Towards Assessing Data Bias in Clinical Trials

Chiara Criscuolo^(✉), Tommaso Dolci, and Mattia Salnitri

Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milan, Italy
{chiara.criscuolo,tommaso.dolci,mattia.salnitri}@polimi.it

Abstract. Algorithms and technologies are essential tools that pervade all aspects of our daily lives. In the last decades, health care research benefited from new computer-based recruiting methods, the use of federated architectures for data storage, the introduction of innovative analyses of datasets, and so on. Nevertheless, health care datasets can still be affected by data bias. Due to data bias, they provide a distorted view of reality, leading to wrong analysis results and, consequently, decisions. For example, in a clinical trial that studied the risk of cardiovascular diseases, predictions were wrong due to the lack of data on ethnic minorities. It is, therefore, of paramount importance for researchers to acknowledge data bias that may be present in the datasets they use, eventually adopt techniques to mitigate them and control if and how analyses results are impacted.

This paper proposes a method to address bias in datasets that: *(i)* defines the types of data bias that may be present in the dataset, *(ii)* characterizes and quantifies data bias with adequate metrics, *(iii)* provides guidelines to identify, measure, and mitigate data bias for different data sources. The method we propose is applicable both for prospective and retrospective clinical trials. We evaluate our proposal both through theoretical considerations and through interviews with researchers in the health care environment.

Keywords: Data bias · Health care · Data management

1 Introduction

In the last decades evidence-based medicine relied consistently on large datasets. New technologies, such as data mining and data lakes, allow to analyze larger and larger datasets obtaining astonishing results that led to new and more reliable therapies [8], such as in the case of diabetic retinopathy detection [16,22]. Yet, reliable results directly depend on high-quality datasets. Biased datasets, i.e., dataset that do not correctly represent the population, inevitably lead to distorted results. This is particularly important for evidence-based medicine, where wrong or not effective therapies might be applied as a consequence of wrong analyses. For instance, in the controversial case of IBM Watson Oncology

system, machine learning algorithms return erroneous recommendations whenever the training of the system is not well planned [15]. The identification and measurement of bias in health care dataset is, therefore, essential to have reliable results and high-quality health care.

Even though bias and its impacts are well studied in specific research fields and application sectors of Computer Science, such as automated-decision making for hiring [23], justice [2] or image recognition [29], they are not frequently considered for health care datasets. Unluckily, information on bias is not provided with these datasets, leaving researchers and doctors clueless on the possible impacts on their analysis and with no means to mitigate them [10]. In order to get the benefits of event-based medicine and, therefore, to provide the best possible health care, information bias need to be considered as an essential part of datasets. This issue affects both retrospective clinical trials, which use already collected datasets that may be affected by data bias unknown at the time of data collection, yet relevant for the current trial, and prospective clinical trials, where datasets are collected for a specific clinical experiment.

This paper provides an insightful definition of the types of data bias that may affect health care datasets for retrospective clinical trials, and a method that will help health care researchers in: (i) identifying and measuring bias for health care datasets; (ii) allowing them to filter datasets based on bias information that will be stored on metadata; (iii) choosing the appropriate mitigation action to reduce bias before using the dataset. This method can be used for perspective clinical trials by monitoring the data collection and guide researchers towards the mitigation of identified bias. To reach these objectives, this paper defines a set of guidelines that will guide researcher on the application of the method. The method proposed in this paper has been validated through interviews with health care researchers and experts in bioethics and philosophy.

The paper is structured as it follows. Section 2 gives an overview of the context of bias and health care datasets. Section 3 describes the baseline of this paper, while Sect. 4 provides the most relevant research work on data bias in health care datasets. Sections 5 and 6 are the contribution of this paper: they describe the method proposed, the types of data bias involved, measurements and mitigation actions. Section 7 illustrates the initial validation of the method proposed. Section 8 concludes the paper.

2   Data Bias in Health Care

Recently, in the context of health care research, the employment of new information technologies is steadily increasing, driven by the admirable goal of improving effectiveness and efficiency of various aspects of medicine. For instance, machine learning proved to be extremely effective when used for detecting diabetic retinopathy [16,22] and skin cancer [13], with results equal to or greater than those of human physicians. Combining big data and machine learning algorithms has been shown to significantly improve the accuracy of analyses [8].

However, despite the enormous advantages of using computer technologies in health care, serious ethical issues emerged on several occasions. For example, it

has been shown that systems for algorithmic decision-making for medical diagnosis and treatment recommendations, despite improving accuracy, also increase opacity and uncertainty, especially when using deep learning techniques [15,36]. Opacity largely affects aspects such as accountability and fairness: famous is the case of IBM Watson Health, whose algorithms for cancer treatment recommendation turned out to be erroneous due to the small amount and the synthetic nature of the training data [15]. In this context, many of the emerged risks are related to data used to train the model or validate the underlying algorithms. In particular, the presence of bias in the data heightens the risk of unfairness and disparity. For instance, whenever a model is used for patients whose data differ from the data on which the model has been trained, results of the clinical trial are biased [31]. Also, bias may arise when the study population is not well selected [14].

In recent years, many ethical risks concern ethnic groups, which are often under-represented in clinical trials. Among the problems were the systematic misclassification of variants as pathogenic for Africans [25] and the inconsistent performance of risk calculators for cardiovascular disease in multi-ethnic studies, especially for ethnic minorities such as African Caribbeans and South Asian women [35]. Additionally, new frontiers of medicine such as genome sequencing tend to exclude ethnic minorities in a worrying way, as evident from the fact that by 2009 fewer than one percent of genome investigations included Africans [39].

3   Preliminaries

Considerable research attention is being paid to aspects of responsible data analysis and *Data Science Ethics*, which lays the foundation for data bias analysis. Data Science Ethics highlights the need for ethical analyses, focusing on concepts such as fairness, diversity, accountability, transparency and privacy. It analyzes the meaning and nature of computational operations, the interactions between hardware, software and data, considering the variety of digital technologies that make them possible. Data Science Ethics emphasizes the complexity of the ethical challenges posed by data science. The rest of this section gives definitions of fundamental concepts used in the rest of the paper: *Bias*, *Fairness* and *Diversity*.

Bias is defined as an "inclination, or prejudice for, or against one person, or group, especially in a way considered to be unfair" [30, p.4]. Sometimes bias is *desirable* and part of the correct system functionality. For example, in a system that predicts the likelihood of a person to have a specific disease, the group of individuals who actually has the disease should be systematically attributed a higher probability than the group who does not have it. An *undesired bias* is "a bias that is considered problematic, possibly unfair by the stakeholders of the system or other persons impacted by the system" [5, p.3]. Typically, this is observed when bias relates to a protected attribute, which is defined as "a sensitive attribute for which non-discrimination should be established" [38, p.2]. In the last few decades, many studies focused on the types of data bias that can be present in data [26,30].

Fairness, in the context of decision-making, is related to the concept of "absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics" [32, p.100]. It is also associated with the so-called "lack of bias" [33]. However, there is no universal definition of fairness, mainly because it is a broad concept that intersects different scenarios, thus in literature there are many formulations. To assess fairness there exist many formal metrics [38], such as statistical ones that evaluate the distribution of positive predictions among instances in the dataset.

Finally, **diversity** is "a general term used to capture the quality of a collection of items, or of a composite item, with regards to the variety of its constituent elements" [12, p.1]. There are many metrics to quantify diversity; some of them are based on the notion of distance and rely on pair-wise similarity between elements, while others are based on coverage, that measures the extent to which the elements cover a predefined number of aspects [12].

4   Related Work

In this section, we separate related work concerning methods for addressing data bias in generic datasets using Computer Science techniques, and, in the second part, methods that address data bias in health care.

4.1   Data Bias in Computer Science

Most of the works in Computer Science, and specifically in Data Science Ethics, focus on ensuring fairness, generally by algorithmic data bias discovery and mitigation in prediction tasks, more specifically for classification algorithms [1, 9, 37].

There are three possible approaches to enforce fairness in data analysis applications: *(i) pre-processing techniques*, i.e., procedures to verify that the training data are fair before the application of the algorithm; *(ii) in-processing techniques*, i.e., procedures to ensure that, during the learning phase, the algorithm does not inherit the bias present in the data, and *(iii) post-processing techniques*, i.e., procedures to correct the algorithm predictions, and consequently decisions, and make them fairer.

In this context, one of the most notable work is *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias* [9], an open-source framework to reach algorithmic fairness for classifiers. AI Fairness 360 quantifies and mitigates data bias using a variety of statistical measures, exploiting pre-processing, in-processing and post-processing techniques.

A work that studies both fairness and diversity is *Nutritional Labels for Data and Models* by Stoyanovich and Howe [34]. The authors develop an interpretability and transparency tool based on the concept of *Nutritional Labels*, drawing an analogy to the food industry, where simple and standardized labels convey information about ingredients and production processes. *Ranking Facts* [34] is a system based on nutritional labels that visually expresses fairness through statistical measures, and diversity through the distribution for each category.

One more recent work is [3], a system to assess diversity, and more specifically coverage, for a given dataset over multiple categorical attributes. Based on new efficient techniques to identify the regions of the attribute space not adequately covered by data, this system can determine the least amount of data that must be added to solve the lack of diversity.

4.2 Data Bias in Health Care

In the area of health care, the majority of work focuses on theoretical and philosophical aspects, and only few of them give more precise guidelines about data bias identification and measurement.

A famous work is [11]: the authors analyze major challenges when implementing predictive analytics in health care settings. They make broad recommendations for overcoming challenges raised in the four phases of the lifecycle of a predictive analytics model: acquiring data to build the model, building and validating it, testing it in real-world settings, and disseminating and using it more broadly. According to the authors, it is essential for predictive analytics models to be constantly evaluated, updated, re-implemented, and reevaluated, because the presence of bias threatens the trust of patients, providers and the public. Nonetheless, the authors remain optimistic that predictive analytics can help build stronger and more dynamic systems.

Another notable work is [31], that presents a conceptual framework of how different types of bias relates to one another. From our point of view, it is particularly relevant the classification of bias that arise from machine learning algorithms in the health care area. The authors provide a list of data bias definitions divided into four categories: *bias in model design*, *bias in training data*, *bias in interactions with clinicians* and *bias in interactions with patients*. They also reveal that the interactions of model predictions with clinicians and patients may exacerbate health care disparities.

As far as we know, there is no method in literature that specifically guides the user in searching and measuring data bias in the health care context, using the instruments from Computer Science and particularly Data Science Ethics.

5   Method

We propose a framework that consists in a pipeline for identifying and measuring bias in health care datasets. This work targets health care researchers and provides a set of application guidelines to be used during clinical trials.

Our method answers to the following questions that health care researchers should have before adopting a health care dataset for clinical trials.

– Q1. Is there any data bias in the dataset?
– Q2. Which ones are they and how to measure them?
– Q3. How to reduce their impact?

Several actors will benefit from this method. First, health care researchers, who mainly exploit medical data to carry out clinical trials; they also represent the target users of our method. Secondly, physicians and patients, since they are directly or indirectly involved in the application of results from clinical trials, and in the usage of automated systems for medical diagnosis based on machine learning techniques. Finally, ethics committees will benefit from having guidelines to apply in the health care research context, to extend the ethical principles of medicine also to new, not yet widespread computer-based tools.

The data management pipeline defined in [5] delineates five main stages of the lifecycle of decision support systems: requirements, system design, construction, testing, and maintenance. The method we propose will be used after data have been extracted from health care datasets, at the beginning of the construction phase and before the analysis phase. This temporal position allows our method to be applied after data have been—at least partially—collected, but before the construction of the analysis algorithm. In fact, in our opinion, clinicians need to modify data before the beginning of the construction phase.

The method uses the data bias identified in [31] and described in Sect. 4, where a list of data bias in the context of health care are defined. The method we are proposing considers bias in training data, since bias in model design and related to clinicians and patients cannot be analyzed from a computer scientist perspective, due to the lack of measurement strategies. In particular, considered data bias are: *minority, missing data, informativeness* and *selection bias.*

5.1   Input/Output

The input of the method consists of a dataset and information on its application context. The method targets reasonably large datasets that are going to be used in machine learning analytics for evidence-based medicine. The output of the method consists of a list of biases that have been identified in the data source, their quantification through appropriate metrics, and possible mitigation actions to solve them. This output allows to: *(i)* inform the researcher about the level of fairness associated to the input dataset by using a variety of bias metrics; *(ii)* increase awareness about the limits of the outcome using the input dataset in clinical trials and consequently, the risks connected to its application; *(iii)* have adequate information for further uses of the datasets, that can be stored in metadata; *(iv)* choose a mitigation strategy that, by applying it, can produce a cleaned version of the input data source.

5.2   Steps

Figure 1 shows the pipeline of the method proposed in this paper. All steps, described below, will be executed by the health care researcher.

1. *Bias Identification* includes the study and analysis on the presence of different types of data bias.

2. *Bias Measurement*, where each data bias is mapped with its corresponding measurement techniques.
3. *Bias Mitigation*, where each data bias is associated to possible mitigation strategies.

The rest of this section provides a description of each step, while the next section details each bias, the respective metrics and mitigation actions.
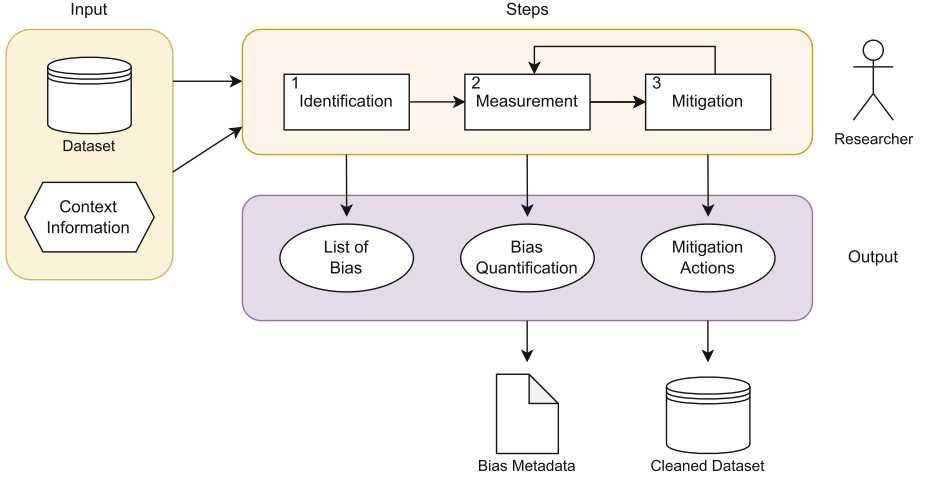
Fig. 1. The pipeline of the proposed method.

Bias Identification. This step consists in identifying data bias that may affect the dataset, and that need to be mitigated if present. Possible data bias are: minority, missing data, informativeness and selection bias. For each of them, we report the definition and a list of questions that the researcher should answer to acquire knowledge of the specific bias, in particular on the risks deriving from their presence. For this reason, the method we propose takes the form of guidelines for health care researchers to perform bias analysis on health data sources, to understand the risks connected to their usage and better know how to mitigate them.

Bias Measurement. This step consists in quantifying the identified data bias, in order to understand the magnitude of the possible impact on the results of the clinical trial. For each data bias we give an overview of the most suitable metrics, taken from the Data Science Ethics literature, to recognize and quantify it. The results of this step can be stored in metadata associated to the dataset, to be reused for future clinical trials. In fact, it is common for a dataset to be used for additional clinical trials [21,28]. Therefore, it is important to save the previously

extracted information on data bias for further analyses not originally planned. To do so, our method stores information on the presence and magnitude of bias.

Bias Mitigation. The last step of the pipeline consists in choosing a mitigation action based on the previously used metric, the context and the goal of the clinical trial. The choice of the mitigation action is strictly related to the context and the goal of the trial, since each action potentially has different consequences on the dataset. The best solution does not always consist in adopting a mitigation action that modifies the dataset: there are cases, such as breast cancer, in which the disease has a different incidence, so the dataset is naturally skewed towards a part of the population. In fact, sometimes the bias is not relevant for the goal of the clinical trial, so the questions in the first step should guide the user in this step as well; on the contrary, when the bias is undesirable or important, the user should perform a mitigation action. In this work, we propose only pre-processing actions, allowing health care researchers to modify the dataset before the beginning of the construction phase. The mitigation actions presented are taken from Computer Science and Data Science Ethics literature. Additionally, the application of a mitigation action needs to be carefully considered, since lowering a value of a metric to mitigate data bias, may raise other metrics values of the same or different data bias. For this reason, the method we propose is iterative: after the mitigation step, the user applies the measurement step again in order to evaluate the cleaned dataset. However, we will not further discuss this aspect since it is not the scope of this paper.

6   Bias Identification and Measurement

This section explores the bias introduced in the previous Sect. 5, giving the definition, the measurement and the possible mitigation strategies for each one.

6.1   Minority Bias

Definition. The protected groups may have insufficient numbers of patients for a model to learn the correct statistical patterns [31, p.3].

Identification. During this step, it is important to focus on the attributes and features that are in the dataset. The following questions help health care researchers to identify this data bias.

– Which attributes and features are present in the dataset?
– Which are more relevant for the study?
– Are there protected attributes for the study? Are gender or ethnicity present in the dataset?

A group can be identified by specifying one or more characteristics of the attributes present in the dataset. Some examples of groups are: females, males,

Caucasian, Black, Asiatic, Caucasian females, Black females, etc. If at least one of these characteristics is protected, the entire group is also protected.

After answering the questions, the health care researcher starts combining the information given by the context and the goal of the clinical trial, with the identified groups. If she concludes that, given the clinical trial, the dataset is not affected by minority bias, because, for example, there is only one group, then the risks and the issues illustrated in Sect. 2 are prevented. For this reason, the identification of different groups, particularly the protected ones, is fundamental.

Measurement. To measure the number of people in the (protected) group and, consequently, the minority bias, we identify three different measurements: density, diversity and fairness. Each one studies different aspects of data representation and distribution.

First, *Density* is the degree to which different entities occur in the dataset. It is the number of occurrences of a value [27], so it represents the easiest metric to study. Figure 2 shows the density for a simple example where there are two attributes: "sex" has two values "Female" and "Male" and "ethnicity" has three values "Caucasian", "Black" and "Asiatic". It is evident from the plot that the minorities are females, Black people and Asiatic people.
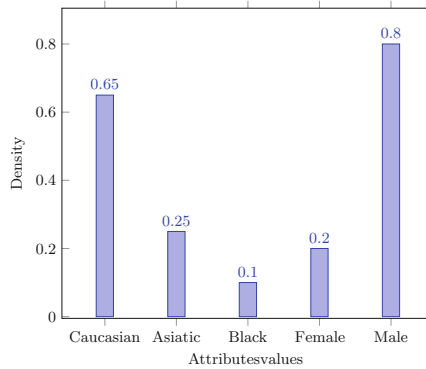
Fig. 2. Density plot

Another measurement is *diversity* that "ensures that different kinds of objects are represented in the data" [12, p.1]. The most popular metrics for diversity are based on the notions of distance or coverage [12]. A system that defines and studies coverage is given in [4], briefly described in Sect. 4. The authors developed efficient algorithms to identify uncovered regions of data. Figure 3, taken from [4], shows a sample dataset with two attributes x_1 and x_2. Every tuple in the dataset is represented by a black dot, and the uncovered region is colored in red. A green zone is a region that has a sufficient number of samples, while a red zone evidences a lack of elements and, consequently, the presence of minority bias.
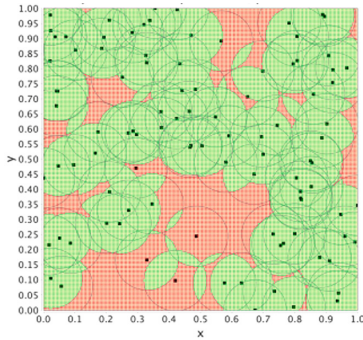
Fig. 3. Diversity plot

sex	
Privileged group	*Male*
Unprivileged group	*Female*
Minority bias against unprivileged group detected in	*2 out of 5 statistical metrics*
ethnicity	
Privileged group	*Caucasian*
Unprivileged groups	*Black, Asiatic*
Minority bias against unprivileged group detected in	*4 out of 5 statistical metrics*

Fig. 4. Fairness plot

To measure *fairness*, the most used metrics are statistical ones [38]: they evaluate the predictions, given by a classification model based on the dataset, compared with the actual instances of the dataset. An example of fairness analysis is given in Fig. 4. In this example we selected five statistical metrics and we analyzed the outcome of the predictions for the different groups. It is worth noticing that the metrics detect minority bias against unprivileged groups, namely "Female", "Black" and "Asiatic" groups. There are tools off-the-shelf that can provide this analysis, such as [9], briefly described in Sect. 4.

Bias Mitigation. Different techniques can be adopted to mitigate minority bias. The choice depends on the goal of the clinical trial and on the metric chosen by the health care researcher. Indeed, if the clinical trial only focuses on the majority groups, without the goal of studying the behavior of minorities, there is no need to change the group distributions, and the researcher should not modify the dataset. Otherwise, she should perform a mitigation action. For space limitation, we present the ones described in Sect. 4.

If the researcher is interested in density and diversity, the mitigation action to apply is [3], that determines the least amount of data that the researcher must add to solve the lack of representation previously discovered. If she is interested in fairness, the dataset can be modified using one of the pre-processing techniques presented in [9], such as *reweighing*, that weights the examples in each (group, label) combination differently to ensure fairness before classification, or *optimized pre-processing*, that learns a probabilistic transformation to modify the features and the labels in the data.

6.2   Missing Data Bias

Definition. Data may be missing for protected groups in a nonrandom fashion, which makes an accurate prediction hard to render [31, p.3].

Identification. "A missing value is a value that exists in the real world, but is not available in a data collection" [6, p.7]. The following questions may help health care researchers to identify this data bias.

– Are there any missing values? If so, why?

It is important to understand the reason why a value is missing. In fact, missing values are very frequent in questionnaires for medical scenarios [20]. There are three main types of missing values: *(i)* missing completely at random, i.e., the value does not depend on observed attributes or missing value, *(ii)* missing at random, i.e., it depends on observed attributes, *(iii)* missing not at random, when the value that is missing depends on the value itself [24]. To identify this data bias, the user has to concentrate on the last type of missing value. It is critical to understand the types of missing values that the dataset may have, because the user has to understand the reason why the value is missing. Unfortunately, rarely researchers can determine the type of missing value solely based on the data at hand, so context information must be considered as well [20].

Measurement. In large datasets, identifying missing values can be difficult and often requires domain knowledge. Completeness, defined as "the ratio between the number of not null values over the total number of values" [6, p.19], can be used to measure the amount of missing values. It can be computed on values, on tuple, or on attributes [7]. Given the example used for the minority bias, Table 1 shows a completeness analysis with the percentage of missing values for each attribute value.

Bias Mitigation. Three main methods can be used to mitigate the presence of missing values [19]: *(i) Deletion*, in which the researcher discards all the data with a missing values; *(ii) Single Imputation* in which the researcher converts the missing values into a new value, for example substituting the mean, the mode or other statistical values; *(iii) Model-Based*, in which the researcher assign a new value, based on a model created on the rest of the dataset. The researcher should decide the best strategy that yields the least biased estimations in the results, taking into account the context and the goal of the clinical trial. Moreover, fixing all missing values may require a considerable amount of effort. Therefore, tuples can be marked based on utility contribution, and prioritized based on the research goals. Additionally, to improve cost/effectiveness, the researcher may consider fixing only attributes that can contribute more to the prediction [6].

6.3   Informativeness Bias

Definition. Features may be less informative to render a prediction in protected groups [31, p.3].

Identification. The following questions help health care researchers to identify
this data bias.

Table 1. Completeness analysis

| Attribute value	% of missing values
Black	15%
Caucasian	5%
Asiatic	30%
Female	25%
Male	0%

– Which features are more important for predicting the result?
– Are there features for which more information is needed?

To understand which features are used to predict the outcome and how much
they are important, the researcher has to analyze the prediction algorithm. Given
an input dataset, the output of an algorithm mainly depends on the technique
adopted.

Measurement. In the machine learning field, advanced techniques such as neural
networks create very complex models, and the process that lead to results is not
easy to understand; for this reason, they are said to be opaque. The purpose of
explainability is to make explicit the interactions among the model, the learning
technique adopted to produce it, and the data on which it operates; this is
relevant both when we want to understand how the model works, and when we
want to explain the outcome of the model for a single individual or group.

Thus, it is fundamental to study *Explainable Machine Learning*, which "high-
lights decision-relevant parts of the used representations of the algorithms and
active parts in the algorithmic model" [17, p.3]. If we consider a prediction
algorithm like logistic regression, to explain the output of the model and con-
sequently measure this data bias, feature importance techniques can be used.
In these tasks, the importance of each feature is given by the coefficient that
the model associates to each feature during the training phase: these coefficients
indicate how much each attribute value impacts the final output. Therefore, in
order to evaluate a model with respect to the informativeness bias, it is sufficient
to study and compare the feature coefficients, and to control whether the most
important attributes also exhibit unfair behaviors. This way we can establish
whether the learned model is informative for all groups or not.

Figure 5 shows the feature importance of the previous example: Caucasian,
Male and Asiatic group have a higher probability to get a positive outcome with
respect to Black and Female that, due to the negative coefficients, will probably
receive a negative one.
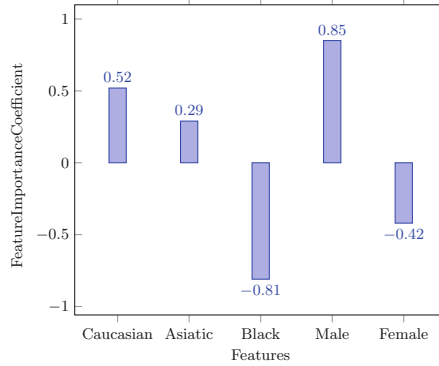
Fig. 5. Feature importance

Bias Mitigation. Given the goal and the context of the clinical trial, if all the features are informative enough, the researcher can leave the model as it is. However, as said before, not all algorithms are explainable, and feature importance is not always computable. In this case, two approaches can be adopted: *(i) post-hoc systems* provides local explanations for a specific decision and make it reproducible on demand, instead of explaining the whole system behavior, *(ii) ante-hoc systems* implements algorithms interpretable by design towards specific approaches [18], such as linear regression and decision trees. According to the context information and to the goal of the clinical trial, the researcher can adopt one of these two techniques to solve the problem of informativeness bias.

6.4   Selection Bias

Bias Definition. The training data may not be representative of the population, or the deployment data may differ from the training data. [31, p.3].

Identification. In this step, the researcher answers to the following questions.

– Are there training and testing phases for the study?
– Has the model been evaluated? If yes, how?

Given the whole available data, the procedure to construct a data analysis algorithm is the following: in a first phase, named "training phase", the model is built using a portion of the data, then in the second phase named "testing phase" the model is evaluated with the remaining data. In some cases, a part of the available data is reserved to develop the model before production: this is called "validation set" or "development set". For the researcher it is useful to understand how these sets are produced, and whether they are similar and representative of the population.

Measurement. Given the sets used for training, testing and the eventual validation, in order to measure this type of data bias, the researcher analyzes and compares the distributions of these sets. A general *Data Exploration* of the sets based on plots, summary statistics and correlation analysis is highly recommended to understand the shape of the data. Figure 6 plots the training and test sets distributions, with respect to a feature Y, from the example described above. The plot suggests that the two distributions are not similar, so in this case the training set and the test set are not equally representative of the population.

To precisely measure the difference between two data distributions and to compare them, *statistical tests* can be adopted. In general, a statistical test defines a null hypothesis H0 (in our example H0 is "the two distributions are equal"), an alternative hypothesis H1 (in this case "the two distributions are different"), and a significance level α, i.e. the probability of rejecting H0. The sampling distribution of the statistical test is computable either exactly or approximately, and the value obtained is the p-value. If the p-value is less than or equal to the chosen significance level α, then H0 is rejected in favor of H1; otherwise, if the p-value is greater than α, the null hypothesis is accepted. Usually, values of α such as 0.05 or 0.01 are chosen.
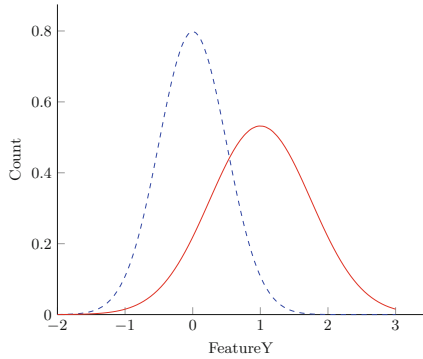
Fig. 6. Training and test sets distributions

Bias Mitigation. If the measurement step showed some inconsistencies, a mitigation procedure is needed. The split between training and test set should be performed avoiding selection bias. The most common option is to assign two third of the data for training and one third for testing. However, for small or "unbalanced" datasets, samples might not be representative. A more advanced option is to use *K-fold cross-validation* [40]: after splitting data into K subsets of equal size, each subset is used for testing one at a time, with the remaining portion of data being used for training. Standard method for evaluation is to use "repeated K-fold cross-validation", in which the K-fold cross-validation is repeated multiple times. Usually K is chosen to be 5 or 10. The special case, when $K = n$, is called "leave-one-out cross-validation", where the testing set comprises

a single point and the remaining data is used for training purposes [40]. Table 2 shows a summary of all types of data bias presented with the corresponding measurement and mitigation techniques.

Table 2. Data bias summary table

Type of data bias	Measurement	Mitigation
Minority	Density Diversity Fairness	Diversity enhancement Reweighing Optimized pre-processing
Missing value	Completeness	Deletion Single imputation Model-based methods
Informativeness	Explainable ML Feature importance	Ante-hoc systems Post-hoc systems
Selection	Data exploration Statistical tests	K-fold cross-validation

7   Validation

A preliminary validation of our method consisted in semi-structured interviews with subjects working in the health care sector. We interviewed 5 subjects: *(i)* an oncologist and experienced health care researcher; *(ii)* a philosopher with expertise in bioethics working in a research institute; *(iii)* an experienced imaging and neuro-anatomy technician working in a research laboratory; *(iv)* a mathematician and philosopher, expert in data analysis and health care ethics; *(v)* a physician and pharmacologist working in a research institute and member of ethics committee.

In these interviews, we described our method, being as impartial as possible, and asked the following questions:

– Is the temporal placement of the method correct?
– Is the list of data bias complete?
– Is data bias correctly defined?
– Is the list of metrics complete for each data bias?
– Do the metrics correctly represent the data bias?
– Is the list of mitigation actions complete?
– Is the pipeline functional to the objectives of the method?
– Are the outputs of the method valuable for the health care researchers?

All subjects answered positively to the questions, confirming the validity of the method and underlining its importance for health care researchers and

datasets. Through their answers, the subjects highlighted relevant aspects for our research such as: the seriousness of the consequences of data bias in clinical trials, the importance of contextual information and the goal in clinical trials in the application of the method, the need of a common procedure for identifying, measuring and possibly mitigating data bias. Furthermore, subjects highlighted that, for the effective usage of the method, a collaboration with a computer scientist becomes fundamental, considering that a health care researcher may not be familiar with mitigation actions. This is especially true since not all of the mitigation actions are available off-the-shelf or easily usable, due to the complexity of understanding them.

From these valuable interactions we have been able to enrich and improve our research, even correcting and adapting the method to the demands and needs of the researchers.

8   Conclusion

Clinical trials based on big data and machine learning techniques, such as data lakes and data mining, allow to analyze larger and larger datasets obtaining precise results through efficient procedures, leading to new and more reliable therapies. However, it has been shown that the complexity of their operations significantly reduces the transparency of the results obtained from their use. Therefore, it is of paramount importance to analyze and measure data bias in health care datasets, to raise the trustworthiness of clinical trials and avoid relying on erroneous results.

This paper proposes a method that guides health care researchers in achieving this objective with a pipeline that defines three steps for the identification of data bias, their measurement and mitigation. The method has been validated with semi-structured interviews that underlined the importance of the method and its outputs.

Future work will include: *(i)* further analyze the relation between data bias and mitigation actions, improving the questions provided to health care researchers; *(ii)* structure metadata for health care datasets, and integrate them in our method; *(iii)* conduct experiments on real health care datasets using our method; *(iv)* build a software tool that supports our method.

References

1. Adebayo, J.A., et al.: FairML: toolbox for diagnosing bias in predictive modeling. Ph.D. thesis, Massachusetts Institute of Technology (2016)
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. In: Ethics of Data and Analytics, pp. 254–264. Auerbach Publications (2016)

3. Asudeh, A., Jin, Z., Jagadish, H.: Assessing and remedying coverage for a given dataset. In: 2019 IEEE 35th International Conference on Data Engineering, pp. 554–565. IEEE (2019)

4. Asudeh, A., Shahbazi, N., Jin, Z., Jagadish, H.: Identifying insufficient data coverage for ordinal continuous-valued attributes. In: Proceedings of International Conference on Management of Data, pp. 129–141 (2021)

5. Balayn, A., Lofi, C., Houben, G.-J.: 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. VLDB J. **30**(5), 739–768 (2021). https://doi.org/10.1007/s00778-021-00671-8

6. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. **41**(3), 1–52 (2009)

7. Batini, C., Scannapieco, M.: Data and Information Quality. DSA, Springer, Cham (2016). https://doi.org/10.1007/978-3-319-24106-7

8. Beam, A.L., Kohane, I.S.: Big data and machine learning in health care. Jama **319**(13), 1317–1318 (2018)

9. Bellamy, R.K., et al.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J. Res. Dev. **63**(4/5), 1–4 (2019)

10. Char, D.S., Shah, N.H., Magnus, D.: Implementing machine learning in health care—addressing ethical challenges. N. Engl. J. Med. **378**(11), 981 (2018)

11. Cohen, I.G., Amarasingham, R., Shah, A., Xie, B., Lo, B.: The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Affairs **33**(7), 1139–1147 (2014)

12. Drosou, M., Jagadish, H.V., Pitoura, E., Stoyanovich, J.: Diversity in big data: a review. Big Data **5**(2), 73–84 (2017)

13. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)

14. Gerhard, T.: Bias: considerations for research practice. Am. J. Health Syst. Pharm. **65**(22), 2159–2168 (2008)

15. Grote, T., Berens, P.: On the ethics of algorithmic decision-making in healthcare. J. Med. Ethics **46**(3), 205–211 (2020)

16. Gulshan, V., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama **316**(22), 2402–2410 (2016)

17. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **9**(4), e1312 (2019)

18. Holzinger, A., Plass, M., Holzinger, K., Crisan, G.C., Pintea, C.M., Palade, V.: A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop. arXiv preprint arXiv:1708.01104 (2017)

19. Ibrahim, J.G., Chen, M.H., Lipsitz, S.R., Herring, A.H.: Missing-data methods for generalized linear models: a comparative review. J. Am. Stat. Assoc. **100**(469), 332–346 (2005)

20. Ibrahim, J.G., Chu, H., Chen, M.H.: Missing data in clinical studies: issues and methods. J. Clin. Oncol. **30**(26), 3297 (2012)

21. Knoppers, B.M.: International ethics harmonization and the global alliance for genomics and health. Genome Med. **6**(2), 1–3 (2014)

22. Krause, J., et al.: Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology **125**(8), 1264–1272 (2018)

23. Lambrecht, A., Tucker, C.: Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. Manag. Sci. **65**(7), 2966–2981 (2019)
24. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data, vol. 793. John Wiley & Sons, Hoboken (2019)
25. Manrai, A.K., et al.: Genetic misdiagnoses and the potential for health disparities. N. Engl. J. Med. **375**(7), 655–665 (2016)
26. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 1–35 (2021)
27. Naumann, F., Freytag, J.C., Leser, U.: Completeness of integrated information sources. Inf. Syst. **29**(7), 583–615 (2004)
28. van Ommen, G.J.B., et al.: BBMRI-ERIC as a resource for pharmaceutical and life science industries: the development of biobank-based expert centres. Eur. J. Hum. Genetics **23**(7), 893–900 (2015)
29. Papakyriakopoulos, O., Mboya, A.M.: Beyond algorithmic bias: a socio-computational interrogation of the google search by image algorithm. Soc. Sci. Comput. Rev. (2021). https://doi.org/10.1177/08944393211073169
30. Pitoura, E.: Social-minded measures of data quality: fairness, diversity, and lack of bias. J. Data Inf. Qual. **12**(3), 1–8 (2020)
31. Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., Chin, M.H.: Ensuring fairness in machine learning to advance health equity. Ann. Internal Med. **169**(12), 866–872 (2018)
32. Saxena, N.A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D.C., Liu, Y.: How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. Artif. Intell. **283**, 103238 (2020)
33. Stoyanovich, J., Abiteboul, S., Miklau, G.: Data, responsibly: fairness, neutrality and transparency in data analysis. In: International Conference on Extending Database Technology (2016)
34. Stoyanovich, J., Howe, B.: Nutritional labels for data and models. IEEE Data Eng. Bull. **42**(3), 13–23 (2019)
35. Tillin, T., et al.: Ethnicity and prediction of cardiovascular disease: performance of qrisk2 and Framingham scores in a UK tri-ethnic prospective cohort study (sabre—southall and brent revisited). Heart **100**(1), 60–67 (2014)
36. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. **25**(1), 44–56 (2019)
37. Tramer, F., et al.: Fairtest: discovering unwarranted associations in data-driven applications. In: IEEE European Symposium on Security and Privacy, pp. 401–416 (2017)
38. Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (fairware), pp. 1–7 (2018)
39. Wapner, J.: Cancer scientists have ignored African DNA in the search for cures. Newsweek Magazine (July 2018). https://www.newsweek.com/2018/07/27/cancer-cure-genome-cancer-treatment-africa-genetic-charles-rotimi-dna-human-1024630.html. Accessed 23 June 2022
40. Zaki, M.J., Meira Jr, W.: Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Cambridge University Press, Cambridge (2020)