Fairness in Missing Data Imputation

Anonymous Authors[1]

Abstract

Missing data are ubiquitous in the era of big data and, if inadequately handled, are known to lead to biased findings and have deleterious impact on data-driven decision makings. To mitigate its impact, many missing value imputation methods have been developed. However, the fairness of these imputation methods across sensitive groups has not been studied. In this paper, we conduct the first known research on fairness of missing data imputation. By studying the performance of imputation methods in three commonly used datasets, we demonstrate that unfairness of missing value imputation widely exists and may be associated with multiple factors. Our results suggest that, in practice, a careful investigation of related factors can provide valuable insights on mitigating unfairness associated with missing data imputation.

1. Introduction

Missing data, as well-recognized, have significant impact on analysis of real data in many fields. One of the most popular approaches for handling missing data is missing data imputation. With the development of statistics and machine learning, different approaches have been adopted for the imputation task (Gondara & Wang, 2018; Li et al., 2019; Stekhoven & Bühlmann, 2012; Van Buuren & Oudshoorn, 1999; Yang et al., 2019; Yoon et al., 2018). While these machine learning imputation methods show preferable performances, potential concern about unfairness of the algorithms is non-negligible. Machine learning algorithms have been shown to inherit the bias and unfairness that human have in decision making. Various studies are conducted in the context of computer vision and natural language processing (Bolukbasi et al., 2016; Brunet et al., 2019; Buolamwini & Gebru, 2018; Gonen & Goldberg, 2019; Klare et al., 2012; Ryu et al., 2017; Zhang & Liu, 2020; Zhao et al., 2018; 2019b), illustrating the wide existence of unfairness and possible remedies. Algorithmic fairness, focusing on non-discrimination of decision outcomes, comes to the fore in the research community.

As discussed in Chouldechova & Roth (2018), Martínez-Plumed et al. (2019) and Pessach & Shmueli (2020), the causes of unfairness in machine learning mainly come from bias in dataset, missing values, prediction algorithms as well as imbalance of populations among different sensitive groups. In particular, it's discussed in Bakker et al. (2020), Martínez-Plumed et al. (2019), Rajkomar et al. (2018) and Gianfrancesco et al. (2018) that missing values contribute to the bias of algorithms. Empirical analysis in Martínez-Plumed et al. (2019) brings to the fore the issue of correctly handling missing data in the sense of fairness, instead of dropping out corresponding samples directly. Inspired by this work, we study the fairness of some representative imputation methods in three datasets*: COMPAS recidivism, Alzheimer's Disease Neuroimaging Initiative (ADNI) and atherosclerosis cardiovascular disease (ASCVD). By looking into imputation results and corresponding predictions for each imputation method and missing data mechanism, we observe non-negligible bias among different gender and race groups among all the imputation methods compared.

In the context of the impact missing data have on fairness, Martínez-Plumed et al. (2019) studies fairness of existing real datasets with missing values; Wang & Singh (2021) studies the impact of missing values in categorical data on fairness; Goel et al. (2020) illustrates the recoverability of outcome's distribution from complete cases, through a causal modeling of data missingness. Our work is the first work to systematically study the fairness associated with the process of imputing missing data.

Our contributions: In this paper, we propose the first fairness notion in imputation, *imputation accuracy parity difference*. We also study fairness of prediction model built on imputated datasets. Our empirical studies show that (1) Severe unfairness exists in both imputation and prediction after imputation. (2) Different imputation methods have non-negligible impact on fairness. (3) Unfairness in both imputation and prediction can be associated with the sample

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

*A detailed description of the datasets can be found in Appendix A imbalance and missing data mechanism. Our work provides valuable insight into how to effectively handle missing values while guaranteeing fairness of the learning tasks.

2. Preliminaries

2.1. Fairness notions

Discussion on fairness are based on pre-specified sensitive attributes A. In real data experiments, we use gender and ethnicity as sensitive attributes, which are widely used in social science study. For gender, we identify male as the majority group and female as the minority group. For ethnicity we identify white as the majority group and black (or "other races") as the minority group. Regarding the fairness notions, Equalized Odds (EO) (Hardt et al., 2016) is widely used. For a decision making procedure, EO aims to let algorithm's prediction $\hat{y}$ independent of sensitive attribute A, conditioned on the true outcome y. Consider the learning task in which data $\mathbf{Z}$ contains predictor X and response y. When response y is binary, false positive rate (FPR) and false negative rate (FNR) for classifier h in group a are defined as $\text{FPR}_a(h) := \sum_{A_i=a} \mathbf{1}\{h(x_i) = 1\} / \sum_{A_i=a} \mathbf{1}\{y_i = 0\}$ and $\text{FNR}_a(h) := \sum_{A_i=a} \mathbf{1}\{h(x_i) = 0\} / \sum_{A_i=a} \mathbf{1}\{y_i = 1\}$. We define *equalized odds difference* as the fairness notion in classification:

Definition 2.1 (equalized odds difference) *Equalized odds difference for classifier h is defined as $EOD(h) := |FPR_{maj}(h) - FPR_{min}(h)| + |FNR_{maj}(h) - FNR_{min}(h)|$.*

Here the subscripts "maj" and "min" mean the quantities are for majority and minority sensitive groups, respectively. The notion is a measure of how close the prediction algorithm h is to the equalized odds, in which EOD equals to 0.

In addition, in this paper we propose a novel notion that measures the fairness of imputation results. We narrow our scope to the scenario when sensitive attributes are binary $A = a \in \{0, 1\}$. Suppose the complete data matrix, without missing values, is denoted by $\mathbf{Z} = (z_{ij}) \in \mathbb{R}^{n \times p}$, and the missing indicator is denoted by $\mathbf{R} = (r_{ij}) \in \mathbb{R}^{n \times p}$, with $r_{ij} = \mathbf{1}\{z_{ij} \text{ is observed}\}$. Let $\mathbf{Z}_{\text{obs}} = \{z_{ij} | r_{ij} = 1\}$ denote the observed data and $\mathbf{Z}_{\text{miss}} = \{z_{ij} | r_{ij} = 0\}$ denote the missing data. We further let $\mathbf{Z}^a = (z^a_{ij})$ and $\mathbf{R}^a = (r^a_{ij})$ denote the complete data matrix and missing indicator matrix in sensitive group $A = a$ (so that $\mathbf{Z} = \mathbf{Z}^0 \bigcup \mathbf{Z}^1$, $\mathbf{R} = \mathbf{R}^0 \bigcup \mathbf{R}^1$). In group a, the data matrix imputed by model g is denoted by $\widehat{\mathbf{Z}}^a(g) = (\hat{z}^a_{ij}(g))$. Assume that both sensitive groups contain missing data, we define mean square imputation error of g in group a as

$$\text{MSIE}_a(g) = \frac{\sum_{(i,j)} (\hat{z}^a_{ij}(g) - z^a_{ij})^2 (1 - r^a_{ij})}{\sum_{(i,j)} 1 - r^a_{ij}}$$

Now we define *imputation accuracy parity*, a novel notion that measures fairness of imputation model g:

Definition 2.2 (imputation accuracy parity difference) *Imputation accuracy parity for imputation model g is defined as $IAPD(g) = MSIE_{maj}(g) - MSIE_{min}(g)$.*

Imputation accuracy parity is similar to the fairness notion *accuracy parity* adopted in multiple literature (Friedler et al., 2016; Zafar et al., 2017; Zhao et al., 2019a). Consider the learning tasks where data $\mathbf{Z}$ contains predictor X and binary response y. When only y contains missing values, imputation can be regarded as a prediction task. In such case, accuracy parity for imputation method g can be regarded as a finite-sample version of the accuracy parity.

2.2. Missing data mechanism

The missing data mechanism (Little & Rubin, 2019) can be classified into three types: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). A missing mechanism is said to be MCAR if missingness is independent of both observed and missing data. When missingness is only dependent on observed data, the mechanism is said to be MAR. For the MNAR, missingness can be associated with both observed and unobserved data. In this paper, we conduct real data experiments to assess fairness associated with imputation in which missing data are artificially generated under all three mechanisms. Specifically, for each real data set used, we normalize all the features and then generate missing values in the first L features in the dataset, with a pre-specified L. Throughout, we use the following 11 models for generating missing values. Given a sample $\mathbf{z} = (z_1, \cdots, z_p)$, the probability that z_j is missing is given by the following values[†], for $\forall j \in \{1, \ldots, L\}$:

MCAR		MAR		MNAR	
0.1	(1a)	$0.1 + 0.8\mathbf{1}_{\text{male}}$	(2a)	$0.5 - z_j$	(3a)
0.5	(1b)	$0.1 + 0.8\mathbf{1}_{\text{female}}$	(2b)	$0.5 - 0.2z_j$	(3b)
0.9	(1c)	$0.5 - 0.5z_{L+j}$	(2c)	$0.5 + 0.2z_j$	(3c)
		$0.5 + 0.5z_{L+j}$	(2d)	$0.5 + z_j$	(3d)

2.3. Imputation methods

The imputation methods investigated include MICE (Buuren & Groothuis-Oudshoorn, 2010), missForest (Stekhoven & Bühlmann, 2012), K-nearest neighbor (KNN) imputation, two matrix completion methods SoftImpute (Hastie et al., 2015) and OptSpace (Keshavan et al., 2010), and also two deep learning methods Gain (Yoon et al., 2018) and Misgan (Li et al., 2019). A more thorough review on existing imputation methods is provided in Appendix B.

[†]the values are truncated inside the unit interval $[0, 1]$

	MSIE								IAPD regarding gender								IAPD regarding race							
Method	I_1	I_2	I_3	I_4	I_5	I_6	I_7	Var	I_1	I_2	I_3	I_4	I_5	I_6	I_7	VarD_g	I_1	I_2	I_3	I_4	I_5	I_6	I_7	VarD_r
MCAR (1a)	0.68	0.68	**0.67**	0.81	0.88	0.73	0.99	0.99	0.45	0.46	0.47	0.50	0.48	**0.36**	0.51	0.55	-0.46	-0.48	-0.48	-0.49	-0.53	**-0.32**	-0.50	-0.53
MCAR (1b)	0.90	0.82	**0.80**	0.91	0.87	0.86	0.95	1.00	0.58	0.51	0.51	0.52	0.50	**0.42**	0.52	0.56	-0.57	-0.56	-0.55	-0.52	-0.54	**-0.42**	-0.49	-0.54
MCAR (1c)	**0.85**	1.09	0.88	0.96	0.88	0.95	0.95	1.00	0.54	0.58	0.53	0.53	0.51	**0.44**	0.47	0.56	-0.55	-0.73	-0.59	-0.53	-0.55	**-0.43**	-0.46	-0.54
MAR (2a)	**0.89**	1.08	0.85	1.03	0.90	1.00	1.04	1.09	0.53	0.79	0.57	0.56	**0.50**	0.60	**0.50**	0.55	-0.63	-0.76	-0.64	-0.61	-0.62	**-0.51**	-0.57	-0.61
MAR (2b)	**0.47**	0.53	0.49	0.62	0.57	0.70	0.69	0.72	0.43	0.35	0.41	0.49	0.49	**0.14**	0.51	0.55	-0.23	-0.27	-0.25	-0.24	-0.28	**-0.15**	-0.22	-0.27
MAR (2c)	0.42	**0.36**	0.54	0.51	0.53	0.80	0.62	0.57	0.15	0.15	0.13	0.18	0.19	**0.11**	0.17	0.21	-0.12	-0.22	**-0.04**	-0.21	-0.23	**-0.04**	-0.06	-0.25
MAR (2d)	**1.27**	1.66	1.35	1.36	1.34	1.37	1.48	1.44	0.86	0.85	0.92	0.83	0.79	**0.73**	0.84	0.83	-0.72	-0.87	-0.81	-0.64	**-0.60**	-0.65	-0.77	-0.62
MNAR (3a)	1.08	1.16	2.44	**0.26**	0.36	0.65	0.64	0.12	0.23	0.21	0.14	**0.00**	-0.01	**0.00**	0.01	0.01	-0.35	-0.47	-0.30	0.03	0.02	**0.01**	0.08	0.02
MNAR (3b)	0.67	0.53	0.67	**0.28**	0.35	0.60	0.52	0.27	0.25	0.16	0.14	0.02	**0.01**	**-0.01**	0.03	0.05	-0.27	-0.32	-0.29	**0.00**	0.03	0.02	-0.05	-0.03
MNAR (3c)	1.50	**1.42**	1.62	1.59	1.43	1.67	1.85	1.59	1.09	1.06	1.14	1.04	1.00	**0.98**	1.04	0.95	-1.01	-1.02	-1.13	-1.01	-1.01	**-0.87**	-0.97	-0.89
MNAR (3d)	2.71	2.44	2.53	2.20	**1.88**	1.95	2.11	1.71	1.59	1.47	1.56	1.35	1.28	**1.01**	1.03	1.00	-1.18	-1.32	-1.25	-1.20	-1.20	**-0.83**	-0.88	-0.90

Table 1. Imputation fairness on COMPAS recidivism dataset. Number of features (besides sensitive attributes) is 10, L = 5. Here imputation methods are encoded as: I_1: MICE; I_2: missForest; I_3: KNN; I_4: SoftImpute; I_5: OptSpace; I_6: Gain; I_7: misGAN. Results are average values over 50 repeated experiments. Var denotes the variance of missing data, VarD_g denotes the difference of missing value's variances between two gender groups and VarD_r denotes that between two race groups.

3. Fairness in imputation accuracy

In this section, we investigate the imputation fairness for different existing methods under various missing mechanisms. Each experiment is repeated for 50 times and results shown are the average values. Let VarD_g and VarD_r denote the difference of missing value's variances between two gender groups and between two racial groups, respectively. Both of the quantities serve as baselines of imputation performance, since the variance difference is equivalent to IAPD when using imputing missing values with mean observed values in two groups respectively. We set $L = 5$ in the experiment of COMPAS dataset and report the results in Table (1). Results for two other real datasets are provided in Appendix C and D.

Observation 1: Severe imputation unfairness widely exists

We observe that in all the 11 missing mechanisms, almost all the imputation methods have positive imputation IAPD regarding gender and negative IAPD regarding race. This implies that regarding gender, all the imputation methods consistently provide more accurate imputation result for female group compared with male group. Meanwhile, all the imputation methods give more accurate imputation result for white people compared with black people. Severe imputation unfairness are observed in this dataset among all the imputation models. In Appendix C and D, we also observe imputation unfairness in ADNI and ASCVD datasets.

Observation 2: Imputation fairness can be influenced by imbalance of missingness

The missingness in MAR (2a) and (2b) are approximately 0.5, which is in the similar level as MCAR (1b). However, in mechanism (2a), about 90% of first 5 features in male group are missing while only 10% of that in female group are missing. This difference can cause significant influence on imputation fairness: The more missingness one group has, the larger imputation error it appears to have. From Table 1, for many imputation methods we found that imputation IAPD regarding gender in MAR (2a) is consistently larger (taking sign into account) than that in mechanism MCAR (1b). This indicates male group's imputation error becomes relatively larger in MAR (2a). Meanwhile, all the imputation methods gives smaller imputation IAPD regarding gender in MAR (2b) than that in mechanism MCAR (1b). These empirical observation indicates that for a fixed overall missingness, imbalance of missingness between two sensitive groups can also influence imputation fairness effectively. Additional evidence is also observed in the experiments of ADNI dataset, shown in Appendix C.

Observation 3: Imputation unfairness tends to be enlarged as missingness increases

Among 3 MCAR mechanisms, from (1a) to (1c) the missingness increases (from 0.1 to 0.9). A trend of increasing imputation unfairness (IAPD regarding gender and race) is also observed for most imputation methods. This implies that degree of missingness contributes to the imputation fairness. Intuitively, the bias in imputation can be amplified by increasing missingness.

Besides, in Appendix C we have an additional observation that imputation unfairness can be associated with imbalance of sample size. Of all 649 patients in ADNI gene dataset, 642 patients are white people. We observe that the IAPD regarding race groups is almost always negative, implying imputation methods' preferences towards white people during imputation. This effect of sample imbalance is also observed in ASCVD dataset (Appendix D), where population in white group is two times that in the other group and that IAPD regarding race groups is also consistently

	Prediction Accuracy								EOD regarding gender								EOD regarding race							
Method	I_1	I_2	I_3	I_4	I_5	I_6	I_7	CC	I_1	I_2	I_3	I_4	I_5	I_6	I_7	CC_g	I_1	I_2	I_3	I_4	I_5	I_6	I_7	CC_r
MCAR (1a)	0.71	0.71	0.71	0.63	0.65	0.60	0.71	**0.72**	0.18	0.16	0.16	**0.05**	0.06	0.06	0.17	0.16	0.29	0.29	0.29	**0.03**	0.05	0.04	0.30	0.29
MCAR (1b)	0.68	0.70	0.68	0.63	0.65	0.58	0.70	**0.71**	0.17	0.16	0.14	0.05	0.06	**0.03**	0.17	0.16	0.27	0.29	0.26	**0.03**	0.04	0.05	0.31	0.29
MCAR (1c)	0.67	0.66	0.66	0.64	0.65	0.64	0.64	**0.70**	0.17	0.13	0.13	**0.06**	**0.06**	0.08	0.12	0.15	0.23	0.24	0.23	**0.04**	0.07	0.15	0.20	0.29
MAR (2a)	0.69	0.69	0.69	0.64	0.65	0.63	0.69	**0.70**	0.14	0.15	0.13	0.06	**0.05**	0.06	0.14	0.26	0.24	0.24	0.25	**0.03**	0.04	0.11	0.24	0.27
MAR (2b)	0.70	0.70	**0.71**	0.64	0.65	0.61	0.70	**0.71**	0.17	0.15	0.16	0.06	**0.05**	**0.05**	0.17	0.16	0.31	0.30	0.29	0.04	**0.03**	0.04	0.30	0.29
MAR (2c)	0.46	0.46	0.46	0.64	0.65	0.52	0.46	**0.71**	0.01	**0.00**	0.02	0.06	0.06	0.03	0.03	0.16	0.02	**0.00**	0.02	0.03	0.07	0.15	0.06	0.30
MAR (2d)	0.57	0.61	0.59	0.54	0.65	0.50	0.54	**0.71**	0.01	0.06	0.03	**0.00**	0.07	0.01	0.01	0.17	**0.00**	0.05	0.02	**0.00**	0.05	0.10	0.03	0.29
MNAR (3a)	0.68	0.66	0.62	0.64	0.66	0.61	0.64	**0.69**	0.12	0.12	0.09	0.05	0.06	0.04	**0.01**	0.14	0.32	0.24	0.16	**0.03**	0.35	**0.03**	0.08	0.25
MNAR (3b)	0.67	0.69	0.69	0.64	0.65	0.62	0.67	**0.71**	0.12	0.16	0.16	**0.05**	0.08	**0.05**	0.06	0.16	0.26	0.28	0.26	**0.03**	0.17	0.04	0.16	0.29
MNAR (3c)	0.70	0.69	0.68	0.64	0.65	0.57	0.69	**0.71**	0.17	0.13	0.12	0.06	0.06	**0.03**	0.09	0.16	0.31	0.26	0.25	0.04	**0.03**	0.05	0.18	0.28
MNAR (3d)	0.69	0.68	0.69	0.64	0.65	0.58	0.69	**0.71**	0.12	0.09	0.09	0.07	0.07	**0.03**	0.15	0.16	0.27	0.26	0.26	**0.04**	0.06	0.05	0.31	0.28

Table 2. Prediction fairness on COMPAS recidivism dataset. Number of features (besides sensitive attributes) is 10, $L = 5$. Here imputation methods are encoded as: I_1: MICE; I_2: missForest; I_3: KNN; I_4: SoftImpute; I_5: OptSpace; I_6: Gain; I_7: misGAN. CC: prediction using complete cases in the training set. Results are average values over 50 repeated experiments. When using complete data for prediction, prediction accuracy is 0.72, accuracy difference regarding gender is 0.16, accuracy difference regarding race is 0.29.

negative. This observation matches intuition that imputation accuracy is positively correlated with the amount of information observed. Imbalance of observed information leads to imputation unfairness.

4. Fairness in prediction accuracy

We study the prediction fairness for different imputation methods on three aforementioned datasets, each of which contains a response y. In each experiment, we firstly conduct a train-test split (80 % and 20 % of the sample size, respectively), artificially generate missing data in the training set, according to the 11 missing mechanisms in Section 2. Next, we impute the missing data using all the imputation models mentioned in Section 2.3. Finally, we train a prediction model h using random forest, based on the original training set (without missing values), complete cases in the training set and imputed datasets. Equalized odds difference (Definition 2.1) on the test set are reported. The results for COMPAS is shown in Table 2 and those for other two datasets are provided in Appendix C and D.

Observation 4: Prediction fairness is associated with missing mechanism

We observe that for each imputation method, i.e., a fixed column, the EOD values vary cross different rows. This implies that different missing mechanisms lead to different prediction fairness when the imputation method is fixed. In particular, in MAR (2c), all the imputation methods has smaller EOD regarding both gender and race, compared with the EODs when using complete data or complete cases to build the prediction model.

Observation 5: Imputation posts a trade-off between accuracy and fairness in prediction

From Table 2, we observe that for a fixed missing mechanism, prediction fairness associated with different imputation methods are different. In particular, we observe that Gain and two matrix completion methods: SoftImpute and OptSpace consistently have smaller EOD compared with other imputation methods (and prediction models without imputation, using complete cases and complete data).

At the same time, prediction algorithms associated with these three methods (i.e., Gain, SoftImput and OptSpace) have lower prediction accuracy. This can be viewed as a trade-off between prediction accuracy and prediction fairness. In fact, we notice from the table that such trade-off widely exists. For an arbitrary missing mechanism, the prediction accuracy associated with an imputation method is lower than that associated with complete data, and most imputation models are also associated with a smaller EOD compared with that associated with the morel built through the complete data.

In addition, in the experiment of ADNI data (shown in Table 4), the EOD for race is larger than 0.15 in most cases. A potential reason is that 624 out of 649 samples are from white patients. This suggests that prediction unfairness can also be influenced by the sample imbalance in some extreme cases, including our experiments of ADNI data.

5. Discussion

In this paper we study the fairness associated with missing data imputation in three real datasets. Our experiments show that imputation unfairness widely exists among different imputation models, representing the first known empirical results in literature. We also demonstrate factors that could contribute to imputation fairness. We further study the impact of imputation on prediction fairness when imputed data are used to build prediction models. This area offers fertile ground for theoretical investigation, as there has been littler exiting work in this area.

References

Angwin, J., Larson, J., Mattu, S., and Kirchne, L. Machine bias. ProPublica. Retrieved from https://www.propublica.org/article/ how-we-analyzed-the-compas-recidivism-algorithm, 2016.

Bakker, M. A., Valdés, H. R., Tu, D. P., Gummadi, K. P., Varshney, K. R., Weller, A., and Pentland, A. Fair enough: Improving fairness in budget-constrained decision making using confidence thresholds. In SafeAI@ AAAI, 2020.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems, pp. 4349–4357, 2016.

Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., and Zemel, R. Understanding the origins of bias in word embeddings. In International Conference on Machine Learning, pp. 803–811, 2019.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pp. 77–91, 2018.

Buuren, S. v. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. Journal of statistical software, pp. 1–68, 2010.

Chouldechova, A. and Roth, A. The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810, 2018.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 797–806, 2017.

Friedjungová, M., Vašata, D., Balatsko, M., and Jirina, M. Missing features reconstruction using a wasserstein generative adversarial imputation network.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236, 2016.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., and Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. JAMA internal medicine, 178(11):1544–1547, 2018.

Goel, N., Amayuelas, A., Deshpande, A., and Sharma, A. The importance of modeling data missingness in algorithmic fairness: A causal perspective. arXiv preprint arXiv:2012.11448, 2020.

Gondara, L. and Wang, K. Mida: Multiple imputation using denoising autoencoders. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 260–272. Springer, 2018.

Gonen, H. and Goldberg, Y. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv preprint arXiv:1903.03862, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Advances in neural information processing systems, pp. 2672–2680, 2014.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. The case for process fairness in learning: Feature selection for fair decision making. In NIPS Symposium on Machine Learning and the Law, volume 1, pp. 2, 2016.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. arXiv preprint arXiv:1610.02413, 2016.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. Matrix completion and low-rank svd via fast alternating least squares. The Journal of Machine Learning Research, 16 (1):3367–3402, 2015.

Kachuee, M., Karkkainen, K., Goldstein, O., Darabi, S., and Sarrafzadeh, M. Generative imputation and stochastic prediction. arXiv preprint arXiv:1905.09340, 2019.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from a few entries. IEEE transactions on information theory, 56(6):2980–2998, 2010.

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., and Jain, A. K. Face recognition performance: Role of demographic information. IEEE Transactions on Information Forensics and Security, 7(6):1789–1801, 2012.

Li, S. C.-X., Jiang, B., and Marlin, B. Misgan: Learning from incomplete data with generative adversarial networks. arXiv preprint arXiv:1902.09599, 2019.

Little, R. J. and Rubin, D. B. Statistical analysis with missing data, volume 793. John Wiley & Sons, 2019.

Martínez-Plumed, F., Ferri, C., Nieves, D., and Hernández-Orallo, J. Fairness and missing values. arXiv preprint arXiv:1905.12728, 2019.

Mattei, P.-A. and Frellsen, J. Miwae: Deep generative modelling and imputation of incomplete data. arXiv preprint arXiv:1812.02633, 2018.

Northpointe. Compas risk & need assessment system: Selected questions posed by inquiring agencies. 2010.

Pessach, D. and Shmueli, E. Algorithmic fairness. arXiv preprint arXiv:2001.09784, 2020.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., and Chin, M. H. Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169 (12):866–872, 2018.

Ryu, H. J., Adam, H., and Mitchell, M. Inclusivefacenet: Improving face attribute detection with race and gender diversity. arXiv preprint arXiv:1712.00193, 2017.

Stekhoven, D. J. and Bühlmann, P. Missforest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1):112–118, 2012.

Van Buuren, S. and Oudshoorn, K. Flexible multivariate imputation by MICE. Leiden: TNO, 1999.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning, pp. 1096–1103, 2008.

Wadsworth, C., Vera, F., and Piech, C. Achieving fairness through adversarial learning: an application to recidivism prediction. arXiv preprint arXiv:1807.00199, 2018.

Wang, Y. and Singh, L. Analyzing the impact of missing values and selection bias on fairness. In International Journal of Data Science and Analytics, 2021.

Yang, Y., Wu, Z., Tresp, V., and Fasching, P. A. Categorical ehr imputation with generative adversarial nets. In 2019 IEEE International Conference on Healthcare Informatics (ICHI), pp. 1–10. IEEE, 2019.

Yoon, J., Jordon, J., and Van Der Schaar, M. Gain: Missing data imputation using generative adversarial nets. arXiv preprint arXiv:1806.02920, 2018.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. 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, 2017.

Zhang, H. and Woodruff, D. P. Medical missing data imputation by stackelberg gan. Carnegie Mellon University, 2018.

Zhang, X. and Liu, M. Fairness in learning-based sequential decision algorithms: A survey. arXiv preprint arXiv:2001.04861, 2020.

Zhao, H., Coston, A., Adel, T., and Gordon, G. J. Conditional learning of fair representations. arXiv preprint arXiv:1910.07162, 2019a.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint arXiv:1804.06876, 2018.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.-W. Gender bias in contextualized word embeddings. arXiv preprint arXiv:1904.03310, 2019b.