models trained on large datasets can encode biases from the prevalence of health outcomes embedded in demographic trends that are driven by complex economic and sociological factors, hidden behind the label of race. Previous research has trained fetal-health risk prediction models but not one of this size on publicly available data, and the area of fairness in computer assisted diagnosis in pregnancy and obstetric care has not been explored yet. We used the same models and methods as the original authors, training Logistic Regression, Boosted Forests, and two Feed Forward Neural Networks on public CDC electronic health record (EHR) data. We find that minority black populations have a far higher fetal morbidity rate and fetal preterm birth rate as compared to other racial categories, which result in poorer accuracy on this population and an over-prediction of stillbirth and preterm risks for the black population. These biases can have dramatic repercussions for under-represented subgroups with poor health outcomes in real-world application, such as the prescription of unnecessary interventions, deterioration of mental health, and an increase in insurance premiums.

Contents

1 Introduction

Context: There has been a rapid growth of interest in applying machine learning methods to healthcare in medical decision making and computer-aided diagnosis (CAD) [Est+19]. Such research has always focused on using predictions to guide healthcare provision to promise more accurate diagnoses and thus better healthcare delivery. However, the framework of applying machine learning equitably and sensibly in the medical context is still nascent [GB20]. This paper aims to discuss the challenges of applying state of the art machine learning models with large data sets to the medical context, and in particular, we discuss the bias of predicting still-birth and preterm pregnancies with the models generated by Koivu and Sairanen in their paper [KS20].

Machine Learning in Pregnancy: Although the field is young, the literature on applying machine learning approaches to pregnancy related healthcare is prolific, with almost 127 papers published spanning four decades of research, as demonstrated by this survey by Davidson and Boland [DB21]. In some cases, these methods have been piloted in clinical trials and mobile health delivery systems, although not all of them are necessarily diagnostic.

Specifically in the field of fetal health prediction, there are about a dozen existing papers that attempt to diagnose fetal health outcomes which included preterm births and stillbirths. [AET18]. Overall, the sample sets involved in those studies are small - ranging from as little as 15 samples to a maximum of 9419 samples. Woolery et al's study, which targeted classifying preterm births, had 18,000 samples but achieved an accuracy of between 53-88%, which according to the authors, outperformed manual classification techniques [WG94].

In contrast, Koivu and Sairanen was the first study to incorporate enough publicly available CDC Electronic Health Record (EHR) data to create a dataset of about 11 million samples, which they then used to classify late stillbirths, early stillbirths and preterm pregnancies using state of the art machine learning models.

Research conducted by the Stillbirth Collaborative Research Network (SCRN) have also conducted in-depth studies into the relevance of Sociodemographic and Physiological risk factors for stillbirths. They found that pregnancy history was the strongest predictor for stillbirth, alongside other risk factors such as being unmarried and having diabetes. The study also notes that prepregnancy risk factors have been found to be more frequent among African Americans, and accounted for roughly 22% of the increase in risk compared to being white [Gro+11]. Koivu and Sairanen cite studies like these in selecting for their feature variables.

Despite the extensive research into the various risk factors contributing to stillbirths or preterm births, including genetic factors, much of the overall burden remains unexplained. [Gro+11] [AET18] [WL02] In particular, it seems like a large reduction in stillbirths have been attributed to more successful delivery methods. [HL06]

Interventions for Pregnancy Outcomes: The consequences of a misdiagnosis of a preterm birth can be fairly dramatic. In the case of a potential stillborn child, which may be indicated by a small-for-gestational-age signal (SGA), interventions include potentially inducing delivery early, expectant management of the fetus, or other therapeutic measures. Existing analysis only detects SGA about 1/4 of the time [Smi15].

The risks of intervention are equally costly for preterm births. These interventions can include hormone therapy through progesterone supplementation, or even cervical cleavage, which is the physical suturing of the cervix to prevent preterm dilation. This invasive procedure is only undertaken if the mother presents with an extensive history of preterm labour.[1]

Of course, the risks of missing a preterm birth or a stillbirth are also problematic, as the chance to prevent the delivery of an underdeveloped or deceased fetus is missed. The debate over how to weight these priorities is still an open one, and we decide in this paper to consider both outcomes as serious and weighty.

There are various tests that can be undertaken during pregnancy to assess the likelihood of a preterm birth or complications with the birth process, including testing for Group B Strep Culture, fetal heart monitoring, amniocentesis and ultrasounds.

Fairness in Medical Machine Learning: Machine Learning is becoming more popular in medical applications. However, (1) medical data is still disproportionately composed of people from different races and other protected attributes [PFS21], and (2) medical data can reflect disparities in care for patients from different vulnerable groups, as well as other forms of structural discrimination such as the way medical data is collected, or the prevalence of certain morbidities or diseases in some populations more than others [PFS21], [McC+20], [Sey+20]. This disparity is likely to cause unfair outcomes when models are trained on this data. Mis-predictions can have profound consequences in the decision to provide treatment or diagnosis to a given individual [GB20], [McC+20].

Motivation to study the problem: In this paper, we focus on a 2020 study that aims to predict birth outcomes based on a large amount of data from the U.S. Centers for Disease Control (”CDC”). In particular, we look at the protected attribute of *race*. A brief data exploration tells us that White people are overly represented (76.5%), whereas Black(15.4%), Asian and Pacific Islander(6.9%), Native American (1.1%) are under represented, or in the case of Latin Americans, not represented at all.

Furthermore, the authors of the paper have described their approach in detail, but have not open sourced their model or data preprocessing, making it difficult for the medical community to verify or test their results. As such, we feel it is important to recreate their process and examine their proposed model for unfair predictions.

The precise implications of models such as these on real world applications are necessarily contingent on their integration with policy and systems-level feedback. However, we can speculate about some of the potential effects of a biased algorithm based on existing applications and interventions.

[1] https://perinatalresources.org/effective-interventions-prevent-preterm-labor/

An imbalanced true positive rate (TPR) may lead to situations where one subgroup is receiving fewer detections of stillbirths and preterm births, which may directly lead to the lack of provision of healthcare disproportionately to these subgroups. Many stillbirths and preterm births already go unidentified, so it would not be difficult to imagine a world in which the situation improves for some subgroups but not others.

However, an imbalanced false positive rate (FPR) may be equally problematic. Given the invasive and dramatic nature of these interventions, a false positive may increase prenatal stress, inducing anxiety and reduced health for the mother, and lead to unnecessary and painful medical interventions. Moreover, a high false positive rate may also drive up insurance premiums for implicated subgroups.

Contributions.

- We present the first fairness analysis on published models specifically trained and evaluated on real-world CDC pregnancy data.
- Empirically, we find the True Positive Rate (TPR), False Positive Rate (FPR), and overall Accuracy of a variety of models trained on this problem: regularized logistic regression, gradient boosted trees, and two deep neural networks. We also conduct this analysis with analogous "race-unaware" models in order to help explain the "race-aware" models and find that disparities in TPR and FPR between demographic groups are diminished in the "race-unaware models".
- Crucially, we provide the code and documentation of our study, both to promote replicability of our findings and to encourage machine learning practitioners to conduct similar analyses in model fairness and explainability.
- We reflect on the policy implications for problems in medical machine learning.

2 Other related works:

This paper is primarily preceded by Pfohl et al [PFS21], who undertook the first broad based study of algorithmic fairness constraints on clinical risk prediction. The paper emphasizes that the appropriateness of fairness is unclear due to both ethical [GB20] and technical considerations. They find that imposing fairness constraints induce nearly-universal degradation of performance metrics, and a heterogenous effect of these constraints on fairness metrics. Beyond the technical difficulties, the authors note that the analysis of fairness in healthcare lack the context and causal awareness necessary to reason about the mechanisms that lead to health disparities. As a result, we have attempted to include that discussion explicitly in the particular domain of fetal-health risk prediction.

[Sey+20] measured bias in deep learning classifiers trained to output diagnoses based on chest X-rays with respect to protected attributes, specifically sex, race, and age. They trained convolutional neural network models on four data sets (three separate data sets and one aggregate from the previous three) and assessed the fairness of the trained models by comparing true positive rate disparities across subgroups of the protected attributes. The paper found non-trivial disparities in different subgroups across all data sets and all classification tasks (possible diagnoses), and that disparities were smallest when using the aggregate data set.

We apply a similar method to pregnancy data and models as in [KS20]. Koivu and Sairanen evaluate the robustness of four machine learning methods, logistic regression, gradient boosting decision tree, and two artificial neutral network models, in predicting early stillbirth, late stillbirth and preterm birth pregnancies. The models are trained on a CDC data set spanning 4 years and covering close to 16 million pregnancies, including demographic data such as race, age, and education level, and externally validated with another data set. They find that machine learning models outperform the control method of using a multivariate logistic regression model on the CDC test set in AUC and TPR at 10% FPR. However, the paper does not explore possible disparities in performance of the models across subgroups; that is, the paper does not investigate any implications for *group fairness* (see e.g. [GP17]).

3 Preliminaries / Problem statement

Our modeling problem is a traditional supervised learning problem, solved through empirical risk minimization (ERM). For each negative birth outcome considered in this paper—early stillbirth, late stillbirth, and preterm pregnancy—we follow [KS20] in training three binary classifiers per model. Hence, one classifier per model per outcome yields twelve total binary classifiers that we train and evaluate. For each classification problem, a label of $Y = 1$ indicates that the negative outcome happened in this case, while $Y = 0$ indicates otherwise.

Models: As with [KS20], who compare a variety of different machine learning approaches in order to identify novel methods for modeling risk in a clinical setting, we compare the same array of machine learning approaches in order to understand how they each perform along the protected subgroup of *race*. In particular, we develop and train the following models:

- Multivariate logistic regression, ℓ_2 regularized, solved by LBFGS optimization [Noc80]. (*"LR"*)
- Gradient boosted trees as implemented by LightGBM [Ke+17], with AUC used as early-stopping criterion. (*"GB"*)
- A deep neural network with two hidden layers, 70 and 80 units respectively, using LeakyReLU activation [Xu+15] and dropout on each hidden layer. (*"Lrelu"*)
- A deep neural network with two hidden layers, with the number of units equal to the dimensionality (features) of the input, using scaled exponential linear units (SELU) and "alpha" dropout on each hidden layer [Kla+17]. (*"Selu"*)

Code: All code for preprocessing the data, all code and parameters used for training the models, and all code for conducting the fairness analysis and producing charts may be found in the following repository: `https://github.com/bradleyyam/fair-child`.

4 Empirical results

4.1 Setup

Datasets: [KS20] used infant birth and death data of pregnancies from 2013 to 2016 as the training set for their models. These files are provided for public use by the Center for Disease Control, National Center of Health Statistics, accessible through the National Vital Statistics System. In reproducing their study, we combined data from the 2013-2016 Natality files and Fetal Death files to form the base of our data set, with a total of 15,883,784 live births and 209,781 fetal deaths. There are several discrepancies between their CDC base dataset and ours, indicating that those files might not have been the exact ones they used. One of the biggest differences is that our data set contains over twice as many cases of fetal death.

We selected 26 feature variables from the base dataset following [KS20]. However, because of changes made to the data collected on birth and death certificates, some fields were unavailable in 2013 and in the fetal death data set. For example, none of the fetal death files contain information on the mother's marital status, preterm birth history, and history for several sexually transmitted infections. The 2013 fetal death data only contained 7 of 26 listed feature variables. Only a few thousand rows contained data on assistive reproductive technology and infertility drug use, so these feature variables were dropped from our data sets.

The data was pre-processed to exclude mothers of less than 18 years of age, and live births with less than 12 weeks of gestational age. Many variables were normalized to zero mean and unit standard deviation: BMI, height, daily cigarettes smoked before pregnancy, parity, and number of previous terminations. One-hot encodings generated for nominal variables race and education.

The dataset was split into train, test, validation, and feature selection sets following a .7/.1/.1/.1 split. Although the authors dropped all rows with missing values, doing the same in our dataset would result in all data on fetal death being dropped from the data set, so we did not take this measure. Instead, a different number of feature variables were used for the stillbirth predictions as compared with preterm birth predictions. Since the fetal death dataset was missing fields, the dataset to train classifiers to predict stillbirths dropped the marital status, infections, and previous preterm births variables, as well as all rows with missing variables. In order to keep using the feature variables dropped from the stillbirth dataset, the preterm dataset dropped all rows from the fetal death dataset, as well as all other rows with missing variables.

Fairness through unawareness: To further investigate the role of the race/ethnicity variables in our model's predictions, we retrained all models via "fairness through unawareness." Per [GP17], "[a] predictor is said to achieve fairness through unawareness if protected attributes are not explicitly used in the prediction process." Hence, we dropped the binary race columns within the training data and retrained all models. As [GP17] note, fairness through unawareness is not sufficient to avoid discrimination when other features are available that may encode (or be correlated with) race; indeed, "various discriminatory practices have been documented following race-blind approach in education, housing, credit, criminal justice system."

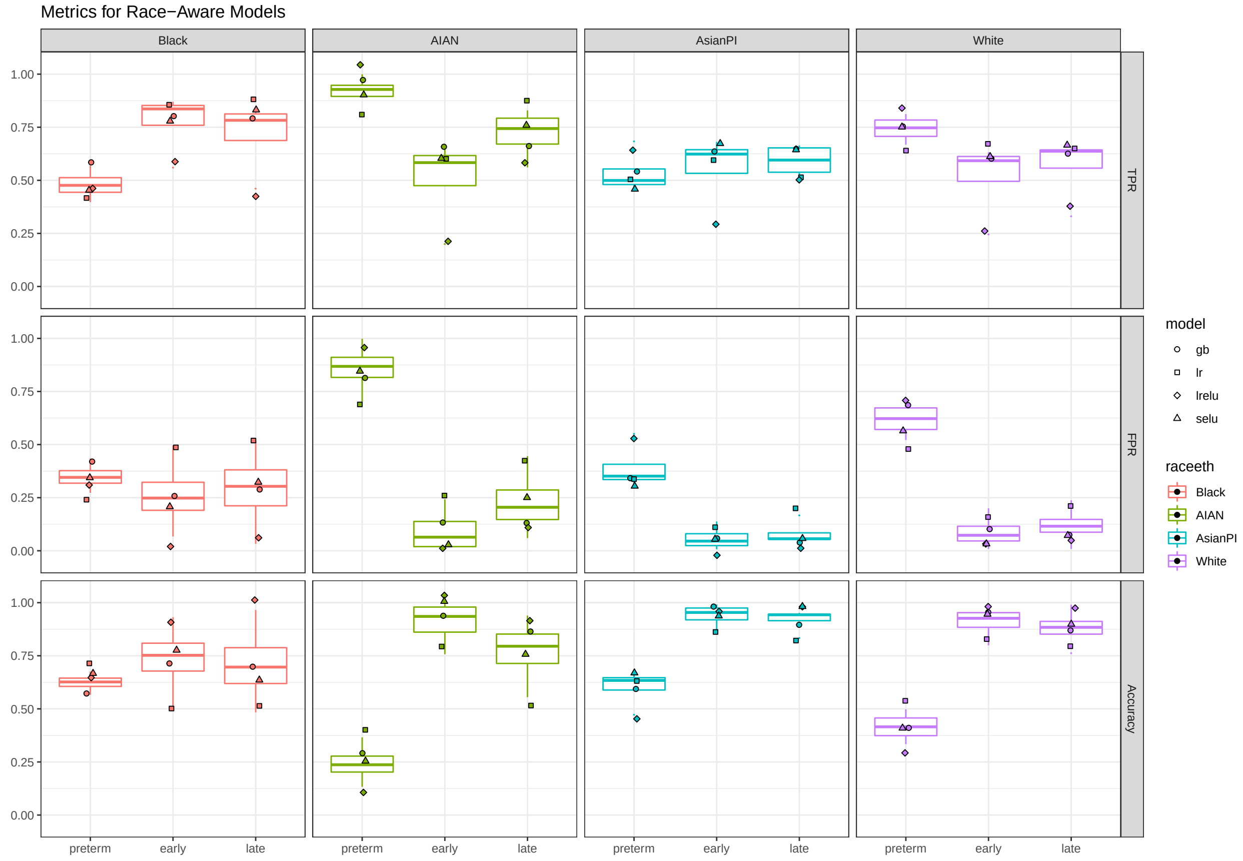


Figure 1: Metrics for the race-aware model. Columns represent each race variable, rows represent each metric, data points represent each model trained for each binary classification task of either preterm birth, late stillbirth, or early stillbirth. A small amount of jitter is applied to the data points.

4.2 Results

Figure 1 displays the race-aware results from our model evaluation, separated by certain racial/ethnic subgroups as defined by the U.S. census: AIAN (”American Indian and Alaskan Native”), Asian/Pacific Islander, Black, and White. The metrics AUC and ”TPR@10” follow the metrics reported by [KS20], where AUC indicates the area under the ROC curve of sensitivity-fallout (TPR-FPR), and TPR@10 indicates the True Positive Rate along the ROC curve at its point of 10% FPR. We newly report the models' True Positive Rate, False Positive Rate, and Accuracy. As discussed, we look to both TPR and FPR to evaluate model outcome-based group fairness, while Accuracy helps to confirm the workings of each model.

Despite taking on possibly different parameters than the models in [KS20], our models generally perform at least as well as those we were aiming to replicate. AUC and TPR@10 seem competitive or higher for both stillbirth classification tasks, while AUC and TPR@10 are somewhat lower for

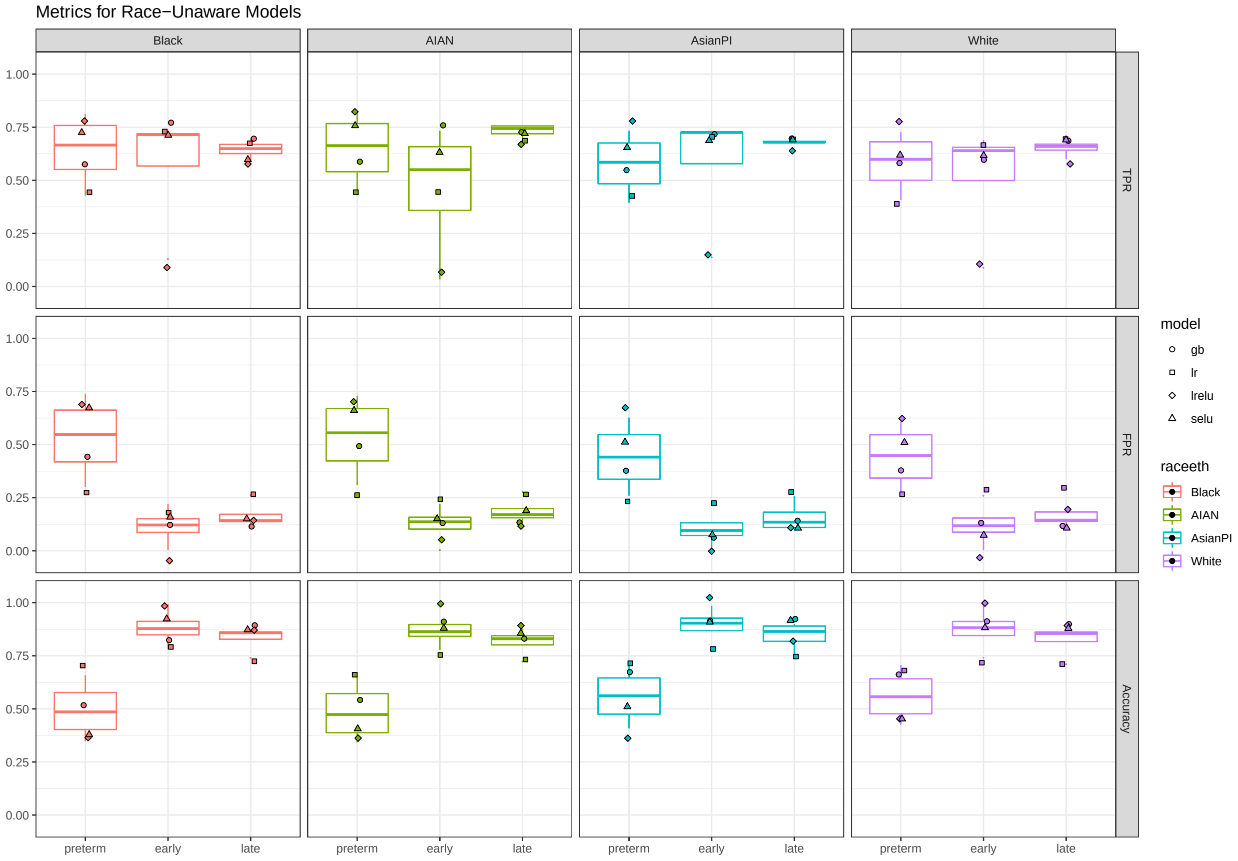


Figure 2: Metrics for the race-unaware model. Columns represent each race variable, rows represent each metric, data points represent each model trained for each binary classification task of either preterm birth, late stillbirth, or early stillbirth. A small amount of jitter is applied to the data points.

the preterm birth task.

Considering first the race-aware models, we note concerning trends in the positivity rates across multiple racial groups and tasks. Black and AIAN patients both were predicted to have a higher TPR compared to White patients, with Asian/PI in between. Further, Black and AIAN patients also were predicted to have lower FPR and somewhat lower (but still relatively high) accuracy. The results are less clear for preterm birth compared to the two stillbirth cases, as preterm birth is much less common in the data and generally a more difficult prediction task.

This trend seems to indicate that the models would mainly predict 1s in the case of Black and AIAN patients. Further, the reduced accuracy that we see alongside disproportionately mixed TPR (and FPR) rates seems to indicate that many of our models are placing high predictive weight on the race of the patient or race-correlated features.

Figure 2 displays the race-unaware results from our model evaluation. Interestingly, we note a significant equalizing of odds: TPR and FPR (and, hence, accuracy) metrics are now much more in line for all racial groups.

4.3 Discussion

The models trained with the race variable (race-aware models) seem to over-emphasize the significance of the race variable – although the proportion of adverse birth outcomes in the Black demographic is higher than in other demographics, the models attribute adverse birth outcomes to Black patients at a rate greater than the original disparity (see Figure 4). Exploration of the CDC dataset (shown below in Figure 3) suggests that the race_Black feature variable (binary, 1 or 0 to indicate if the patient is or is not Black) reflects greater structural inequalities and serves as an aggregate of the negative effects of feature variables correlated with having adverse birth outcomes. In particular, the Black race variable was most strongly correlated with a higher BMI index and participating in the Special Supplemental Nutrition Program for Women, Infants, and Children, which provides supplemental foods, healthcare referrals, and nutritional education to low-income women who are pregnant or breastfeeding. These in turn were strongly correlated with having pre-pregnancy diabetes, a known contributing factor to stillbirth outcomes.

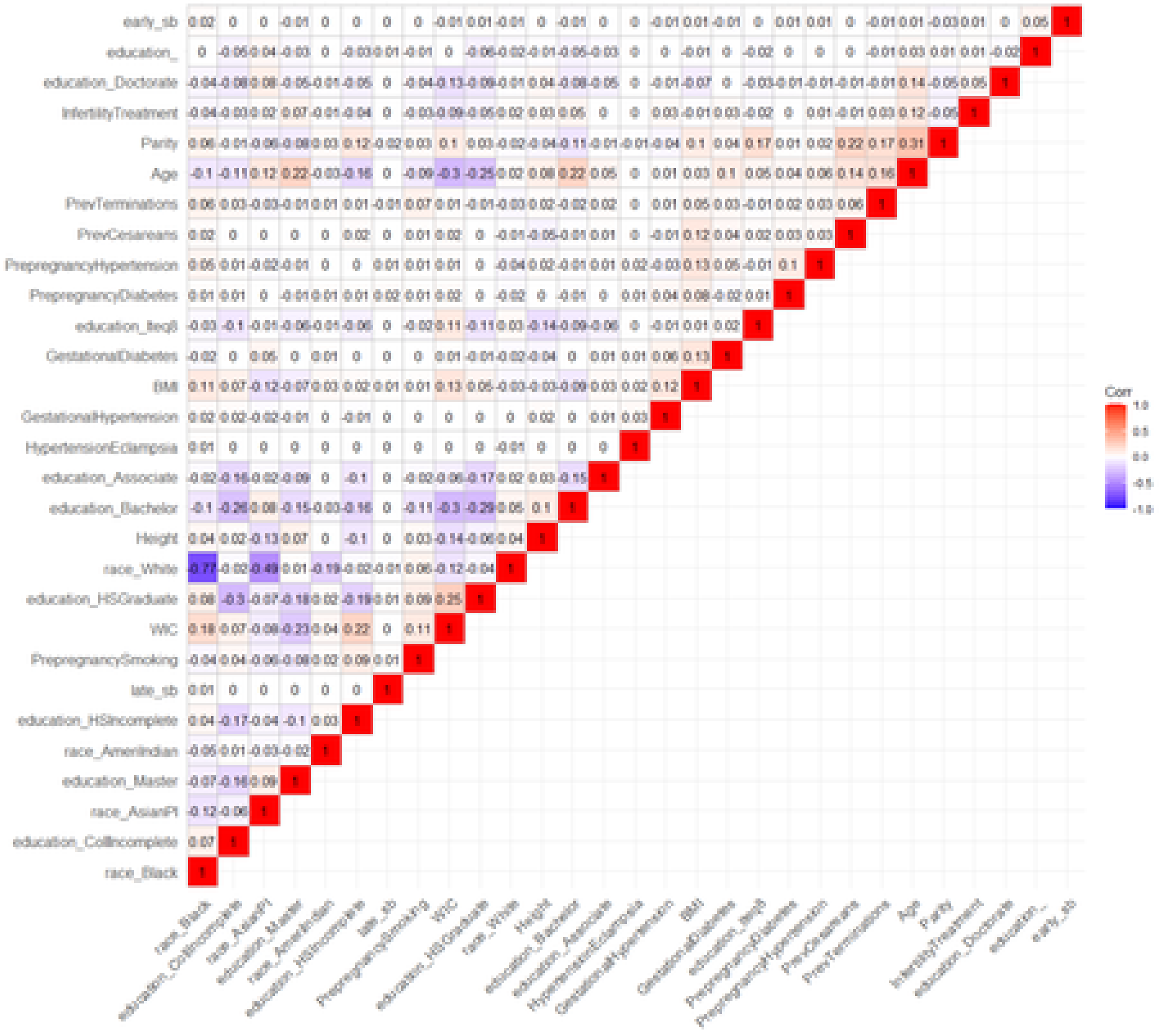
Furthermore, the Black race variable was positively correlated with variables indicating a highest education level of a high school diploma or partial completion of a high school or college diploma, while being negatively correlated to variables that indicated an education level of a Bachelor's degree or higher. In comparison, other races tended to be more positively correlated with variables that indicated higher levels of education, with the biggest difference being between the Black and White race variables. Although having a lower education level in and of itself likely does not have an effect on birth outcomes, it in turn may be correlated with lower income or access to medical resources. These trends suggest that the Black race variable could be acting as an amplifier for multiple other feature variables that have a small correlation individually with adverse birth outcomes. In contrast, the race_White variable had almost the exact opposite correlations with these sociodemographic and educational factors as compared to the race_Black variable.

Removing race as a feature variable entirely seems to mitigate disparities in TPR and FPR across racial cross-sections. These results suggest that this reduces the models' over-reliance on race as a predictor and that the Black variable was acting as an aggregate proxy for other features as opposed to vice versa. However, it is also worth noting that the outcome variables (in the figure denoted as late_sb and early_sb have relatively small correlations with all feature variables, suggesting that these features may be limited in predictive power.

5 Policy Implications

Pfohl et al's paper notes that predictive models may not be appropriately contextualized in terms of the heterogeneous impact of the complex policy interventions that they enable, and that model

Figure 3: Correlation plot between feature variables used to predict stillbirths. Each colored square represents the correlation between the feature variables of its column and row, labeled with significance level of the correlation.



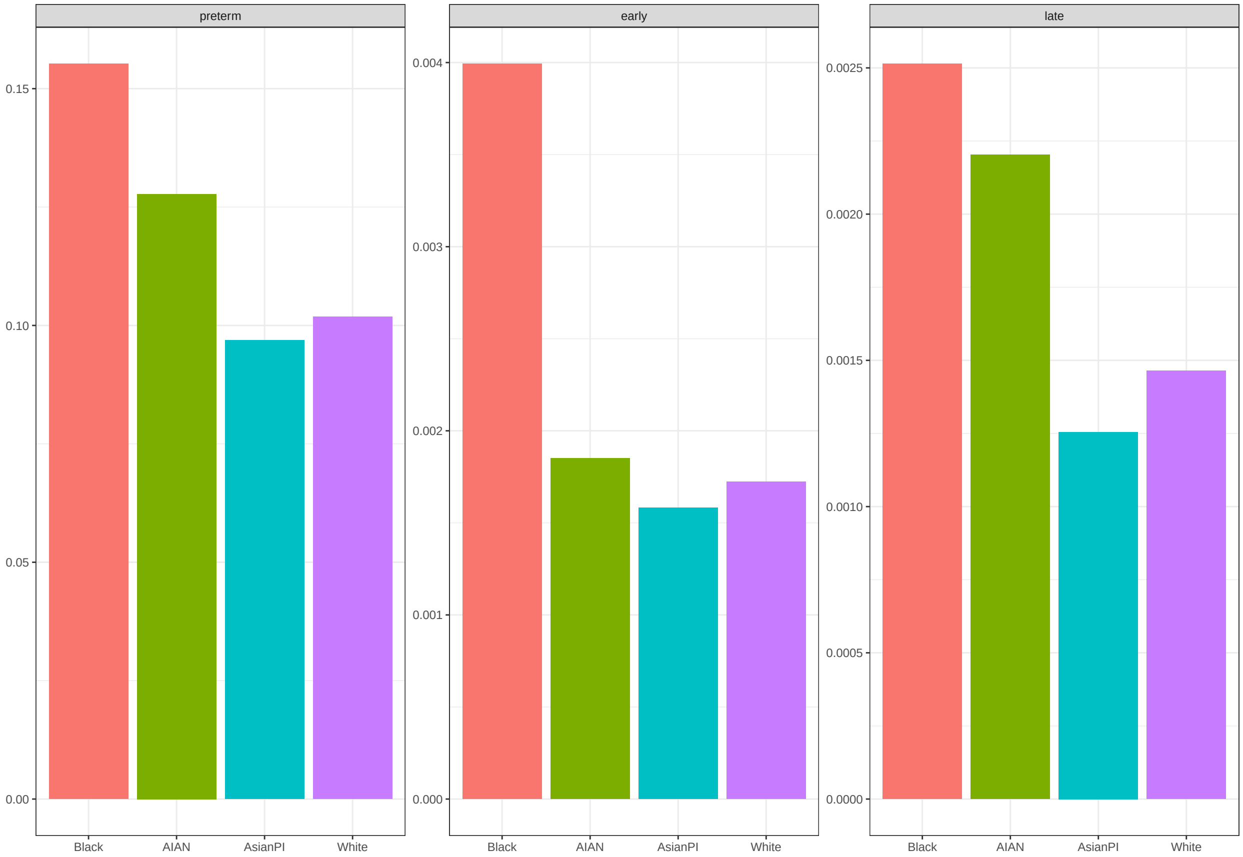


Figure 4: Breakdown of the positivity rates in the training data, for each task and for each race/ethnicity. Note that the y-axes have different scaling factors.

performance should not be conflated with accrual of benefit.[PFS21] In other words, a model should be assessed in terms of how its application creates (potentially disparate) benefits for real individuals. Likewise, a model's disparity in sensitivity or equalized odds is important insofar as it creates disparate benefits.

Rajkomar et al provides a helpful taxonomy of biases in the context of machine learning in healthcare and their resulting effects.[18] In our case, we are most concerned with allocation discrepancy, in which a difference in positive or negative predictions can lead to medical services being over or under prescribed for that group.

The implications of applying models biased toward over-predicting stillbirths and preterm births for minority populations can be dramatic in practice. The medical interventions for stillbirth and preterm births are non-trivial, traumatic and potentially fatal to both mother and child. In this specific medical context, we cannot afford to over-treat certain populations. Financially, this may also lead to insurance premiums being artificially driven up for minority populations based on race. In the worst possible case, an ignorant application of these models that result in over-treatment can generate a cycle that results in poorer fetal-health outcomes for minority populations, which generates even more biased data, which further reinforces the same kind of biases in newly trained models, resulting in a devilish self-fulfilling risk prophetic feedback loop.

The appropriate paradigm of distributive justice for rectifying an allocation bias is the equal allocation or demographic parity, in which resources are proportionally allocated to patients in the protected group with adjustments for comorbidities. In other words, patients in protected groups should receive resources that are proportional to their disease burden.

6 Conclusions, Limitations and Implications

6.1 Summary

In this study, we study the fairness of four classification algorithms trained to predict risk of early stillbirth, late stillbirth, and preterm birth based on pregnancy data. We find that minority black, and to some extent American Indian and Alaskan Native, populations that are under-represented in the dataset as a whole are over-represented in having higher risks for fetal morbidity and preterm birth. This is a reflection of complex historical, economic and sociological factors that are associated with the race label. All our models are overestimate the risk of stillbirth and preterm births for those specific minority populations as a result.

6.2 Limitations

In our attempt to replicate the work of [Sey+20], we discovered several discrepancies in model parameters and the original dataset. Most notably, marital status, history of previous preterm births, and history of several sexually transmitted diseases were missing from all fetal death data. Furthermore, our dataset contained more than twice as many cases with a stillbirth outcome. Our final dataset also contained slightly different demographic proportions. Because of these discrepancies, the fairness analysis of discrepancies in our trained models may not accurately reflect the fairness of performance of theirs. Missing data limits the ability of our models to train on those features, possibly causing them to rely more heavily other features that could be less relevant.

6.3 Concluding Remarks

Our analysis reveals that simply training race-unaware models can alleviate disparities in model predictions across protected groups. However, it also demonstrates that the inclusion of the race variable tends to be problematic in domains where race is highly correlated with other sociodemographic and health factors that collectively skew predictions, and where no other factor or groups of factors have a strong predictive power. Therefore, another sensible approach is collect more natality data during the pregnancy, including biological markers like genetic and proteomic data, while remaining sensitive to how race implicitly encodes many other factors. We explored race-unaware predictions in order to better-understand the predictions of the race-aware models, but we do *not* recommend that ”fairness through unawareness” be applied in the clinical machine learning context. Ultimately, even if model predictions are race unaware, applying models to policy must be race aware in order to pursue a more equitable allocation of medical resources.

6.4 Future directions

Our study leaves several directions for future work. In general, classifying stillbirths and preterm births with EHR data is already a difficult problem. In order to verify our hypotheses that the models put a disproportionate amount of weight on the predictive power of race, more causal exploration of the model is needed to be done. Emerging methods for ”explainable AI” (XAI) such as LIME may prove fruitful for improving understanding of a model's predictions [RSG16].

Beyond this study, classification models are applied to predict risk of other health outcomes besides birth outcomes. Furthermore, fairness could be evaluated with respect to other relevant protected attributes such as age and sex. It would be interesting to extend the analysis done in this paper to models trained in other contexts with respect to other protected attributes.

Most broadly, it remains a problem for clinicians, ML practitioners, ethicists, and policymakers to establish ethical and effective principles for applying machine learning methods in a way that is both fair and effective. It also to be seen if any fairness-constrained optimization techniques can further eliminate the disparities while preserving accuracy.

References

[Noc80] Jorge Nocedal. “Updating quasi-Newton matrices with limited storage”. In: *Mathematics of computation* 35.151 (1980), pp. 773–782.

[WG94] Linda K. Woolery and Jerzy Grzymala-Busse. "Machine Learning for an Expert System to Predict Preterm Birth Risk". In: *Journal of the American Medical Informatics Association* 1.6 (Nov. 1994), pp. 439–446. ISSN: 1067-5027. DOI: 10.1136/jamia.1994.95153433. eprint: https://academic.oup.com/jamia/article-pdf/1/6/439/2101525/1-6-439.pdf. URL: https://doi.org/10.1136/jamia.1994.95153433.

[WL02] Ronald J Wapner and Dawnette Lewis. "Genetics and metabolic causes of stillbirth". In: 26.1 (2002), pp. 70–74.

[HL06] Gary D.V. Hankins and Monica Longo. "The Role of Stillbirth Prevention and Late Preterm (Near-Term) Births". In: *Seminars in Perinatology* 30.1 (2006). Optimizing Care and Outcomes for Late Preterm (Near-Term)Infants: Part 1, pp. 20–23. ISSN: 0146-0005. DOI: https://doi.org/10.1053/j.semperi.2006.01.011. URL: https://www.sciencedirect.com/science/article/pii/S0146000506000127.

[Gro+11] Stillbirth Collaborative Research Network Writing Group et al. "Association between stillbirth and risk factors known at pregnancy confirmation". In: *JAMA: the journal of the American Medical Association* 306.22 (2011).

[Smi15] Gordon Smith. "Prevention of stillbirth". In: (2015).

[Xu+15] Bing Xu et al. "Empirical evaluation of rectified activations in convolutional network". In: *arXiv preprint arXiv:1505.00853* (2015).

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning". In: *arXiv preprint arXiv:1606.05386* (2016).

[GP17] Pratik Gajane and Mykola Pechenizkiy. "On formalizing fairness in prediction with machine learning". In: *arXiv preprint arXiv:1710.03184* (2017).

[Ke+17] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017), pp. 3146–3154.

[Kla+17] Günter Klambauer et al. "Self-normalizing neural networks". In: *arXiv preprint arXiv:1706.02515* (2017).

[AET18] Akhan Akbulut, Egemen Ertugrul, and Varol Topcu. "Fetal health status prediction based on maternal clinical history using machine learning techniques". In: *Computer methods and programs in biomedicine* 163 (2018), pp. 87–100.

[18] "Ensuring Fairness in Machine Learning to Advance Health Equity". In: *Annals of Internal Medicine* 169.12 (2018). PMID: 30508424, pp. 866–872. DOI: 10.7326/M18-1990. eprint: https://www.acpjournals.org/doi/pdf/10.7326/M18-1990. URL: https://www.acpjournals.org/doi/abs/10.7326/M18-1990.

[Est+19] Andre Esteva et al. "A guide to deep learning in healthcare". In: *Nature medicine* 25.1 (2019), pp. 24–29.

[GB20] Thomas Grote and Philipp Berens. "On the ethics of algorithmic decision-making in healthcare". In: *Journal of medical ethics* 46.3 (2020), pp. 205–211.

[KS20] Aki Koivu and Mikko Sairanen. "Predicting risk of stillbirth and preterm pregnancies with machine learning". In: *Health Information Science and Systems* 8.1 (2020). DOI: 10.1007/s13755-020-00105-9.

[McC+20] Melissa D McCradden et al. "Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning". In: *Journal of the American Medical Informatics Association* 27.12 (2020), pp. 2024–2027.

[Sey+20] Laleh Seyyed-Kalantari et al. "CheXclusion: Fairness gaps in deep chest X-ray classifiers". In: *Biocomputing 2021* (2020). DOI: 10.1142/9789811232701_0022.

[DB21] Lena Davidson and Mary Regina Boland. "Towards deep phenotyping pregnancy: a systematic review on artificial intelligence and machine learning methods to improve pregnancy outcomes". In: *Briefings in Bioinformatics* (Jan. 2021). bbaa369. ISSN: 1477-4054. DOI: 10.1093/bib/bbaa369. eprint: https://academic.oup.com/bib/advance-article-pdf/doi/10.1093/bib/bbaa369/35460043/bbaa369.pdf. URL: https://doi.org/10.1093/bib/bbaa369.

[PFS21] Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. "An empirical characterization of fair machine learning for clinical risk prediction". In: *Journal of biomedical informatics* 113 (2021), p. 103621.