Overcoming Data Biases: Towards Enhanced Accuracy and Reliability in Machine Learning

Jiongli Zhu, Babak Salimi
University of California San Diego
{jiz143, bsalimi}@ucsd.edu

Abstract

The pervasive integration of machine learning (ML) across various sectors has underscored the critical challenge of addressing inherent biases in ML models. These biases not only undermine the models' fairness and accuracy but also have significant real-world consequences. Traditional approaches to mitigating these biases often fail to address their root causes, leading to solutions that may superficially seem fair but do not tackle the underlying problems. This review paper explores the role of causal modeling in enhancing data cleaning, preparation, and quality management for ML. By analyzing existing research, we demonstrate how causal reasoning can effectively identify and rectify data biases, thus improving the fairness and accuracy of ML models. We advocate for the increased adoption of causal approaches in these processes, emphasizing their potential to significantly enhance the integrity and reliability of data-driven technologies.

1 Introduction

Machine Learning has become integral to sectors such as healthcare, finance, and law enforcement, spotlighting the importance of addressing biases and inaccuracies in ML models. These critical issues necessitate the development of ML models that are reliable, accurate, and fair, given their significant impact on individuals and communities. Consequently, substantial research efforts have been dedicated to mitigating algorithmic bias, aiming to enhance the robustness, reliability, accuracy, and fairness of ML models [3, 57].

Despite numerous efforts to address data biases in ML, current strategies often focus on alleviating the symptoms rather than confronting the underlying causes of these biases. This approach may inadvertently lead to "fair-washing," where superficial measures worsen the problems they intend to solve [96]. In the realm of developing fair ML models, prevalent methods include: (1) integrating fairness metrics into the optimization process during training, known as in-processing [10, 12, 42, 44, 88, 89], and adjusting the model's output post-training, referred to as post-processing [34, 41, 67, 82]; and (2) modifying the data before training, or pre-processing, to achieve a more balanced distribution [11, 26, 40, 74, 87]. However, these approaches often operate under the assumption that the training data is representative of the actual distribution [37], a premise that is frequently flawed. Data biases, such as confounding, measurement, and selection biases, along with other data quality issues, distort the data distribution [13, 27, 57, 61, 62, 96], often leading to training datasets that do not

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

accurately represent the target population. This mismatch poses challenges in preprocessing the data to obtain a representative sample. Consequently, ML models trained with such biased data are likely to underperform, being unfair and inaccurate when applied to the target population and during the inference process.

Considerable efforts have been directed toward mitigating data biases, including selection bias [17, 20, 36, 52, 70] and labeling errors [39, 90, 93]. Yet, these initiatives often hinge on significant assumptions—like the presence of an unbiased sample or specific presumptions regarding data quality issues—that are challenging to verify in real-world settings. Such reliance introduces complications, rendering these strategies less effective for practical applications. Furthermore, traditional data cleaning techniques [43, 48, 56, 69] may falter in restoring the ground truth or in ensuring datasets accurately reflect their target domains. Occasionally, it may be inherently unfeasible to obtain a representative sample to efficiently counter data biases. Overlooking these pivotal concerns may inadvertently perpetuate existing biases within the data [31, 61, 75].

In this paper, we explore data biases through a causal lens, integrating concepts from ML , causal inference, and data management. Our primary objective is to highlight the significant potential of causal reasoning in enhancing data cleaning techniques, with a particular focus on data quality management research. Causal reasoning facilitates a more thorough examination and validation of the assumptions underlying data collection processes and data provenance, thereby increasing transparency. By reviewing recent studies that employ causal inference for debiasing data, we aim to showcase the considerable impact of this methodology. Our analysis focuses on incorporating causal methods into existing frameworks for data quality management and cleaning, with the goal of reducing biases and improving both the fairness and accuracy of ML models. This specific investigation contributes to the expanding field of data quality management research, an essential component of data management. We advocate for ongoing research and development aimed at forging more robust, unbiased, and effective data-driven technologies, achieved through the refinement of data management practices.

2 Data Biases

Data bias refers to systematic errors within datasets that lead to outcomes that are either inaccurate, unfair, or unreliable. These biases often manifest as uncertainties and incompleteness in data and systematic deviations in the data distribution, compromising its representation of the actual phenomena under study. In the context of knowledge extraction, these biases can lead to analyses that yield incorrect conclusions and false discoveries. In the context of ML, if these biases are not adequately addressed, they can be learned and perpetuated by downstream models, impacting their accuracy and fairness. In this section, we explore the most common sources of data bias in real-world applications, with a specific focus on challenges such as bias due to missing data, confounding variables, and erroneous measurements [5, 28, 57, 65, 95]. Understanding and addressing these factors is vital for assessing the quality and reliability of data used in ML.

Bias due to Missing Data: During data collection, certain portions of data may be missing for various reasons, such as high data collection costs for specific sub-populations or historical discrimination [1]. This missing data can manifest as either *missing values within tuples* or *entirely missing tuples*. It is particularly challenging when entire tuples are missing non-randomly, leading to selection bias [15]. This bias occurs when the data collection process or the selection of training data is influenced by specific attributes, resulting in a subset that does not accurately represent the entire population. Even in scenarios where recovery is theoretically possible [1], such as cases of data missing completely at random or when an unbiased sample is available, the existing approaches for dealing with missing data imputation or selection bias typically only provide asymptotic guarantees [45, 72, 76]. In practical applications with finite data, these methods might display unpredictable behavior and still lead to biased samples. The non-random nature of missing data thus presents significant challenges in obtaining unbiased

[1] Recoverability of a distribution from missing data, biases, or data quality issues, in principle, refers to the capacity to accurately and consistently (asymptotically) estimate the underlying probability distribution or statistical properties of a dataset, even in the presence of such data quality challenges.

data that accurately reflects the underlying distribution, highlighting a crucial concern in data management and ML. We illustrate with two examples:

Example 1 (Missing Attribute Values) *Missing data presents a formidable challenge in critical areas such as healthcare and finance, characterized by its non-random occurrence and complex mechanisms. In pediatric health studies, for example, in cancer research, parents' hesitancy to divulge sensitive prognosis details, like life expectancy, results in crucial information being omitted. Studies have shown that such omissions correlate with poorer survival outcomes in comprehensive cancer registries [68, 71]. Similarly, financial ML applications face missing data, particularly in loan application datasets, where information on repayment potential for rejected applicants or those with restricted financial access is often absent. This gap, largely due to historical and racial biases, distorts data distribution. If not addressed, this distortion leads to inaccurate estimations and perpetuates biases in these sectors [22, 54, 68].*

Example 2 (Selection Bias) *Selection bias is prevalent in many sensitive domains, such as health care, finance, and predictive policing. In predictive policing, selection bias may occur when historical crime data, which often reflects past law enforcement and societal biases, is used to train ML models. This can lead to a cycle where certain communities are over-policed based on biased data, further perpetuating the bias in future decision makings [9, 53]. In covid-19 studies, selection bias can arise when the data is collected from a population of individuals who are hospitalized or have tested positive, leading to a false association between the test positive rate and ethnic minorities due to barriers in healthcare access [30]. In finance, selection bias can manifest in credit scoring where historical lending data may disproportionately represent certain socio-economic groups, such as individuals from higher income areas. This can lead to unfair or inaccurate credit decisions when the model is applied to populations from diverse economic backgrounds, including underdeveloped regions [4, 80].*

Bias due to Latent Confounding: Confounding bias arises in ML when unobserved confounders affect both predictors and outcomes, leading to spurious correlations and misinterpreted causal relationships [64]. This bias can distort conclusions, making data associations that seem causal when they are not. In ML, models trained on such data may base predictions on these unreliable correlations, resulting in inaccuracies and poor generalization across real-world scenarios [2, 35, 84].

Example 3 (Confounding Bias) *Confounding bias significantly impacts ML applications in healthcare and social media analytics. In healthcare, for instance, ML models trained on skin cancer images may falsely associate surgical markings with disease severity, misguiding the diagnosis [23, 81]. Similarly, pneumonia detection models may inaccurately correlate device fingerprints with the disease by using data pooled from hospitals with varying pneumonia rates, leading to misidentifications based on hospital systems rather than the disease itself [86]. In social media analytics, complex relationships between various factors and self-harm tendencies create biased associations between social media use and self-harm, complicating the analysis [79].*

Bias due to Measurement Error: Measurement errors arise when there is a discrepancy between the true value of a variable and the value obtained through measurement or observation. When these errors are not random but systematically affect certain sub-populations, this results in skewed data distribution, a situation known as measurement bias [49, 58, 65]. A prevalent form of measurement bias is label bias. Label bias arises when irrelevant factors, such as sensitive demographic information, influence the assigned labels during the data collection process.

Example 4 (Measurement Bias) *In epidemiological studies estimating cardiovascular risk from dietary habits, reliance on self-reported dietary intake questionnaires can introduce measurement bias. Participants often*

misreport consumption—understating unhealthy and overstating healthy foods due to social desirability—skewing data away from true dietary patterns. This misalignment can lead models to underestimate the benefits of healthy diets on heart disease prevention [63]. Similarly, in computer vision or natural language processing, crowdsourced data labeling can embed label bias. For example, facial recognition models may perform poorly on certain ethnic groups if labels are influenced by unconscious stereotypes, undermining the model's accuracy and fairness in applications like surveillance [33].

3 Causal Modeling of Data Biases

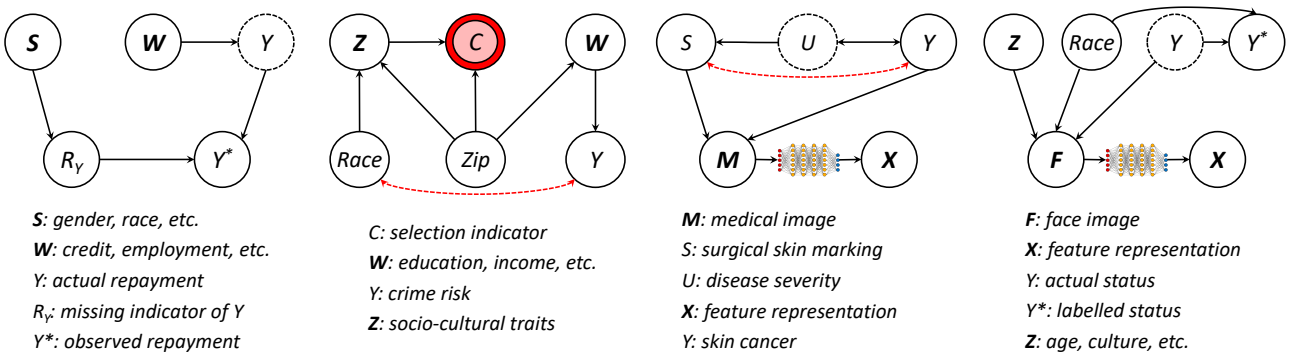
In this section, we demonstrate the essential role of causal modeling in addressing various data biases. Causal modeling provides a structured framework for understanding and capturing the provenance of data collection processes, along with their intricacies. This approach is crucial in identifying the sources of bias and plays a key role in informing the development and implementation of data debiasing and cleaning algorithms. By leveraging causal relationships, these algorithms are better equipped to tackle the root causes of bias, rather than merely addressing their symptoms. Such an approach leads to the creation of a more robust and reliable dataset, which is vital for building fair and accurate ML models.

Causal Diagrams: A causal diagram or causal graph is a directed graph that represents the causal relationships between a collection of observed or unobserved (latent) variables and models the underlying process that generates the observed data. Each node in a causal diagram corresponds to a variable, and an edge between two nodes indicates a potential causal relationship between the two variables. To illustrate, consider the causal diagram shown in Figure 1b. In this graph, the edge from the various factors such as education and income ($\boldsymbol{W}$) to the crime risk (Y) indicates that these factors of a person causally influence their risk of committing crimes.

d-separation and Conditional Independence: Causal diagrams encode a set of conditional independences that can be read off the graph using d-separation [64]. Two nodes are d-separated by a set of variables $\boldsymbol{V}_m$ in causal diagram G, denoted $(V_l \perp\!\!\!\perp V_r \mid_d \boldsymbol{V}_m)$ if for every path between them, one of the following conditions holds: (1) the path contains a chain ($V_l \to V \to V_r$) or a fork ($V_l \leftarrow V \to V_r$) such that $V \in \boldsymbol{V}_m$, and (2) the path contains a collider ($V_l \to V \leftarrow V_r$) such that $V \notin \boldsymbol{V}_m$, and no descendants of V are in $\boldsymbol{V}_m$. A distribution is said to be Markov compatible with a causal graph if d-separation within the graph implies conditional independence in the data distribution, i.e., $(V_l \perp\!\!\!\perp V_r \mid_d \boldsymbol{V}_m) \Rightarrow (V_l \perp\!\!\!\perp V_r \mid \boldsymbol{V}_m)$. Continuing with the causal diagram in Figure 1b, the graph encodes the d-separation statement $(Y \perp\!\!\!\perp \text{Zip} \mid_d \boldsymbol{W})$. For any distribution that is Markov compatible with this graph, this d-separation implies that crime risk (Y) and neighborhood (Zip) are independent, conditioned on education and income ($\boldsymbol{W}$). In this paper, assuming Markov compatibility, we consider d-separation to always imply conditional independence and use these terms interchangeably.

Next, we model each of the data biases using causal diagrams. Our discussion primarily centers on three specific types of biases: non-random missing values and selection bias as instances of bias due to missing data, confounding bias resulting from variable omission, and label bias as a manifestation of measurement errors. In addition, we explore existing research that addresses various forms of data biases in ML applications and discuss recent works that utilize the conditional independences encoded in causal diagrams for building fair ML models.

Algorithmic Fairness: Fairness in ML centers around a model h producing an output $h(\boldsymbol{x})$ and considering a protected attribute S, like gender or race. Many existing definitions of fairness require some form of statistical independence between the model's output and the protected attribute, which is sometimes conditioned on a third set of variables [57]. For instance, *statistical parity* ([21]) necessitates equal positive and negative prediction rates across different groups, formalized as $(S \perp\!\!\!\perp h(\boldsymbol{x}))$. *Equalized odds* ([34]) aims for parity in false positive and negative rates across groups, denoted as $(S \perp\!\!\!\perp h(\boldsymbol{x}) \mid Y)$. Meanwhile, *conditional statistical parity* seeks consistent positive classification probabilities across groups when accounting for certain permissible attributes $\boldsymbol{A}$, which are considered non-discriminatory factors in decision-making, expressed as $(S \perp\!\!\!\perp h(\boldsymbol{x}) \mid \boldsymbol{A})$. Notably,



(a) (b) (c) (d)

Figure 1: (a) A causal diagram for modeling missing values in pediatric health studies. (b) A causal diagram for modeling selection bias in predictive policing. (c) A causal diagram for modeling confounding bias in medical imaging. (d) A causal diagram for modeling label bias in facial image labeling. The elements of a causal diagram are: ○ representing observed attributes, ⋯ denoting unobserved attributes, ⟶ illustrating a causal dependency between two variables, ⟷ indicating a correlation due to common parent variables not included in the diagram, and ⋯⟶ signifying a spurious correlation due to data biases.

conditional statistical parity is a more general fairness concept compared to the others. When the set of admissible features $\boldsymbol{A}$ is empty, it simplifies to statistical parity, and when $\boldsymbol{A}$ includes the outcome label Y, it becomes equivalent to equalized odds. Many other associational and causal fairness criteria can also be expressed as conditional independence constraints [73].

3.1 Modeling Bias due to Missing Data

3.1.1 Missing Values

Missing values within a variable U can be efficiently represented using a binary missing indicator variable R_U, which denotes the presence or absence of data in U. Specifically, $R_U = 1$ denotes a non-missing (present) value, whereas $R_U = 0$ indicates a missing value. Furthermore, let U^* denote the observed dataset from U, where missing entries are filled with a placeholder (e.g., null). We assume that only U is subject to missing values, with other variables' actual values being completely observed. The interaction between U, R_U, and U^* can be formally depicted as follows:

$$U^* = \begin{cases} U, & \text{if } R_U = 1, \\ \text{null}, & \text{if } R_U = 0. \end{cases}$$

The subset of data that contains no missing values can be considered a sample from the distribution $\Pr(\boldsymbol{V} \mid R_U = 1)$, where $\boldsymbol{V}$ denotes the set of all variables. This implies that the subset of data without missing values is representative of the underlying distribution only if $\Pr(\boldsymbol{V} \mid R_U = 1) = \Pr(\boldsymbol{V})$. The condition for this equality is that the occurrence of missing values is completely random, denoted as $(R_U \perp\!\!\!\perp \boldsymbol{V})$, which suggests that R_U is causally independent of all other variables. However, in cases where missingness is not random, R_U is causally influenced by other variables. Such influence can be depicted in a causal graph with edges from influencing variables to R_U, thus capturing the missingness pattern. Causal modeling, therefore, provides a comprehensive framework to explicitly identify the sources of non-random missing values and to understand their effects on the data distribution. It also aids in studying the sufficient and necessary conditions for the recoverability of missing data, thereby enhancing the robustness and applicability of data analysis in various contexts.

Next, we use a concrete example in credit risk assessment to show how non-random missing values can be modeled using causal diagrams and how this would affect the downstream ML model.

Example 5 (Credit Risk Assessment) *The causal diagram in Figure 1a depicts the scenario of missing values in loan application data, as discussed in Example 1. This graph indicates that the actual loan repayment label Y is independent of demographic factors $\boldsymbol{S}$, i.e., ($\boldsymbol{S} \perp\!\!\!\perp Y$), suggesting that demographic information ($\boldsymbol{S}$) does not correlate with loan repayment (Y) in the underlying distribution ($\Pr(Y \mid \boldsymbol{S}) = \Pr(Y)$). The observed version of Y, denoted as Y^*, exhibits missingness influenced by individual demographics, reflecting that records from certain demographic groups are more prone to incompleteness due to historical biases. This relationship is captured in the causal diagram by the missingness variable R_Y, which is dependent on the demographic information $\boldsymbol{S}$. Given the high correlation between the occurrence of missing values and demographics, any imputation method with errors could lead to a biased dataset. Consequently, the imputed labels Y^*_{imp} could become strongly correlated with demographic factors $\boldsymbol{S}$. This outcome demonstrates the challenges in handling missing data, particularly when such missingness is non-randomly linked with demographic attributes. Consequently, models trained on this observed data are likely to be unfair, perpetuating historical biases.*

Data imputation methods in practice often assume that missing data occurs either completely at random (MCAR) or at random (MAR), which suggests that the mechanism of missingness does not depend on the actual values of the variable that is missing [29]. However, these methods may introduce bias when the missingness mechanism is not at random (MNAR), meaning the missingness of a variable is influenced by its own actual values or other latent variables. Such conditions render traditional imputation strategies prone to producing biased data as the original, true values of the data are typically not recoverable [29, 32, 50, 66]. Consequently, ML models trained on this biased, imputed data inherit and perpetuate the bias, leading to unfair and unreliable outcomes. To mitigate these challenges, causal modeling has been instrumental in identifying both the necessary and sufficient conditions for effectively recovering from data missingness. Additionally, it aids in pinpointing which statistics or parts of the distribution can be recovered, or in determining the external information necessary for such recovery [59]. The key to this approach lies in leveraging the invariance encoded by conditional independencies within the causal graph.

Fairness and Missing data: Recent studies investigating the impact of imputation on algorithmic fairness under different missingness mechanisms reveal significant gaps. For instance, [92] presents theoretical results on fairness guarantees in the analysis of incomplete data, while [38] highlights common disparities in imputation quality across different demographic groups. Causal modeling has been pivotal in examining the relationship between fairness and the need to consider data missingness to achieve algorithmic fairness. In this vein, recent research has harnessed the power of causal modeling to unravel multivariate dependencies in datasets with missing data, exploring the sufficient and necessary conditions for recoverability of the distribution especially when multiple variables suffer from missing data [28, 59, 60]. In particular, [28] underscores that neglecting missing data can compromise the fairness of ML models, especially in high-stakes situations like loan decision-making. The authors of this study propose a novel algorithm with a decentralized decision-making process that only leverages recoverable conditional distributions when the joint data distribution is not recoverable.

3.1.2 Selection Bias

The sampling or selection of tuples in a dataset can be modeled through a selection variable C. This binary variable indicates whether a tuple is selected, i.e., the observed data can be viewed as a random sample from the distribution $\Pr(\boldsymbol{V} \mid C = 1)$, where $\boldsymbol{V}$ represents the set of all variables. In the case of a completely random selection mechanism, where C is independent of $\boldsymbol{V}$ (i.e., $C \perp\!\!\!\perp \boldsymbol{V}$), the sampled data distribution $\Pr(\boldsymbol{V} \mid C = 1)$ is representative of the underlying distribution $\Pr(\boldsymbol{V})$. However, in the presence of selection bias, where the selection process is non-random, the selection variable C becomes dependent on other variables (i.e., $C \not\perp\!\!\!\perp \boldsymbol{V}$). This dependency is depicted in the causal graph by edges from variables that affect the selection of data to the

variable C, capturing factors influencing data selection. As a result, the sampled data becomes biased and not representative of the underlying distribution, as indicated by $\Pr(\boldsymbol{V}) \neq \Pr(\boldsymbol{V} \mid C = 1)$.

Example 6 (Predictive Policing) *Figure 1b presents a simplified causal graph that captures the data collection process in predictive policing, where ML models are applied to predict crime. The graph encodes that crime risk Y is influenced by causal factors $\boldsymbol{W}$ such as education and income, but is independent of Race. However, the graph also highlights the bias in police data, which often reflects biases from individuals' interactions with the police, influenced by socio-cultural traits and patrol frequency in their neighborhoods [9, 53]. This reflects a case of non-random data selection, where the selection variable C is influenced by both the neighborhood (Zip) and socio-cultural traits ($\boldsymbol{Z}$), as depicted in Figure 1b. As a result, the police data can be viewed as a sample from $\Pr(\boldsymbol{V} \mid C = 1)$, where $\boldsymbol{V} = \{Race, \boldsymbol{Z}, Zip, \boldsymbol{X}, Y\}$ represents the set of all variables. Due to selection bias, conditioning on C introduces a spurious correlation between race and crime ($Race \not\perp\!\!\!\perp Y \mid C = 1$) in the training data, a phenomenon known as collider bias, which is depicted by bidirectional dotted red arrows between them in the graph. Training an ML model on this biased dataset to predict crime risk is likely to learn and propagate this spurious correlation, utilizing race in predicting crime, leading to unfair and inaccurate outcomes.*

Significant efforts in ML have been directed towards mitigating selection bias, employing various techniques including causal modeling to establish when it is fundamentally possible to recover from such biases [5, 6]. Within this scope, a prominent manifestation of selection bias is termed covariate shift, which occurs when there is a discrepancy in the distribution of features $\boldsymbol{X}$ between the training and test data, while the conditional distribution $\Pr(Y \mid \boldsymbol{X})$ remains constant. This phenomenon often arises when training data suffers from selection bias where the selection mechanism is independent of the label Y. This implies that the selection variable does not directly depend on the training label Y and is d-separated from it by $\boldsymbol{X}$ in the causal diagram.

Fairness and Selection bias: Recent work in ML has focused on the interaction between algorithmic fairness and selection bias [17, 20, 36, 52, 70]. These works, including inverse propensity scoring and density ratio estimation, often rely on specific assumptions about the underlying data distribution or the need for access to unbiased samples, a requirement that can be restrictive in practical scenarios. This challenge is particularly pronounced in sensitive areas such as predictive policing, healthcare, and finance, where inherent biases in these fields make obtaining unbiased data samples impossible. However, it is often more practical to acquire background knowledge about the data collection process in these domains. Such knowledge can be effectively represented through causal diagrams. In this vein, [78] introduces a method that uses causal diagrams to mitigate model unfairness, especially under covariate shift scenarios, although this method is applicable primarily to addressable graphs that satisfy certain graphical conditions.

To overcome the limitations encountered in previous methods, a recent study CRAB [96] introduces an approach for constructing fair ML models in the presence of selection bias, without the need for an unbiased dataset. Instead of relying on stringent assumptions or unbiased samples from the underlying distribution, CRAB only requires partial knowledge about the data collection process. This approach makes it more practical compared to other methodologies that necessitate more restrictive conditions. Next, we will review CRAB as a case study to illustrate how causal reasoning can be effectively utilized to develop ML models that maintain fairness in the underlying distribution, even when faced with selection bias.

3.1.3 Consistent Range Approximation for Building Fair Models under Selection Bias

CRAB presents a framework for developing fair models under selection bias, tailored to enforce fairness definitions that can be captured by conditional independence constraints, such as conditional statistical parity, equality of odds, and predictive parity [85]. Central to this framework is fairness queries, which assess the fairness of a classifier h, which will be reviewed next.

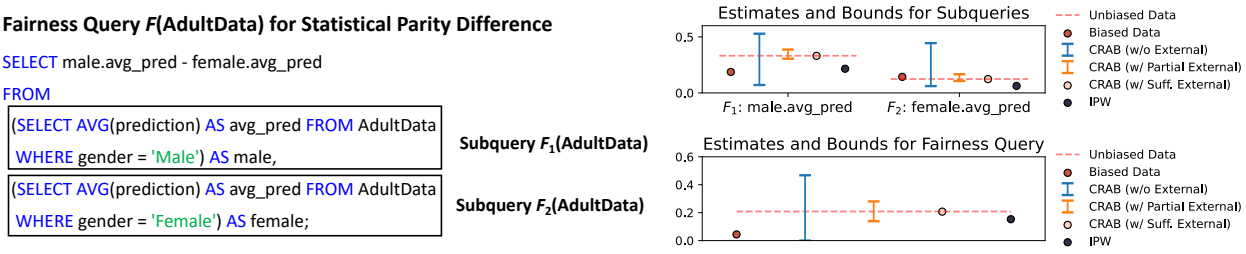


Figure 2: Comparative analysis of the consistent ranges obtained through CRA, alongside various estimates and the ground truth for the fairness query in the presence of selection bias in training data. In these plots, the red lines denote the fairness queries calculated using unbiased data. The fairness queries computed using biased data serve as a biased estimate of the unbiased fairness query. Another estimate of unbiased fairness query uses inverse propensity scores to re-weigh the data when evaluating fairness queries (IPW) [17]. The consistent ranges derived by CRAB with varying availability of external data are shown. Specifically, given sufficient external data, the consistent upper and lower bounds overlap [96].

Fairness query: Let h be a binary classifier with a protected attribute $S \in \boldsymbol{X}$, the fairness query is a measure used to assess the fairness violation of a model h wrt. the conditional statistical parity. It is defined based on a set of admissible attributes $\boldsymbol{A}$ and a population Ω with support $\boldsymbol{X} \times Y$:

$$F(\Omega) = \frac{1}{2\,|\boldsymbol{A}|} \sum_{\substack{y \in \textsc{Dom}(Y),\\ \boldsymbol{a} \in \textsc{Dom}(\boldsymbol{A})}} \left| \Pr_\Omega(h(\boldsymbol{x}) = y \mid s_1, \boldsymbol{a}) - \Pr_\Omega(h(\boldsymbol{x}) = y \mid s_0, \boldsymbol{a}) \right|.$$

It can be easily verified that for a model h to satisfy conditional statistical parity in a target population Ω, it must fulfill the condition $F(\Omega) = 0$. In this context, a fairness query is essentially the average dependency between the model's output and sensitive attributes, once adjustments have been made for admissible attributes. In practice, given a data D_Ω sampled from the distribution Ω, the fairness query $F(\Omega)$ can be computed through the empirical fairness query $\hat{F}(D_\Omega)$, which can be seen as an empirical estimate of $F(\Omega)$. Specifically, in the context of binary classification, $\hat{F}(D_\Omega)$ can be calculated by:

$$\hat{F}(D_\Omega) = \frac{1}{|\boldsymbol{A}|} \sum_{\boldsymbol{a} \in \textsc{Dom}(\boldsymbol{A})} \left| \frac{\sum_{\boldsymbol{x} \in \mathsf{N}^+_{s_1,\boldsymbol{a}}} h(\boldsymbol{x})}{|\mathsf{N}^+_{s_1,\boldsymbol{a}}|} - \frac{\sum_{\boldsymbol{x} \in \mathsf{N}^+_{s_0,\boldsymbol{a}}} h(\boldsymbol{x})}{|\mathsf{N}^+_{s_0,\boldsymbol{a}}|} \right|$$

where $\mathsf{N}^+_{s,\boldsymbol{a}}$ denotes the set of data points in D_Ω with positive labels, protected attribute values $S = s$ and admissible attributes value $\boldsymbol{A} = \boldsymbol{a}$. For example, Figure 2 presents an empirical fairness query that measures the model's violation of statistical parity on the adult data. In order to avoid sampling variability, in the subsequent, we assume samples are sufficiently large such that $\hat{F}(D_\Omega) \approx F(\Omega)$ and use them interchangeably.

Building models that are fair on the target population Ω requires to achieve $F(\Omega) = 0$. However, in practice, we only have access to the biased data D_Δ sampled from the population Δ that suffers from selection bias. Using this biased data D_Δ to evaluate fairness query gives $\hat{F}(D_\Delta)$, which is a biased and inaccurate estimate of the actual unfairness $F(\Omega)$. Furthermore, mitigating unfairness based on this biased estimate will result in a model that is fair on the biased training data ($F(\Delta) = 0$), while being unfair when deployed to the unbiased target population ($F(\Omega) \neq 0$). Nevertheless, without external data about the unbiased target population Ω, it's almost impossible to accurately estimate $F(\Omega)$.

In addressing the challenge of answering fairness queries from data affected by selection bias, the situation is akin to query answering on incomplete datasets, where complete and accurate responses are unattainable due to missing information. This challenge is tackled using an approach inspired by the concepts of possible worlds and consistent query answering [16, 24, 25, 47]. CRAB utilizes this methodology by considering every conceivable

underlying population or "possible world" from which the training data could have been obtained. Conceptually, CRAB computes the fairness query in each possible world and then uses these computations to establish a range for the fairness query by determining upper and lower bounds for unbiased fairness query answers. To generate tight and meaningful ranges, CRAB incorporates auxiliary information, which helps in narrowing down the range of potential underlying distributions. This approach is crucial for accurately evaluating fairness in models where the data is compromised by selection bias, ensuring a more reliable and valid assessment of fairness.

Auxiliary information and possible repairs: The auxiliary information that CRAB incorporates includes the causal diagram that represents the collection process of the biased data, and a set of external data sources that can potentially provide partial information about the underlying distribution Ω. Intuitively, the causal diagram encodes the causes of selection bias, i.e., which variables affect the tuple selection, while the external data source can be used to compute unbiased statistics about the underlying distribution. CRAB captures the space of possible unbiased, complete data using the notion of possible repairs. Formally, given a biased dataset D_Δ, the set of possible repairs of D_Δ, denoted as $\mathsf{Repairs}(D_\Delta)$, is defined as the set of all datasets D with the same schema as D_Δ such that: (1) $D \supsetneq D_\Delta$ and (2) D is consistent with $\mathcal{G}$ and $\mathcal{A}_\Omega$, i.e., it satisfies the constraints posed by the auxiliary information. Specifically, all repairs in $\mathsf{Repairs}(D_\Delta)$ must adhere to the conditional independences encoded in the causal diagram, and the unbiased statistics derived from $\mathcal{A}_\Omega$. Note that CRAB does not compute each of the possible repairs. Instead, the concept of possible repairs is used as a framework for addressing the incompleteness of information in the presence of selection bias. The problem of consistent range approximation is built upon the concept of possible repairs.

Consistent range approximation: The consistent range approximation (CRA) computes the consistent upper bound (CUB) and consistent lower bound (CLB) of the fairness query $F(\Omega)$. Similar to consistent query answering in databases [8, 19], CRA considers the space of all possible repairs, which stands for possible ways to complete the biased data D_Δ. Specifically,

$$\text{CLB} = \min_{D \in \mathsf{Repairs}(D_\Delta)} \hat{F}(D), \quad \text{CUB} = \max_{D \in \mathsf{Repairs}(D_\Delta)} \hat{F}(D)$$

As mentioned, CRA does not compute each of the possible repairs, but utilizes the conditional independence conditions encoded in the causal diagram, which every possible repair must satisfy, to derive closed-form solutions for the range of fairness query answers. This ensures that the actual unfairness of the model on the underlying distribution will fall within this computed range, i.e. $F(\Omega) \in [\text{CLB}, \text{CUB}]$. This range is referred to as the consistent range. Furthermore, CRA can integrate varying levels of external data sources about the underlying distribution, enabling the derivation of more precise consistent ranges. This property makes CRAB a practical solution for addressing selection bias.

It is worth noting that the external data source is not mandatory for CRA. In the absence of external data sources, [96] provides the closed-form CUB and CLB leveraging merely the conditional independence condition encoded in the causal diagram. We use the example of police data to demonstrate CRA in the absence of external data sources. For simplicity, we illustrate the CRA of fairness query wrt. statistical parity, where $\boldsymbol{A} = \emptyset$.

Example 7 (CRA on the Predictive Policing Data) *Continuing with Example 6, assume the protected attribute $Race \in \{white, non\text{-}white\}$ and the label, crime risk $Y \in \{low\ risk, high\ risk\}$. In this case, the fairness query wrt. statistical parity notion can computed by:*

$$F(\Omega) = \Pr_\Omega(low\ risk \mid white) - \Pr_\Omega(low\ risk \mid non\text{-}white). \tag{3}$$

The CUB of $F(\Omega)$ can be derived by combining the upper bound of $\Pr_\Omega(low\ risk \mid white)$ and the lower bound of $\Pr_\Omega(low\ risk \mid non\text{-}white)$. First, we show how $\Pr_\Omega(low\ risk \mid white)$ is upper bounded. As presented in Figure 1b, the selection variable C is influenced by ZIP and $\boldsymbol{Z}$. Let $\boldsymbol{U} = (C) = \{\boldsymbol{Z}, ZIP\}$, we have the conditional independence condition encoded in the causal diagram: $(C \perp\!\!\!\perp \boldsymbol{V} \mid \boldsymbol{U})$, where $\boldsymbol{V}$ is the set of all variables. The following holds due to this conditional independence:

$$\Pr_\Omega(low\ risk \mid white, \boldsymbol{u}) = \Pr_\Omega(low\ risk \mid white, \boldsymbol{u}, C = 1) = \Pr_\Delta(low\ risk \mid white, \boldsymbol{u}). \tag{4}$$

The upper bound of $\Pr_\Omega(low\ risk \mid white)$ *can be derived by applying the law of total probability and Eq. 4:*

$$\begin{aligned}
\Pr_\Omega(low\ risk \mid white) &= \sum_{\boldsymbol{u} \in \text{DOM}(\boldsymbol{U})} \Pr_\Omega(low\ risk \mid white, \boldsymbol{u}) \Pr_\Omega(\boldsymbol{u} \mid white) \\
&= \sum_{\boldsymbol{u} \in \text{DOM}(\boldsymbol{U})} \Pr_\Delta(low\ risk \mid white, \boldsymbol{u}) \Pr_\Omega(\boldsymbol{u} \mid white) \\
&\leq \sum_{\boldsymbol{u} \in \text{DOM}(\boldsymbol{U})} (\max_{\boldsymbol{u}^* \in \text{DOM}(\boldsymbol{U})} \Pr_\Delta(low\ risk \mid white, \boldsymbol{u}^*))\ \Pr_\Omega(\boldsymbol{u} \mid white) \\
&= \max_{\boldsymbol{u}^* \in \text{DOM}(\boldsymbol{U})} \Pr_\Delta(low\ risk \mid white, \boldsymbol{u}^*) \sum_{\boldsymbol{u} \in \text{DOM}(\boldsymbol{U})} \Pr_\Omega(\boldsymbol{u} \mid white) \\
&= \max_{\boldsymbol{u}^* \in \text{DOM}(\boldsymbol{U})} \Pr_\Delta(low\ risk \mid white, \boldsymbol{u}^*).
\end{aligned} \tag{5}$$

Similarly, one can derive a lower bound for $\Pr_\Omega(low\ risk \mid non\text{-}white)$*, resulting in the subsequent formulation for the CUB of the fairness query:*

$$\mathit{F}(\Omega) \leq \mathit{CUB} = \max_{\boldsymbol{u}^* \in \text{DOM}(\boldsymbol{U})} \Pr_\Delta(low\ risk \mid white, \boldsymbol{u}^*) - \min_{\boldsymbol{u}^* \in \text{DOM}(\boldsymbol{U})} \Pr_\Delta(low\ risk \mid non\text{-}white, \boldsymbol{u}^*).$$

Furthermore, if sufficient external data sources which enable computing the unbiased statistics $\Pr_\Omega(\boldsymbol{u} \mid white)$ *are available, CRA is able to directly estimate* $\mathit{F}(\Omega)$*.*

The above results demonstrate how CRA gives consistent ranges with no or sufficient external data. In practice, one may have access to a level of external data that falls in between these two extremes. For instance, we might not have access to the external data about socio-cultrual traits $\boldsymbol{Z}$, thus only being able to compute the unbiased probabilities $\Pr_\Omega(ZIP \mid Race)$. CRAB also provides closed-form consistent ranges when having partial access to external data, including this case. Next, we empirically compare the various estimates of the fairness query with the CLBs and CUBs obtained through CRA on real-world data. We focus on the CLBs and CUBs computed when having no or sufficient external data, as they have been introduced in Example 7.

Example 8: Figure 2 presents the comparison between consistent ranges and the estimates of the model's unfairness on the unbiased distribution. The adult data [51], which contains financial and demographic data to predict if an individual's income exceeds $50K$, is used for model training and testing. Specifically, the training data is injected with selection bias, where the selection depends on gender, age, and relationship. In the example, the consistent range of the fairness query can be computed based on the consistent ranges of its sub-queries. When unbiased external data is unavailable, the fairness query computed using biased data shows significant inaccuracy, especially for subquery F_2. Nevertheless, in the absence of unbiased external data, CRAB guarantees to upper and lower bound the actual query answer on the underlying distribution. When the unbiased external data is leveraged, IPW still deviates from the unbiased fairness query. In contrast, given sufficient external data (a subset of unlabeled data used by IPW), the consistent upper and lower bounds derived by CRAB overlaps, resulting in an accurate estimate of the unbiased fairness query. In addition, the consistent ranges obtained with partial external data demonstrate the effectiveness of incorporating limited unbiased external data for deriving tighter consistent ranges. The results imply that (1) the consistent ranges always guarantee to bound the actual unfairness of the ML model, and (2) given external data about unbiased distribution, CRAB is able to derive tighter bounds or estimates of the unbiased fairness query.

The CUBs of fairness queries can be seen as the models' worst-case unfairness given available information about the underlying distribution. Therefore, CUBs can be used to train certifiably fair ML models by incorporating them into the loss function. In addition to the CRAB system, [96] also presents a theoretical analysis of the impact of selection bias on the fairness of ML models and establishes necessary and sufficient graphical conditions on the data collection causal diagram under which the selection bias leads to unfair ML models.

3.2 Confounding Bias

Confounding bias presents challenges in ML when a latent variable C confounds some observed features S with the training label Y, distorting their association. For example, in healthcare data, suppose S represents lifestyle factors or genetic predispositions, Y is the disease training label, and C encompasses unrecorded environmental factors like exposure to pollutants or access to healthcare facilities. Reliance on S for predicting Y can render ML models unreliable due to unstable correlations across different settings [91]. Furthermore, when S includes sensitive attributes, confounding bias can introduce biases that unfairly impact certain groups, especially if C relates to socioeconomic factors such as income level or education, thereby exacerbating disparities.

Example 9 (Medical Imaging) *Continuing with the application of skin cancer detection in Example 3. The causal modeling of confounding bias is shown in Figure 1c. In the causal diagram, the presence of a surgical skin marking (S) does not causally contribute to skin cancer (Y) as there are no edges between them. However, they become correlated in the data due to the confounding of disease severity (U).*

Since ML models learn correlation instead of causation, this non-causal spurious correlation between the presence of surgical skin markings and skin cancer will be learned and lead to inaccurate predictions. In particular, the model will have a high false positive rate on patients with other severe diseases, who are also likely to have surgical skin markings.

Fairness, Robustness, and Confounding Bias: Confounding bias poses a significant challenge across the board, particularly impacting the robustness and fairness of algorithmic models. The crux of efforts in algorithmic fairness is to ensure that sensitive attributes and training labels remain independent, conditioned on a subset of observed features, thus aiming to nullify spurious correlations brought about by unobserved confounding biases [26, 55, 74]. Achieving such independence ($S \perp\!\!\!\perp Y \mid X$), as exemplified in Example 9, is vital for preventing reliance on non-causal features like surgical markings for predictions, which enhances both the fairness and robustness of models. A variety of approaches have been developed to enforce conditional independence, ranging from feature selection methods that mitigate spurious correlations [26], counterfactual data augmentation techniques that elucidate causal relationships and generate varied counterfactual scenarios [55], to minimal repair strategies such as *Capuchin* for data adjustment in compliance with Multivalued Dependency (MVD) [74]. Furthermore, in-processing techniques play a crucial role, incorporating strategies such as integrating conditional mutual information into the loss function [77], employing adversarial mechanisms for confounding-invariant feature extraction [94], and developing feature representations that achieve conditional independence [83]. Ultimately, causal inference stands as a foundational strategy for modeling confounding bias and securing the requisite conditional independence, thus bolstering the efforts to enhance fairness and ensure robustness against confounding bias and spurious correlations in algorithmic models.

3.3 Measurement Bias

Given a variable U affected by the measurement error, we can create a variable U^* indicating the collected or observed values, while the actual variable U is unobserved. When measurement errors are non-random, the values of the observed variable U^* often depend on its actual value U and other variables. The observed data suffering from this measurement bias can be seen as a random sample from $\Pr(\boldsymbol{V} \setminus \{U\}, U^*)$ where $\boldsymbol{V}$ denotes the set of all variables. It is only representative of the underlying distribution when $\Pr(\boldsymbol{V}) = \Pr(\boldsymbol{V} \setminus \{U\}, U^*)$, which rarely holds in practice. In the context of ML, the label variable Y often suffers from mismeasurement and appears to be biased, which degrades the performance of downstream ML models [39]. Next, we will discuss an example of label bias existing in the medical imaging data.

Example 10 (Modeling Label Bias) *Continuing with the scenario of facial identification in Example 10. Figure 1d presents the causal modeling of label bias in this application. Ideally, the actual label Y and the sensitive attribute $Race$ are independent ($Y \perp\!\!\!\perp Race$). However, due to the inadvertent bias during the labeling, the observed label Y^* is influenced by both $Race$ and Y, resulting in the correlation between race-related facial features and labels in the observed data ($Y^* \not\perp\!\!\!\perp Race$). Consequently, models trained on this biased observed data will predict based on race-related facial features, leading to inaccuracies and unfairness.*

The general problem of measurement bias has been studied recently, particularly in the context of causal inference. Through structural equation modeling, [46] detects measurement bias in longitudinal health-related data. In contrast, [7] applies Bayesian factor analysis to effectively detect both uniform and non-uniform measurement bias, with high detection rates in cases where an observed violator is present. To eliminate the systematic bias induced by measurement errors, [65] highlights several algebraic and graphical methods that work under different assumptions about the error mechanism. Beyond the broader issue of measurement bias, a range of research specifically targets the challenge of unfairness stemming from label bias.

Fairness and Label Bias: Addressing label bias in ML necessitates innovative optimization and modeling strategies. [39] tackles this by positing that the distribution of biased labels should closely match the true distribution in terms of KL divergence, subject to the observed level of unfairness. They approach this through a constrained optimization problem, adjusting data weights to mitigate label bias while aiming for minimal alteration. [90] approaches fairness through a label-flipping optimization problem, designed to adjust labels for individual fairness with minimal changes, formulated as a mixed-integer quadratic programming problem. This is further refined to an integer linear programming challenge, with [90] providing approximate yet theoretically grounded solutions. On another front, [93] focuses on identifying label inaccuracies by associating low self-confidence in model predictions with potential errors, utilizing confidence intervals for selective data refinement. These methods, while effective, often rely on simplifying assumptions, such as a minimal number of mislabeled instances, and do not fully confront measurement bias directly. However, advancements in causal modeling offer a principled approach to constructing fair and accurate models by accounting for measurement bias. [14] leverage the concept of conditional independence between unbiased labels and other variables, informed by facial action units, to tailor loss functions that enhance fairness in facial expression recognition. Similarly, [18] explores various strategies for remedying label bias, emphasizing the crucial role of accurate causal diagrams in developing unbiased algorithmic risk assessments without compromising fairness.

4 Conclusions and Future Directions

This paper has investigated the significant challenges posed by data biases in machine learning (ML), emphasizing the critical role of causal modeling in addressing these complexities. By analyzing data biases resulting from missing data, confounding variables, and measurement errors, we have highlighted their substantial impact on the fairness, accuracy, and reliability of ML models. Adopting a causal perspective not only helps in mitigating the symptoms of data biases but also in directly tackling their root causes. This approach is key to developing more robust and equitable ML applications, illustrating the importance of understanding data generation processes to effectively minimize algorithmic bias.

Our exploration underscores the need for ongoing research and improvement in data-centric methods to enhance fairness, robustness, and accuracy in ML. We advocate for better data management practices, emphasizing their vital role in advancing ML and ensuring its benefits to society. Future research directions are poised for significant advances through the integration of data bias considerations with various aspects of data quality management in databases, particularly in terms of information incompleteness and inconsistency. Data biases inherently lead to these issues, suggesting that insights from data management research could significantly contribute to developing new approaches for data cleaning and quality management in ML. This includes devising strategies for training ML models in the presence of incomplete and uncertain data.

Moreover, effectively addressing data biases involves focusing on various constraints that capture the statistical properties of data, similar to integrity constraints in data management. Conditional independence constraints, for example, are a critical category of statistical integrity constraints vital for learning de-confounded predictive models, eliminating spurious correlations, and ensuring fairness in predictive modeling. The pursuit of research in developing data cleaning methods with respect to conditional independence constraints, investigating the interplay between these constraints and database dependencies, and formulating efficient maintenance, validation, and repair techniques is imperative. Such initiatives are poised to significantly enhance data fairness and model reliability in ML, paving the way for more accountable and transparent AI systems.

References

[1] Karen Antman, David Amato, William Wood, J Carson, Herman Suit, Karl Proppe, Robert Carey, J Greenberger, R Wilson, and E Frei 3rd. Selection bias in clinical trials. Journal of Clinical Oncology, 3(8):1142–1147, 1985.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.

[3] Agathe Balayn, Christoph Lofi, and Geert-Jan Houben. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. The VLDB Journal, 30(5):739–768, 2021.

[4] John Banasik, Jonathan Crook, and Lyn Thomas. Sample selection bias in credit scoring models. Journal of the Operational Research Society, 54(8):822–832, 2003.

[5] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In Artificial Intelligence and Statistics, pages 100–108. PMLR, 2012.

[6] Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[7] M. Barendse, C. Albers, F. Oort, and M. Timmerman. Measurement bias detection through bayesian factor analysis. Frontiers in Psychology, 5, 2014.

[8] Leopoldo Bertossi. Consistent query answering in databases. ACM Sigmod Record, 35(2):68–76, 2006.

[9] Sarah Brayne, Alex Rosenblat, and Danah Boyd. Predictive policing. Data & Civil Rights: A New Era Of Policing And Justice, pages 2015–1027, 2015.

[10] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 21(2):277–292, 2010.

[11] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. Advances in neural information processing systems, 30, 2017.

[12] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In Proceedings of the conference on fairness, accountability, and transparency, pages 319–328, 2019.

[13] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: Why? how? what to do? arXiv preprint arXiv:2105.12195, 2021.

[14] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 14980–14991, 2021.

[15] SC Choi and IL Lu. Effect of non-random missing data mechanisms in clinical trials. Statistics in medicine, 14(24):2675–2684, 1995.

[16] Marco Console, Paolo Guagliardo, Leonid Libkin, and Etienne Toussaint. Coping with incomplete data: Recent advances. In Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, pages 33–47, 2020.

[17] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In International conference on algorithmic learning theory, pages 38–53. Springer, 2008.

[18] Jessica Dai and Sarah M Brown. Label bias, label shift: Fair machine learning with unreliable labels. In NeurIPS 2020 Workshop on Consequential Decision Making in Dynamic Environments, volume 12, 2020.

[19] Akhil A Dixit and Phokion G Kolaitis. Consistent answers of aggregation queries using sat solvers. arXiv preprint arXiv:2103.03314, 2021.

[20] Wei Du and Xintao Wu. Robust fairness-aware learning under sample selection bias. arXiv preprint arXiv:2105.11570, 2021.

[21] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In ITCS, pages 214–226. ACM, 2012.

[22] Adrien Ehrhardt, Christophe Biernacki, Vincent Vandewalle, Philippe Heinrich, and Sébastien Beben. Reject inference methods in credit scoring. Journal of Applied Statistics, 48(13-15):2734–2754, 2021.

[23] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. nature, 542(7639):115–118, 2017.

[24] Wenfei Fan and Floris Geerts. Foundations of data quality management. Synthesis Lectures on Data Management, 4(5):1–217, 2012.

[25] Su Feng, Boris Glavic, Aaron Huber, and Oliver A Kennedy. Efficient uncertainty tracking for complex queries with attribute-level bounds. In Proceedings of the 2021 International Conference on Management of Data, pages 528–540, 2021.

[26] Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R Varshney. Causal feature selection for algorithmic fairness. 2022.

[27] Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. JAMA internal medicine, 178(11):1544–1547, 2018.

[28] Naman Goel, Alfonso Amayuelas, Amit Deshpande, and Amit Sharma. The importance of modeling data missingness in algorithmic fairness: A causal perspective. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 7564–7573, 2021.

[29] John Graham. Missing data: Analysis and design. New York, NY: Springer. 06 2012.

[30] Gareth J Griffith, Tim T Morris, Matthew J Tudball, Annie Herbert, Giulia Mancano, Lindsey Pike, Gemma C Sharp, Jonathan Sterne, Tom M Palmer, George Davey Smith, et al. Collider bias undermines our understanding of covid-19 disease risk and severity. Nature communications, 11(1):1–12, 2020.

[31] Shubha Guha, Falaah Arif Khan, Julia Stoyanovich, and Sebastian Schelter. Automated data cleaning can hurt fairness in machine learning-based decision making. ICDE, 2022.

[32] Anna Guo, Jiwei Zhao, and Razieh Nabi. Sufficient identification conditions and semiparametric estimation under missing not at random mechanisms. arXiv preprint arXiv:2306.06443, 2023.

[33] Luke Haliburton, Sinksar Ghebremedhin, Robin Welsch, Albrecht Schmidt, and Sven Mayer. Investigating labeler bias in face annotation for machine learning. arXiv preprint arXiv:2301.09902, 2023.

[34] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In NIPS, pages 3315–3323, 2016.

[35] Ziniu Hu, Zhe Zhao, Xinyang Yi, Tiansheng Yao, Lichan Hong, Yizhou Sun, and Ed Chi. Improving multi-task generalization via regularizing spurious correlation. Advances in Neural Information Processing Systems, 35:11450–11466, 2022.

[36] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. Advances in neural information processing systems, 19, 2006.

[37] Maliha Tashfia Islam, Anna Fariha, Alexandra Meliou, and Babak Salimi. Through the data management lens: Experimental analysis and evaluation of fair classification. In Proceedings of the 2022 International Conference on Management of Data, pages 232–246, 2022.

[38] Vincent Jeanselme, Maria De-Arteaga, Zhe Zhang, Jessica Barrett, and Brian Tom. Imputation strategies under clinical presence: Impact on algorithmic fairness. In Machine Learning for Health, pages 12–34. PMLR, 2022.

[39] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In International Conference on Artificial Intelligence and Statistics, pages 702–712. PMLR, 2020.

[40] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. Knowledge and information systems, 33(1):1–33, 2012.

[41] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In 2012 IEEE 12th International Conference on Data Mining, pages 924–929. IEEE, 2012.

[42] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 35–50. Springer, 2012.

[43] Bojan Karlas, Peng Li, Renzhi Wu, Nezihe Merve Gürel, Xu Chu, Wentao Wu, and Ce Zhang. Nearest neighbor classifiers over incomplete information: From certain answers to certain predictions. Proc. VLDB Endow., 14(3):255–267, 2020.

[44] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Proceedings of the 35th International Conference on Machine Learning, pages 2564–2572. PMLR, 2018.

[45] Jae Kwang Kim. Finite sample properties of multiple imputation estimators. 2004.

[46] B. King-Kallimanis, F. Oort, and G. Garst. Using structural equation modelling to detect measurement bias and response shift in longitudinal data. AStA Advances in Statistical Analysis, 94:139–156, 2010.

[47] Paraschos Koutris and Jef Wijsen. Consistent query answering for primary keys and conjunctive queries with negated atoms. In Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, pages 209–224, 2018.

[48] Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J Franklin, and Ken Goldberg. Activeclean: Interactive data cleaning for statistical modeling. Proceedings of the VLDB Endowment, 9(12):948–959, 2016.

[49] Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. Biometrika, 101(2):423–437, 2014.

[50] Katherine J Lee and John B Carlin. Recovery of information from multiple imputation: a simulation study. Emerging themes in epidemiology, 9:1–10, 2012.

[51] M. Lichman. Uci machine learning repository, 2013.

[52] Anqi Liu and Brian Ziebart. Robust classification under sample selection bias. Advances in neural information processing systems, 27, 2014.

[53] Kristian Lum and William Isaac. To predict and serve? Significance, 13(5):14–19, 2016.

[54] Qingwei Luo, Sam Egger, Xue Qin Yu, David P Smith, and Dianne L O'Connell. Validity of using multiple imputation for" unknown" stage at diagnosis in population-based cancer registry data. PLoS One, 12(6):e0180033, 2017.

[55] Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. Learning for counterfactual fairness from observational data. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 1620–1630, 2023.

[56] Mohammad Mahdavi and Ziawasch Abedjan. Baran: Effective error correction via a unified context representation and transfer learning. Proceedings of the VLDB Endowment, 13(12):1948–1961, 2020.

[57] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6):1–35, 2021.

[58] Roger E Millsap and Howard T Everson. Methodology review: Statistical approaches for assessing measurement bias. Applied psychological measurement, 17(4):297–334, 1993.

[59] Karthika Mohan and Judea Pearl. Graphical models for processing missing data. Journal of the American Statistical Association, 116(534):1023–1037, 2021.

[60] Razieh Nabi, Rohit Bhattacharya, Ilya Shpitser, and James Robins. Causal and counterfactual views of missing data models. arXiv preprint arXiv:2210.05558, 2022.

[61] Felix Neutatz, Binger Chen, Ziawasch Abedjan, and Eugene Wu. From cleaning before ml to cleaning for ml. Data Engineering, page 24, 2021.

[62] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data, 2:13, 2019.

[63] L. Page and M. Henderson. Appraising the evidence: what is measurement bias? Evidence Based Mental Health, 11:36 – 37, 2008.

[64] Judea Pearl. Causality. Cambridge university press, 2009.

[65] Judea Pearl. On measurement bias in causal inference. arXiv preprint arXiv:1203.3504, 2012.

[66] Judea Pearl and Karthika Mohan. Recoverability and testability of missing data: Introduction and summary of results. Available at SSRN 2343873, 2013.

[67] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. Advances in neural information processing systems, 30, 2017.

[68] Jennifer K Plichta, Christel N Rushing, Holly C Lewis, Marguerite M Rooney, Dan G Blazer, Samantha M Thomas, E Shelley Hwang, and Rachel A Greenup. Implications of missing data on reported breast cancer mortality. Breast Cancer Research and Treatment, 197(1):177–187, 2023.

[69] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. Holoclean: Holistic data repairs with probabilistic inference. Proc. VLDB Endow., 10(11):1190–1201, 2017.

[70] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian Ziebart. Robust fairness under covariate shift. arXiv preprint arXiv:2010.05166, 2020.

[71] A. Rosenberg, V. Dussel, L. Orellana, T. Kang, J. Geyer, C. Feudtner, and J. Wolfe. What's missing in missing data? omissions in survey responses among parents of children with advanced cancer. Journal of palliative medicine, 17 8:953–6, 2014.

[72] Donald B Rubin. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In Proceedings of the survey research methods section of the American Statistical Association, volume 1, pages 20–34. American Statistical Association Alexandria, VA, USA, 1978.

[73] Babak Salimi, Bill Howe, and Dan Suciu. Database repair meets algorithmic fairness. ACM SIGMOD Record, 49(1):34–41, 2020.

[74] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019., pages 793–810, 2019.

[75] Sebastian Schelter, Tammo Rukat, and Felix Biessmann. Jenga-a framework to study the impact of data errors on the predictions of machine learning models. In EDBT, pages 529–534, 2021.

[76] Nathaniel Schenker and Alan H Welsh. Asymptotic results for multiple imputation. The Annals of Statistics, 16(4):1550–1566, 1988.

[77] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Information-theoretic bias reduction via causal view of spurious correlation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 2180–2188, 2022.

[78] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 3–13, 2021.

[79] Anita Johanna Tørmoen, Martin Øverlien Myhre, Anine Therese Kildahl, Fredrik Andreas Walby, and Ingeborg Rossow. A nationwide study on time spent on social media and self-harm among adolescents. Scientific reports, 13(1):19111, 2023.

[80] Geert Verstraeten and Dirk Van den Poel. The impact of sample bias on consumer credit scoring performance and profitability. Journal of the operational research society, 56:981–992, 2005.

[81] Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA dermatology, 155(10):1135–1141, 2019.

[82] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In Proceedings of the 2017 Conference on Learning Theory, pages 1920–1953, 2017.

[83] Renzhe Xu, Peng Cui, Kun Kuang, Bo Li, Linjun Zhou, Zheyan Shen, and Wei Cui. Algorithmic decision making with conditional fairness. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pages 2125–2135, 2020.

[84] Tal Yarkoni. The generalizability crisis. Behavioral and Brain Sciences, 45:e1, 2022.

[85] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[86] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine, 15(11):e1002683, 2018.

[87] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In ICML (3), volume 28 of JMLR Workshop and Conference Proceedings, pages 325–333. JMLR.org, 2013.

[88] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pages 335–340, 2018.

[89] Hantian Zhang, Xu Chu, Abolfazl Asudeh, and Shamkant B Navathe. Omnifair: A declarative system for model-agnostic group fairness in machine learning. In Proceedings of the 2021 international conference on management of data, pages 2076–2088, 2021.

[90] Hantian Zhang, Ki Hyun Tae, Jaeyoung Park, Xu Chu, and Steven Euijong Whang. iflipper: Label flipping for individual fairness. Proceedings of the ACM on Management of Data, 1(1):1–26, 2023.

[91] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5372–5382, 2021.

[92] Yiliang Zhang and Qi Long. Assessing fairness in the presence of missing data. Advances in Neural Information Processing Systems, 34, 2021.

[93] Yixuan Zhang, Boyu Li, Zenan Ling, and Feng Zhou. Mitigating label bias in machine learning: Fairness through confident learning. arXiv preprint arXiv:2312.08749, 2023.

[94] Qingyu Zhao, Ehsan Adeli, and Kilian M Pohl. Training confounder-free deep learning models for medical applications. Nature communications, 11(1):1–9, 2020.

[95] Sami Zhioua and Rūta Binkytė. Dissecting causal biases. 2023.

[96] Jiongli Zhu, Sainyam Galhotra, Nazanin Sabri, and Babak Salimi. Consistent range approximation for fair predictive modeling. Proceedings of the VLDB Endowment, 16(11):2925–2938, 2023.