# Making hospital readmission classifier fair – What is the cost?

**Sandro Radovanović, Andrija Petrović, Boris Delibašić, Milija Suknović**
University of Belgrade – Faculty of Organizational Sciences
Centre for Business Decision Making
Jove Ilića 154, 11000 Belgrade, Serbia
{sandro.radovanovic, boris.delibasic, milijas}@fon.bg.ac.rs,
apetrovic@mas.bg.ac.rs

***Abstract.*** *Creating predictive models using machine learning algorithms is often understood as a job where Data Scientist provides data to the algorithm without much intervention. With the rise of ethics in machine learning, predictive models need to be made fair. In this paper, we inspect the effects of pre-processing, in-processing and post-processing techniques for making predictive models fair. These techniques are applied to the hospital readmission prediction problem, where gender is considered as a sensitive attribute. The goal of the paper is to check whether unwanted discrimination between female and male in the logistic regression model exists and if exists to alleviate this problem making classifier fair. We employed logistic regression model which obtained AUC = 0.7959 and AUPRC = 0.5263. We have shown that reweighting strategy is a good trade-off between fairness and predictive performance. Namely, fairness is greatly improved, without much sacrificing predictive performance. We also show that adversarial debiasing is a good technique which combines predictive performance and fairness, and Equality of Odds technique optimizes Theil index.*

**Keywords.** Fairness, Machine Learning, Bias Mitigation, Hospital Readmission

## 1 Introduction

Machine learning is one of the most interesting fields of Computer Science. We can define machine learning as a system that has the ability to find rules, or more generally learn patterns, for the problem at hand based on historical data without being explicitly programmed (Witten et al., 2016). The idea is to induce new knowledge about the problem based on examples, observations, direct experience or instructions without or with very limited human intervention. It is increasingly popular in various fields as developed algorithms are created for general purpose. This means that machine learning can be applied in any area using only historical data. New knowledge is used as a decision-making tool which drives and optimize for certain outcomes.

Although machine learning models by its nature perform statistical discrimination, sometimes discrimination is systematic toward a certain group of people thus making classifier unfair. Additionally, often discrimination is made on sensitive attributes such as gender or race which is prohibited in many countries. However, if data used has bias the learned machine learning model will inherit that bias and most probably amplify it (Veale & Binns, 2017). This can lead to unethical outcomes such as discrimination of marginalized subpopulation due to underrepresentation in data, perceived race, ethnicity or other (Zarsky, 2016).

This unwanted bias raises the question of the applicability of machine learning models. Namely, consequences of those models can be catastrophic (i.e. convict certain person because of race or ethnicity (Berk & Hyatt, 2015)). An assumedly good solution to the raising concern is to remove features that can cause unwanted bias. However, removing those features does not guarantee that unwanted bias will be removed. For example, one subpopulation exists solely in one part of the feature space which is specific just for that subpopulation (i.e. certain ethnicity lives in a certain neighborhood). Therefore, other strategies need to be employed.

In this paper, we want to create a 30-day hospital readmission classification model with regard to possible gender bias. Hospital readmission is defined as unplanned hospital admission within 30 days after the previous discharge. The reasons for hospital readmission vary from lack of patient care or lack of hygiene to medical errors or diagnoses complications. Because hospital readmissions can be related to lack of medical care, it poses serious insurance threat for a medical institution. Namely, it is estimated that hospital readmissions costs in US hospitals sum up to $17-$29 billion per year (Zuckerman et al., 2016). Having that in mind, one can try to find a pattern that leads to hospital readmission using machine learning

algorithms. However, one wants to mitigate unwanted gender bias without sacrificing prediction accuracy.

Therefore, in this paper, we employ existing strategies for the mitigation of unwanted bias and show how much predictive performance suffers when unwanted bias is eliminated.

The paper is structured as follows. In Section 2 literature review is provided. In Section 3 we present methodology, with an explanation of data, experimental setup and used mitigation techniques. In Section 4 we discuss obtained results on pediatric hospital readmission dataset. We conclude the paper in Section 5.

# 2 Literature Review

What is fairness in machine learning systems? The answer is hard to formulate both qualitatively and quantitatively. Based on anti-discrimination laws that exist in the majority of countries one can define fairness as unequal treatment based on sensitive attributes, such as race, gender or religion (Žliobaitė, 2017). Most often unfair behavior of the decision model, machine learning or any other decision-making model, originates from bias.

However, formalizing fairness of the decision-making process must have at least two distinct notions of disparate treatment and disparate impact (Barocas and Selbst, 2016). Disparate treatment means that decisions are influenced, in less degree or greater degree, by sensitive attribute. It is often regarded as intentional discrimination or discrimination that is available in data collection process. Disparate impact means that decisions are disproportional between subjects with different values of sensitive attributes (i.e. certain subpopulation have greater benefit compared to other subpopulation). It is often regarded as unintentional discrimination or discrimination that is created by machine decision model. This is not the only part of fairness in machine learning models. Another notion of fairness discusses whether fairness means achieving parity or satisfying the preference (Gajane & Pechenizkiy, 2017).

If fairness is to be achieved in disparate treatment and parity then one can hide sensitive attributes from the learning process. This way algorithm is unaware of sensitive information, thus learned model will not explicitly use sensitive information in the decision-making process. For some specific tasks this approach did work, but there are many unacceptable models in practice (Marx, 2005; Taslitz, 2007). One can find critiques of this approach because sensitive attribute can be deduced from already available non-sensitive attributes. Another approach for disparate treatment and parity fairness is called counterfactual measures (Kusner et al., 2017). Namely, the algorithm is counterfactually fair if prediction remains the same no matter what the value of the sensitive attribute is.

Machine learning fairness is often understood in term Statistical parity which can be interpreted such that predictions should be approximately the same for individuals across the subgroups based on a sensitive attribute (Dwork et al., 2012). This way we are measuring disparate impact and parity. A similar measure is called equality of opportunity (Zafar et al., 2017a) which is defined that true positive rate should be the same across the subgroups regarding sensitive attribute.

For preference-based fairness definitions, one can find preferred treatment and preferred impact (Zafar et al., 2017b). The preferred treatment is satisfied if each group benefits (have a higher or lower value of predictions based on benefit direction) more from predictor compared to any other predictor. The prediction model is said to have preferred impact if the model has as much benefit as another model for all subgroups of subjects based on sensitive attributes.

Mitigation of bias in machine learning algorithm is a topic of increasing importance. Namely, consequences of the decision that origins from the machine learning model can come with a cost which hinders the usability of such models. However, determining the actual impact of an algorithm is very difficult. The impact may arise from limited data, inadequate parameter setting or even hard-coded rules for automatic decision making. In the era, deep learning and deep neural networks learned models introduce uncertainty over how and why decisions are made (Mittelstadt et al., 2016). Additionally, some decisions can be the result of noise which is often called “bug” example. Analysis, whether that example is a bug or systematic bias toward certain subpopulation, is practically impossible (Burrell, 2016).

To the best of our knowledge, there are three approaches to bias mitigation for fair predictions. Those are:

- Pre-processing techniques,
- In-processing techniques, and
- Post-processing techniques.

Pre-processing techniques are dataset transformation techniques that attempt to reduce or remove unwanted bias present in the original dataset. Transformation of the dataset can be suppression of the correlated columns, massaging the label attribute (i.e. changing labels of some data points in order to ensure disparate treatment), instance weighting or sampling (Kamiran & Calders, 2012). Suppression of the correlated columns seems like a good starting point for bias mitigation. However, this approach was abandoned since sensitive attribute can be represented using more complex rules, rather than plain correlation. Massaging the label attribute is rather an intrusive approach that can change the distribution of the data and consequently create less accurate prediction models (Calders et al., 2009; De Laat, 2018). Reweighing of the instances is used much more compared to other approaches because it ensures

complete removal of the unwanted bias using probability theory. Additionally, almost all (at least the most common) learning algorithms have the capability to include instance weights (Glauner & Valtchev, 2018; Suresh & Guttag, 2019). Another interesting approach is to transform data into latent space via learning a fair representation of data (Zemel et al., 2013; Johansson et al., 2016).

In-processing techniques for bias mitigation are a classification or regression models that simultaneously maximize prediction accuracy (or minimizing loss function) and reducing model ability to determine the sensitive attribute based on other attributes. In a paper (Kamishima et al., 2012) adds a regularization term to the loss function which is discrimination-aware regarding sensitive attribute. Adversarial debiasing (Zhang et al., 2018) learns an adversarial network which controls model accuracy and adversary's ability to determine the sensitive attribute from the predictions thus creating the fair classifier.

Post-processing techniques deal with predictions and try to adjust predictions in order to make predictions fair. Most often equal odds technique is employed. This technique solves a linear program to find probabilities with which to change output labels to optimize equalized odds (Pleiss et al., 2017).

# 3 Methodology

The methodology section consists of two parts. First, we give a brief overview of the dataset used and classification problem tackled. Next, we introduce the experimental setup and evaluation measures used.

## 3.1 Problem definition

In this paper, we want to create a classification model for hospital readmission based on patient diagnoses. Hospital readmission is defined as admission to the hospital within 30 days after discharge. In this paper, we utilized data from State Inpatient Database (SID), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality (NIS, 2011). SID database aims to track hospital admissions for all patients. One admission contains, besides demographic and insurance related data, diagnoses associated with admission. One patient can have at most 25 diagnoses for each hospital admission, having each diagnosis presented in ICD-9-CM code. Data preprocessing is performed in such a manner that one row presents one admission, while columns present diagnosis. Therefore, the dataset can contain over 15,000 attributes. We selected pediatric subpopulation of hospital readmission data from California in the period from January 2009 to December 2011. Additionally, we removed rare diagnoses, having selected only diagnoses that appeared in over 0.5% of admissions. The final dataset contains 66.994 hospital admissions and 851 diagnoses. It is worth to mention that label attribute (hospital readmission) is very skewed with the majority of the patients not readmitting to the hospital within 30 days after discharge (~16% readmit to the hospital).

The goal of the paper is to check whether unwanted discrimination between female and male in the logistic regression model exists and if exists to alleviate this problem making classifier fair. Since this will surely hinder classification performance our goal is to check what the cost of making classifier fair is.

## 3.2 Logistic regression

In this paper, we will learn and evaluate the logistic regression model. Logistic regression is a linear classification model (James et al., 2013) that can be presented in the form presented in Eq. 1.

$$\log\left(\frac{p}{1-p}\right) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n \qquad (1)$$

where $\log(p/(1-p))$ present logarithm of the odds ratio, $\theta$ coefficient associated to input attribute $x$ with exeption of $\theta_0$ which is called intercept. Positive value of $\theta$ tells us that increasing attribute $x$ increases odds of hospital readmission, while a negative value decreases the odds. One must optimize $\theta$ so the loss is minimized. Loss function is defined in eq. 2.

$$\min\big(L(\theta)\big) = \sum_{i=1}^{m} \log(1 + \exp(-y_i(x^T\theta + c)) \qquad (2)$$

where $x$ represent input attribute, $y$ binary attribute which is often called output or dependent attribute, $\theta$ coefficient associated with $x$ which are interpreted as weights of input attributes, $m$ number of examples and $c$ random noise. Loss function given in Formula 2 present maximum likelihood loss function obtained from Formula 1. It is convex and continuous, and therefore gradient descent can be applied.

## 3.3 Fairness measures

However, we would like to create fair logistic regression model regarding the gender of the patient. For that purposes, we will measure the disparate impact of the dataset and statistical parity of dataset before creating a prediction model. Disparate impact is defined as presented in Eq. 3.

$$\frac{P(y = 1|gender = female)}{P(y = 1|gender = male)} \qquad (3)$$

which present the ratio of readmitted patients of the female and male gender. Disparate impact ratio between 0.8 and 1.2 is acceptable, but more restrictive boundary can be applied i.e. from 0.9 to 1.1. From the disparate impact, we can derive measure commonly used in fairness definition called statistical parity. It is defined in Eq. 4.

$$P(y = 1|gender = female) - P(y = 1|gender = male) \qquad (4)$$

One can find that statistical parity is often called the mean difference between groups. (Zemel et al., 2013)

Both the disparate impact and statistical parity can be applied before the prediction model and on predictions obtained after the prediction model are created. Besides afore-mentioned measures, we will use additional measures which work only on predictions. First one is the Theil index (Speicher et al., 2018). Theil index is defined in Eq. 5.

$$T = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i}{\mu}\ln(\frac{b_i}{\mu}) \quad (5)$$

where $n$ is a number of examples, $b_i = \hat{y}_i - y_i + 1$ which can be interpreted as individual example deviation. Namely, if the correct value of the individual label is 1 (patient will readmit to the hospital within 30 days) and the predicted value is 0, then $b_i$ will be 0 – 1 + 1 = 0. Similarly, if the correct value is 0, and the predicted value is 1 then $b_i$ will be 1 – 0 + 1 = 0. Symbol $\mu$ present percentage of readmission within a sensitive group. Appropriate values of the Theil index are between 0 and 0.2. However, boundaries that are more restrictive can be used, i.e. from 0 to 0.1.

Finally, we will use the average odds difference between sensitive groups. The formula is presented in Eq. 6.

$$\frac{1}{2}\left((FPR_{s=U} - FPR_{s=P}) + (TPR_{s=U} - TPR_{s=P})\right) \quad (6)$$

where $FPR$ present false positive rate, i.e. the percentage of false alarms and $TPR$ true positive rate, i.e. the percentage of detection of hospital readmissions. One should notice additional information presented in eq. 6. False positive rates and true positive rates are calculated for unprivileged ($s = U$) and privileged ($s = P$) values of the sensitive attribute.

## 3.4 Experimental Setup

In this paper, we used data from January 2009 to December 2010 for training dataset where logistic regression model is learned, while evaluated on data originating from January 2011 to December 2011.

We will first train and evaluate plain logistic regression model. Evaluation of the model is done using afore-mentioned fairness metrics and also using standard evaluation metrics for binary classification which will be defined below. As a sensitive attribute which disturbs fairness, we use the gender of the patient. We further employ different strategies for mitigating bias regarding the gender of the patient.

The first strategy is called Reweighting (Kamiran & Calders, 2012) and this approach falls into preprocessing bias mitigation strategies. The idea of this approach is to assign a weight to an example in order to reduce possible bias presented in data. This compensation is done by weighting example with a ratio of the expected probability of sensitive attribute value and the class gave independence and the observed probability of sensitive attribute value and the class gave independence. The formula is presented in Eq. 7.

$$w_i = \frac{p_{\exp}(S = X(S) \wedge Class = X(Class))}{p_{obs}(S = X(S) \wedge Class = X(Class))} \quad (7)$$

where $i$ is an example, $S$ sensitive attribute, and $Class$ label attribute. Reweighted dataset will ensure that bias will be 0. Therefore, the predictive model which is learned on this dataset will have a higher probability to be discrimination-free (Kamiran & Calders, 2012). We will use weight information and inject it into logistic regression.

The second approach is to use a specialized algorithm for fair classification. For that purposes, we use the adversarial network (Zhang et al., 2018). The predictive model will use the neural network which corresponds to the generative adversarial network with two steps. The first step, called predictor tries to discriminate examples using input features. In this paper we utilize a deep neural network with three levels, each containing convolutional layer, normalization layer and sigmoid layer, followed by a max pooling layer. The second step is called the adversary step which is responsible for the satisfaction of fairness. (Zhang et al., 2018)

Finally, we employ postprocessing technique which finds the optimal value of decision threshold which for maximal satisfaction of fairness. This technique is called Equality of Odds and it solves linear programming problem to find the best decision threshold to optimize equalized odds. (Pleiss et al., 2017)

As evaluation measures, we used the area under the ROC curve (AUC). This measure calculates the area under the curve that is calculated using values of TPR and FPR for every possible decision threshold values. It can be interpreted as the probability that the random positive example (hospital readmission example) has a greater probability of readmission than random negative example (non-hospital readmission example). Since probability (or confidence) score for hospital readmission is not that important for AUC, but that positive example has a higher probability of readmission compared to negative example one can relate AUC with Mann-Whitney U test (Branco et al., 2016). Random classifier would have 0.5 for AUC. Values closer to one are better.

We used additional evaluation measure called area under the precision-recall curve (AUPRC). AUPRC is interpreted using the relation of true positive rate (recall) and precision. Namely, AUPRC can be viewed as a probability of positive example among those examples whose output values exceed a randomly selected threshold (Boyd et al., 2013). Higher values of AUPRC suggest that predictive model has greater power discriminating positive and negative examples, where random classifier would have a value of the ratio of positive class (class imbalance ratio). It is considered that AUPRC is more appropriate for class imbalance problems since true positive examples (which are most likely easy to discriminate due to class

imbalance) raises starting point (left side of AUC plot) when calculating AUC (Saito & Rehmsmeier, 2015).

# 4 Results

The results of the pre-model fairness metrics are presented in Table 1. Pre-model fairness metrics are used to check if there are errors in the data collection process, or if there exists a systematic difference in sensitive groups. If there exists unfairness in the dataset, one must consider not creating a predictive model based on that dataset since obtained predictions could lead to catastrophic consequences.

**Table 1.** Pre-model Fairness metric values

| | Original | Reweighting | Adversarial Network |
|---|---|---|---|
| Disparate Impact | 0.9092 | 1 | 0.9092 |
| Statistical Parity | -0.0160 | 0 | -0.0160 |

We can observe that Reweighted dataset has perfect fairness metrics. This is expected behavior because the reweighting procedure ensures that sensitive attribute will not have an effect on the Original and Adversarial network has the same values for pre-model fairness since they operate with the original dataset. Based on Dispare impact our dataset has a satisfactory level of fairness (acceptable level is between -0.1 and 0.1, but boundaries that are more restrictive are often used i.e. between -0.05 and 0.05). However, based on statistical parity our dataset can be considered just fair with value 0.9092. From Table 1 Equality of Odds is omitted since this technique operates on predictions.

After learning the logistic regression model we obtained performance measures presented in Table 2.

**Table 2.** Performance measures

| | Original | Reweighting | Adversarial Network |
|---|---|---|---|
| AUC | 0.7959 | 0.7953 | 0.7254 |
| AUPRC | 0.5263 | 0.5256 | 0.4411 |

The best performing model is the logistic regression model that uses original dataset. Logistic regression is the best one on both performance measures used, AUC and AUPRC. Obtained AUC is close 0.8 which is a very good performance for hospital readmission problem (Radovanovic et al., 2015). This means that our model discriminates between hospital readmission patients and non-hospital readmission patients, in such manner that ~80% of true hospital readmission patients has a greater probability of readmission compared to non-hospital readmission patients. Also, AUPRC is 0.5263 which is ~3.5 greater than class imbalance ratio. This can be interpreted that the learned predictive model is ~3.5 better than the random model.

Based on the performances of the other two approaches we can suggest that fairness comes with a cost (Zliobaite, 2015). Namely, one has to be aware of the tradeoff between predictive performance and fairness of the predictive model. We can see that Reweighting strategy did not lose on predictive performance. This is mainly due to the fact that the predictive model was logistic regression which utilized additional information in terms of weights. However, Adversarial network decreased in performance by ~7% in AUC and ~8% in AUPRC.

However, in order to inspect whether the predictive model is fair one must inspect post-model fairness. Namely, preparing a dataset to be fair is not good enough effort. If the predictive model is unfair, no matter how fair original dataset is, consequences of usage can be huge (i.e. discriminate a person based on gender or race).

Post-model fairness metrics are presented in Table 3. In this analysis, we include Equal Odds technique which is used to optimize the performance of the logistic regression model.

**Table 3.** Post-model Fairness metric values

| | Orig. | RW. | AN | EO |
|---|---|---|---|---|
| Disp. Impact | 0.9342 | 0.9781 | 0.9268 | 1.2841 |
| Stat. Parity | -0.0051 | -0.0017 | -0.0087 | 0.035 |
| Theil Index | 0.1304 | 0.1302 | 0.1274 | 0.0319 |
| Avg. Odds Diff. | 0.0028 | 0.0092 | -0.0088 | 0.1852 |

Original logistic regression model obtained better values for disparate impact and statistical parity compared to pre-model fairness values. Namely, the improvement was ~3% for disparate impact and 0.01 for statistical parity. This is an indicator that gender does not influence the predictive model with great impact. Also, the average odds difference is near perfect level with 0.0028. This can be interpreted that odds of hospital readmission for female and male pediatric hospital readmissions are nearly the same. However, concerning thing is that the Theil index is 0.1304, which is above a restrictive threshold value. This means that entropy which is caused by the groups is greater than expected. Performance metric values on the reweighted dataset (column RW) is better compared to the original dataset by every metric used. Similarly, Adversarial network (column AN) also has satisfactory fairness metrics, expect Theil index. We can notice that the values of those metrics are lower compared to logistic regression with the original

dataset. The only difference is the Theil index which is slightly better. Finally, predictions are optimized for expected odds in Equality of Odds technique. As we see Theil index is much lower, but with the cost of rising of other fairness metrics.

# 5 Conclusion

In this paper, we wanted to inspect what is the cost of the fairness of the predictive model on hospital readmission application. We tried three approaches for achieving fair model present in the literature. Namely, we tried adjusting dataset in order for dataset to be fair (pre-processing techniques). Then we trained and evaluated specialized predictive models which take into account information notion of fairness and makes an optimal tradeoff between being fair and accurate (in-processing techniques). Finally, we found the optimal decision threshold, which is most fair (post-processing technique).

We have shown that the predictive performance of the logistic regression model on hospital readmission data can be made without much sacrificing the fairness notion. Namely, the reweighting strategy has shown that dataset can be prepared to be perfectly fair before predictive model learning. In addition, post-model fairness measures improve without greater loss of predictive accuracy (in terms of AUC and AUPRC). Namely, predictive performance decreased for less than 0.01%, but fairness metrics increased ~4% even up to ~7%.

Adversarial networks, although mathematically sound, failed to beat logistic regression with AUC equal to 0.7254 and AUPRC equal to 0.4411. However, fairness metrics show that this model does a good job integrating discriminative and fairness notion into a single algorithm.

Finally, Equality of Odds optimizes the Theil index. However, this optimization comes with the cost of deterioration of other fairness metrics. Therefore, in order to use this technique, one must take the effects of post-processing optimization.

As a part of future work we will try to develop framework for integration of fairness metrics into logistic regression optimization procedure. One approach of integration would be through regularization, where coefficients of logistic regression would be controlled for fairness. Another approach where fairness could be achieved is by defining specific loss function while still remaining in regression like models. Namely, multi-goal function can be developed and learned using adversarial learning or meta-heuristics such as genetic algorithm, particle swarm optimization etc.

# Acknowledgments

This work was supported in part by the ONR/ONR Global under Grant N62909-19-1-2008.

# References

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*, 671.

Berk, R., & Hyatt, J. (2015). Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter*, *27*(4), 222-228.

Boyd, K., Eng, K. H., & Page, C. D. (2013, September). Area under the Precision-Recall Curve: Point estimates and Confidence Intervals. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 451-466). Springer, Berlin, Heidelberg.

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, *49*(2), 31.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512.

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009, December). Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops* (pp. 13-18). IEEE.

De Laat, P. B. (2018). Algorithmic decision-making based on machine learning from Big Data: Can transparency restore accountability?. *Philosophy & technology*, *31*(4), 525-541.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214-226). ACM.

Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*.

Glauner, P., Valtchev, P., & State, R. (2018). Impact of Biases in Big Data. *arXiv preprint arXiv:1803.00897*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 112, p. 18). New York: Springer.

Johansson, F., Shalit, U., & Sontag, D. (2016, June). Learning Representations for Counterfactual Inference. In *International Conference on Machine Learning* (pp. 3020-3029).

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1-33.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, September). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 35-50). Springer, Berlin, Heidelberg.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* (pp. 4066-4076).

Marx, S. (2005). Racism Without Racists: Colorblind Racism and the Persistence of Racial Inequality in the United States. *Contemporary Sociology*, *34*(6), 640.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

NIS, HCUP Nationwide Inpatient Sample (2011). Healthcare cost and utilization project (HCUP).

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems* (pp. 5680-5689).

Radovanovic, S., Vukicevic, M., Kovacevic, A., Stiglic, G., & Obradovic, Z. (2015, June). Domain knowledge based hierarchical feature selection for 30-day hospital readmission prediction. In *Conference on Artificial Intelligence in Medicine in Europe* (pp. 96-100). Springer, Cham.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018, July). A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2239-2248). ACM.

Suresh, H., & Guttag, J. V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002*.

Taslitz, A. E. (2007). Racial Blindsight: The Absurdity of Color-Blind Criminal Justice. *Ohio St. J. Crim. L.*, *5*, 1.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 2053951717743530.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017a, April). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1171-1180). International World Wide Web Conferences Steering Committee.

Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., & Weller, A. (2017b). From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems* (pp. 229-239).

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, February). Learning fair representations. In *International Conference on Machine Learning* (pp. 325-333).

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335-340). ACM.

Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*.

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060-1089.

Zuckerman, R. B., Sheingold, S. H., Orav, E. J., Ruhter, J., & Epstein, A. M. (2016). Readmissions, observation, and the hospital readmissions reduction program. *New England Journal of Medicine*, *374*(16), 1543-1551.