Achieving Fair Inference Using Error-Prone Outcomes

Laura Boeschoten**, Erik-Jan van Kesteren**, Ayoub Bagheri, Daniel L. Oberski *

Utrecht University, Department of Methodology & Statistics, Utrecht (The Netherlands)
**Shared first authorship. These authors contributed equally to the current work

Received 16 October 2020 | Accepted 20 December 2020 | Published 22 February 2021

unir
LA UNIVERSIDAD EN INTERNET

Abstract

Recently, an increasing amount of research has focused on methods to assess and account for fairness criteria when predicting ground truth targets in supervised learning. However, recent literature has shown that prediction unfairness can potentially arise due to measurement error when target labels are error prone. In this study we demonstrate that existing methods to assess and calibrate fairness criteria do not extend to the true target variable of interest, when an error-prone proxy target is used. As a solution to this problem, we suggest a framework that combines two existing fields of research: fair ML methods, such as those found in the counterfactual fairness literature and measurement models found in the statistical literature. Firstly, we discuss these approaches and how they can be combined to form our framework. We also show that, in a healthcare decision problem, a latent variable model to account for measurement error removes the unfairness detected previously.

Keywords

Algorithmic Bias, Fair Machine Learning, Latent Variable Model, Measurement Error, Measurement Invariance.

DOI: 10.9781/ijimai.2021.02.007

I. Introduction

Supervised learning is used to guide human decisions across a wide range of different fields. In sensitive areas such as healthcare or criminal justice, a key issue is that these decisions are equitable and fair. To this end, an active area of research investigates how fairness criteria can be incorporated into supervised learning [1]–[6]. This literature has focused on supervised learning for a single objective, assumed to be the target variable of interest.

However, focusing on fair inference for a single objective is not sufficient in many real-world applications. The motivating example for this paper is presented in [7]: a commercial health prediction algorithm, widely used by health insurance companies and affecting millions of patients, exhibits significant racial bias –at a given risk score, black patients are considerably sicker than white patients, as evidenced by signs of uncontrolled illnesses. The bias arises because the algorithm predicts healthcare costs rather than illness, but unequal access to care means that less money is spent caring for black patients than for white patients. Thus, substantial racial biases arise, despite healthcare cost appearing to be an effective proxy for health by some measures of predictive accuracy, and despite these predictions complying with conventional standards of fair inference on outcomes [8]. The situation presented in [7] is but one example of a more general common framework of using a proxy to measure outcomes which cannot be directly measured – another example would be predicting true criminal recidivism using only observed recidivism, which is an error-prone proxy [9]. In this paper, we suggest using an approach from the field of social science: to make use of *multiple* observable proxies to build a measurement model representing the unobserved (latent) variable of interest. We propose to integrate such an approach when developing prediction models. This issue cannot be ignored because fairness is generally conceptualised on a level more abstract than the proxy label [10]; for example, it is reasonable to require that fairness in a healthcare need prediction system should extend to a person's true health status. However, it is challenging to measure a patient's true health status, as such measures are typically impossible to observe directly. In social science, a common approach is to make use of multiple observable indicators to build a measurement model representing the unobserved (latent) variable of interest. We propose to integrate such an approach when developing prediction models.

This paper addresses the problem of prediction unfairness arising from measurement error. By considering the supervised learning problem at the level of a latent variable of interest, we reformulate the problem as one of adequate *measurement modelling*. In effect, instead of requiring perfect measurement to achieve fairness, we propose that researchers developing a prediction model to be used for decision-making collect several independent, possibly error-prone, measures of the variable of interest (e.g. health).These measures act like error-prone labels made by independent annotators, each containing some information about the true health status (similar to, e.g., [11],[12]). We then suggest to combine measurement models from the statistical literature with techniques from the literature on fair ML to assess and ameliorate the problem of unfair predictions in the face of measurement error.

Our contributions are as follows:

- We illustrate that existing methods to examine unfairness in error-prone outcomes are insufficient;
- We suggest a framework, based on the existing measurement modelling literature, to investigate and ameliorate such issues;
- We perform an exemplary analysis to demonstrate the sug-gested approach. In an existing healthcare application, this demonstrates that replacing one proxy with another does not lead to parity, while our approach does.

* Corresponding author.
E-mail addresses: l.boeschoten@uu.nl (Laura Boeschoten), e.j.vankesteren@uu.nl (Erik-Jan van Kesteren), a.bagheri@uu.nl (Ayoub Bagheri), d.l.oberski@uu.nl (Daniel L. Oberski).

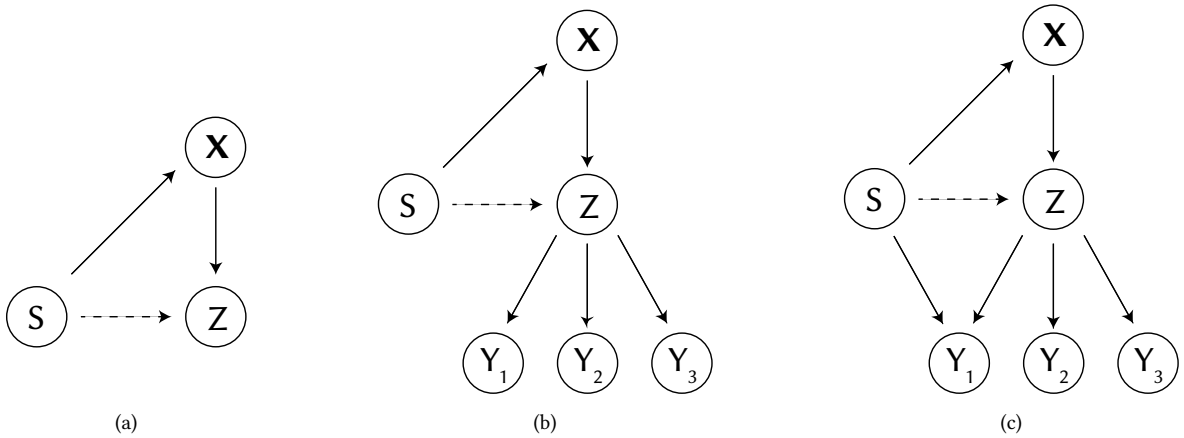

Fig. 1. Graphical representation of causal relations between the sensitive feature (S), the predictors ($\mathbf{X}$), and the error-prone outcome (Z) in the naive case (A), in the measurement error framework (B), and in the measurement error framework with differential item functioning on the Y_1 proxy (C). The dotted arrow indicates the discriminatory causal pathway (as in [8]) which is blocked when performing fair inference, evaluating $E[Z \mid \mathbf{X}, S]$ to compute a risk score $\hat{Z}$.

In Section II, we provide a summary of basic concepts in fairness. In Section III prior approaches with respect to fair inference are discussed. In Section IV, the failure of these approaches is discussed when making use of proxies, and the proposed framework is introduced based on existing measurement models. In Section V the proposed framework is then applied to the exemplary data set provided by [7].

II. Problem Definition

We consider probabilistic classification and regression prob-lems with a set of features $\mathbf{X}$ and true outcome Z. Among the features, there is a sensitive feature $S \in \mathbf{X}$ (e.g. race, gender), with respect to which discriminatory predictions are to be avoided. Furthermore, although the prediction problem is with respect to the true outcome Z – e.g. "health" or "crime" – this outcome is not directly observed; instead, we have observed a set of error-prone proxy variables $\mathbf{Y}$. For example, in practice a proxy for "health", $Y \in \mathbf{Y}$, might be the costs of healthcare or the number of chronic conditions experienced by the patient, whereas, instead of "crime", the number of arrests might be measured. Following [8], we represent the goal of the regression or classification problem as a query on the (generative) joint distribution $p(Z, \mathbf{X})$, potentially after conditioning on a set of "fixed" covariates C, i.e. the (discriminative) conditional joint $p(Z, \mathbf{X}\backslash\mathbf{C} \mid \mathbf{C})$. Typically, this query will be the point prediction $\hat{Z} := E(Z \mid \mathbf{X})$.

Following standard social-scientific measurement theory [13], the fact that $\mathbf{Y}$ is a measurement proxy for Z is reflected by a *causal model*, in the sense of [14], [15], in which $Z \rightarrow \mathbf{Y}$, i.e., the true outcome is a common cause of all available proxy variables. Because Z is an unobserved latent variable, our causal model will be identifiable only through additional assumptions of conditional independence; we discuss these assumptions later. The key point to note here is that, generally,

$$E(Z \mid \mathbf{X}) \neq E(Y \in \mathbf{Y} \mid \mathbf{X}) \tag{1}$$

i. e. predictions using error-prone proxies as labels, $\hat{Y}$, will, of course, differ from the $\hat{Z}$ that would have been obtained had the true labels been available.

III. Related Work

A large and growing literature on fairness of predictions for the error-free outcome Z exists, with divergent and sometimes mutually exclusive definitions of the notion of algorithmic fairness. An excellent overview of this literature can be found in [6], which identified 20 separate definitions. Broadly, a distinction can be made between statistical metrics, distance-based measures, and causal reasoning [6].

Statistical metrics define fairness as the presence or absence of a (conditional) independence in the joint distribution $p(Z, \hat{Z}, S)$. For example, take a classification problem in which the decision is taken as $d := I(Z > \tau)$, where I is the indicator function and τ is some threshold on the predicted score. *Statistical parity* ("group fairness") is then defined as

$$p(d = 1 \mid S = s) = p(d = 1 \mid S = s') \tag{2}$$

for all $s \neq s'$, i.e., the decision should not depend on the sensitive attribute, whereas *predictive parity* is defined as

$$p(Z = 1 \mid d = 1, S = s) = p\,(Z = 1 \mid d = 1, S = s'\,) \tag{3}$$

for all $s \neq s'$—i.e. the positive predictive value should not depend on the sensitive attribute. Further definitions include conditional statistical parity [2], overall accuracy equality [1], and well calibration [4].

Distance-based measures of fairness account for the non-sensitive predictors $\mathbf{X}\backslash S$, in addition to the observed and predicted outcomes and sensitive attribute. The well-known "fairness through awareness" framework [3] generalises several of the preceding notions, such as statistical parity, by defining fairness as "similar decisions for similar people". Consider a population of potential applicants P, and consider any randomised output from the prediction algorithm, $M(x \in P)$. Fairness is achieved whenever the distance among the decisions M made for two people is at least as small as the distance between these people, i.e. when

$$D\,(M(x), \mathrm{M}(y)) \leq d(x, y) \tag{4}$$

for any $x, y \in P$. Here, D and d are arbitrary metrics on the distance between outputs and people, respectively. Careful choice of these metrics can yield some of the above definitions as special cases. Since the fairness condition can be trivially achieved, for example by always outputting a constant regardless of the input, the prediction model should be trained by minimising a loss function under the above constraint.

Finally, in recent years, results from the causal modelling literature have been leveraged to define and achieve "counterfactual" fairness [5], [8]. In these definitions one first considers a causal model involving Y, $X\backslash S$, and S such as Panel A of Fig. 1. This causal model then induces a counterfactual distribution $p_{do(s)}(\hat{Z}|X)$, i.e. the distribution we would observe if S were set to the value s [14]. [5] then defined counterfactual fairness as

$$p_{do(s)}(\hat{Z}|X) = p_{do(s')}(\hat{Z}|X) \tag{5}$$

Note that this definition looks superficially similar to the definition of statistical parity (group fairness), but is distinct because it refers to an individual. This definition has as a disadvantage that *any* causal effect of the sensitive attribute on the prediction is deemed illegitimate. Based on the same framework, [8] suggested a more general definition: some causal pathways originating in S are denoted discriminatory, while others are not. Fairness is then achieved by performing inference on a distribution $p^*(Z, \mathbf{X})$, in which the "fair world" distribution $p^*(Z, \mathbf{X})$ is close in a Kullback-Leibler sense to the original $p(Z, \mathbf{X})$, but all discriminatory pathways have been blocked (up to a tolerance) using standard causal inference techniques. Note that, if all causal pathways originating in S are deemed discriminatory and the tolerance set to zero, the counterfactual fairness criterion by [5] will be satisfied.

IV. Proposed Framework

A. Fair Inference in Error-prone Outcomes

The existing methods from Section III do not consider the target Z to be error-prone. However, in practice, the target feature $Y \in \mathbf{Y}$ in the data set is not a perfect representation of the true underlying outcome Z. There can be several sources for this imperfect representation. For example, the true underlying outcome of interest may not be directly measurable at all (i.e., $Z \neq Y$ for any possible Y). In this case, the outcome of interest will only partially explain any feature used as its proxy. For example, in using healthcare costs Y as a proxy for health Z, the observed value will in part be determined by other factors besides Z, such as the location of residence of the patient. Then, even if the outcome of interest were "true healthcare costs" – thus in principle measurable – the observed feature will in practice still not be an infallible proxy, because health records are never perfect observations and always contain some form of noise [16]. Together, such sources of noise in the observation process are termed "measurement error", and any outcome Z containing measurement error can be considered *latent* [17] and modelled as such.

Crucially, the presence of measurement error may result in unfair inferences for the error-prone outcome, even after applying the procedures presented in Section III to account for unfairness. This is shown in a compelling example by [7], who concluded that commercial algorithms used by insurance companies for patient referral contain a fundamental racial bias. In the algorithm under consideration, healthcare costs $Y \in \mathbf{Y}$ are used as a proxy for health Z. [7] illustrated that although there is no bias in healthcare costs, there is strong racial bias in other proxies of health such as whether patients have chronic conditions. Specifically, in order to be referred to a primary care physician, the true underlying health status Z of black patients was worse than that of white patients.

[7] concluded that fair inference requires selecting a better proxy for health as the outcome variable Z. Indeed, their analyses were possible precisely due to the availability of different proxies of health, such as the number of chronic conditions. However, we note that solving racial bias in a new proxy does not guarantee the absence of racial bias in other proxies indicating other aspects of health. Instead, here we suggest incorporating several proxies, or *indicators* $\mathbf{Y}$ in a measurement model for the unobserved, error-prone outcome Z [18]. In the next section, we introduce the existing literature on measurement models and its approach to fair inference.

B. Fair inference in Measurement Models

When outcomes are thought to be error-prone, an existing literature suggests the use of measurement models [16], [19]. At their core, measurement models describe the causal relationship between observed scores $\mathbf{Y}$ and unobserved "true scores" Z as $Z \rightarrow \mathbf{Y}$. A measurement model adequately represents the empirical conditions of measurement if conditional independence can be assumed [20]. More specifically, measurement models assume that Y_1 and Y_2 are conditionally independent given Z, i.e.,

$$p(Y_1, Y_2 \mid Z) = p(Y_1 \mid Z)\, p(Y_2 \mid Z) \tag{6}$$

A plethora of variations of measurement models assuming conditional independence have been developed, such as latent class models [21], item response models [22], mixture models [23], factor models [24], structural equation models [25], and generalised latent variable models [26].

Measurement models are suggested here as a convenient way to account for a latent variable's relationship to sensitive features. The measurement error of a proxy variable (e.g. Y_1) is then assumed to differ over different groups of S. To account for group differences in proxy variables, a large body of literature is available where this issue is known under different labels. Generally, these approaches are applied within the structural equation modelling (SEM) framework [27], as SEM explicitly separates the measurement model ($Z \rightarrow \mathbf{Y}$) from the structural model ($\mathbf{X} \rightarrow Z$). Approaches for investigating how features S influence Z are investigating item bias [28], Differential Item Functioning (DIF) [29] and measurement invariance [30]. For an extensive overview of the different approaches and their benefits and drawbacks, we refer to [30]–[33].

C. Proposed Method for Fair Inference on Latent Variables

We propose our framework for fair inference on outcomes which are measured only through error-prone proxies in a step-by-step manner. To clarify the framework and make it more comparable to earlier work, we use the running example of health risk score prediction from [7]. Their healthcare data set contains several clinical features $\mathbf{X}$ at time point $t-1$ (e.g., age, gender, care utilisation, biomarker values and comorbidities) which are used to predict healthcare cost Z at time t. In addition, the patient's race is the sensitive feature S, coded as $S=b$ for black patients and $S=w$ for white patients. The relations between these features are shown in panel A of Fig. 1.

Based on X, the expectation of a persons' healthcare cost is used as a risk score $\hat{Z} := E[Z \mid \mathbf{X}, S]$. The risk score is used to make a decision D to refer a patient to their primary care physician to consider program enrolment. More specifically $d=1$ if Z is above the 55^{th} percentile. In this setting, attributes $\mathbf{X}$ can be legitimately controlled. However, conditional on $\mathbf{X}$ both groups in S should have equal probability of being referred:

$$p(d = 1 \mid \mathbf{X} = x, S = b) = p(d = 1 \mid \mathbf{X} = x, S = w) \tag{7}$$

As mentioned in Section A and shown by [7], this procedure leads to bias in other proxies of Z, such as a patient's number of chronic conditions.

Our proposed framework is a SEM implementation of the second and third panels of Fig. 1. The general structure of the model is that of a Multiple Indicator, Multiple Causes (MIMIC) model.

In SEM, a latent variable (a hypothetical construct that is not directly observed) can be related to observable variables, such as indicators and causes of the latent variable, through sets of regression equations [34] and where parameters are typically estimated by means of maximum-likelihood [35]. A MIMIC model is a particular structure of a SEM model where a latent variable is simultaneously related to both observed indicator and cause variables [36]. In our model, the outcome variable Z (e.g., health) has multiple proxy indicators (e.g., chronic conditions, healthcare costs, hypertension), and the $\mathbf{X}$ features predict Z directly (thus the proxies only indirectly). A graphical representation of the MIMIC SEM model is shown in Fig. 2. This implementation imposes additional assumptions on the general causal

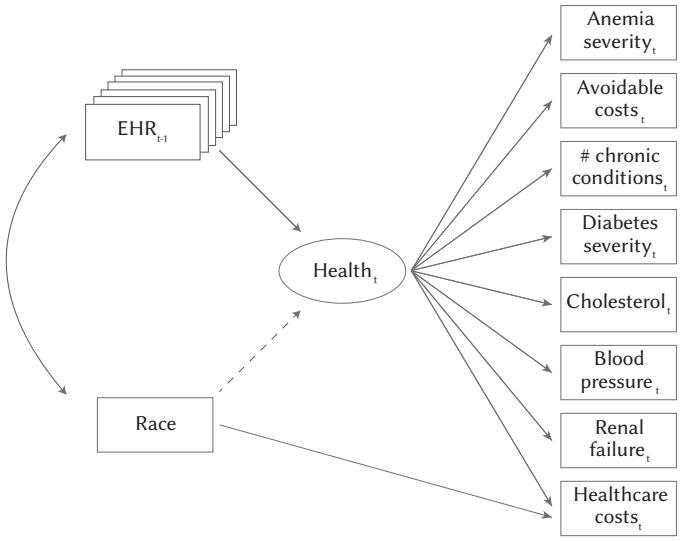

Fig. 2. Structural equation model for the proposed framework on the healthcare data set. For clarity, residual variances of the endogenous variables are not drawn in the diagram. EHR stands for Electronic Health Record. For more information on the variables used in the model, see [7].

graphs, most notably linear relationships between the variables and multivariate Gaussian residuals.

We implement our proposed correction procedure on the outcome variable Z in an existing fair inference approach [8] by means of the following steps:

1. The data-set is split in half to obtain a training set and a test set.
2. Regression parameters ($\mathbf{X}, S \longrightarrow Z$) are estimated on the training set using the MIMIC model.
3. The path from race to health is blocked by setting $S = b$ for all rows in the test set.
4. Predictions are generated for the adjusted test set by using the parameter estimates obtained in step 2.

To summarise, during estimation of the regression parameters ($\mathbf{X} \longrightarrow Z$), health is conditioned on race, but during prediction the path from race to health is blocked by setting $S = b$. Following the notation of [8], this yields a "fair world" distribution $p^*(Z, \mathbf{X})$. The expectation $\hat{Z} = E[Z \mid \mathbf{X}, S]$ is then computed from this distribution, meaning for two participants who differ only on S but not on $\mathbf{X}$, the risk score $\hat{Z}$ will be exactly the same. Because in SEM the latent outcome Z is modelled as a linear combination of the different proxies, the risk score is a reflection of the underlying health rather than only health cost.

V. Experiments

In this section, we evaluate the proposed framework on an application of the procedures discussed in this paper. We first prepare the data set as provided by [7] to create a basic risk score based on healthcare cost similar to the commercial risk score reported in their paper. Then, we illustrate our argument from Section A: we perform fair inference on the proxy measure for health (healthcare cost) to show that this does not solve the issue of unfairness in other proxy measures. This is a reproduction of the results shown by [7]. Next, we use the SEM framework from Section C to show how including a formal measurement model for Z – as in panel B of Fig. 1 – can largely solve the issue of unfairness in the proxies. Last, we show how existing differential item functioning (DIF) methods in the SEM framework – panel C of Fig. 1 – can aid in interpreting the extent to which proxy measures contain unfairness. Fully reproducible R code for this section is available as supplementary material to this paper at the following DOI: 10.5281/zenodo.3708150.

A. Data Preparation and Feature Selection

Log-transformations are applied to highly skewed variables at time-point t, such as costs, to meet the assumption of normally distributed residuals in regression procedures. As an additional normalisation step, the predictors at time-point $t-1$ are re-scaled to homogenise their levels of variance. The data set is then split into a training and a test set. In this section, estimation is always done on the training set and inference is done on the test set.

To simplify our proposed framework for the purpose of this application, we select a subset of features at time-point $t-1$ for prediction of the target of interest at time point t, health. We want our procedure to be comparable to the commercial algorithm which produces the risk scores described in [7]. If the features we select are the same features used by the commercial algorithm, then our procedure would yield very similar results upon generating a risk score. Unfortunately, the predicted risk scores used by [7] cannot be replicated exactly using the provided data set.

To select the subset of predictor features for further use in our procedure, we performed a LASSO regression [37] where all available features at time-point $t-1$ are used as predictor variables, and the provided algorithmic risk score at time-point t is used as a target. Following the guidelines by [38], we used cross-validation to select the optimal λ penalty value. This yields a set of non-zero predictors which predict the algorithmic risk score well.

Superman's rank correlation between the commercial and the replicated risk score is high $\rho = .82$, indicating that the commercial and replicated risk scores perform similarly in the rank-based cutoff applied in [7]. The predictors selected in this model are used as predictors X in the structural equation models of the following sections.

B. Fair Inference on Cost as a Proxy of Health

Pane A of Fig. 1 illustrates conditional statistical parity as defined by [6]. To perform standard statistical parity correction, the outcome Z is conditioned on sensitive feature S when estimating the coefficients of the prediction model ($X \longrightarrow Z$), and during prediction all subjects are assumed to have the same level of S, e.g., $S = b$, such that

$$p\,(Z = z \mid \mathbf{X} = x, S = b) = p\,(Z = z \mid \mathbf{X} = x, S = w) \tag{8}$$

However, in the current situation we do not measure Z directly, but only a proxy $Y \in \mathbf{Y}$. Standard parity correction for this proxy does not necessarily mean the parity is achieved for other proxies [7]. The

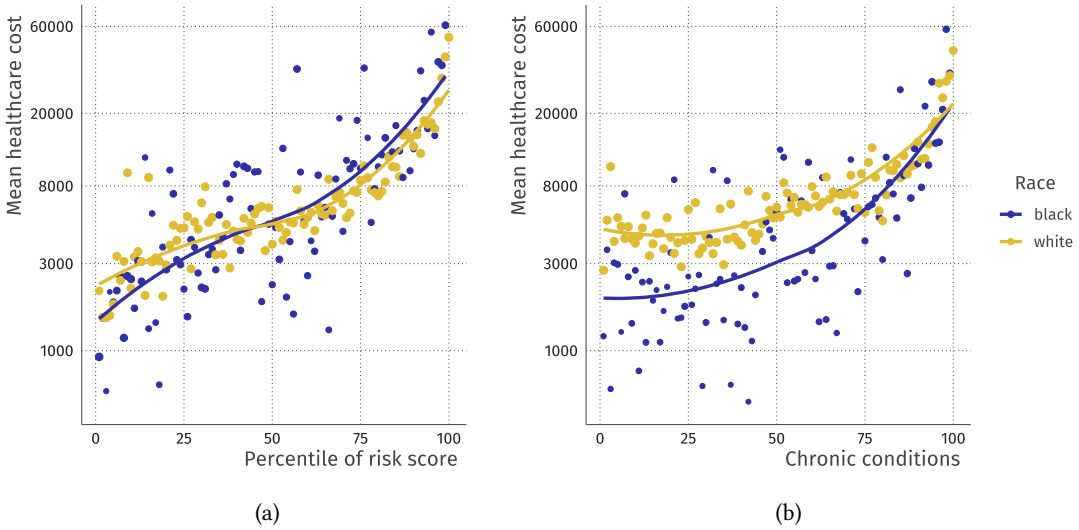

Fig. 3. Although the risk score displays statistical parity on healthcare costs (no differences between the lines in panel A), these costs conditional on health (as measured by chronic illness) depends on race (panel B). This causes statistical disparity for the risk score on the level of health (Fig. 4, panel B). Figure replicated from [7].

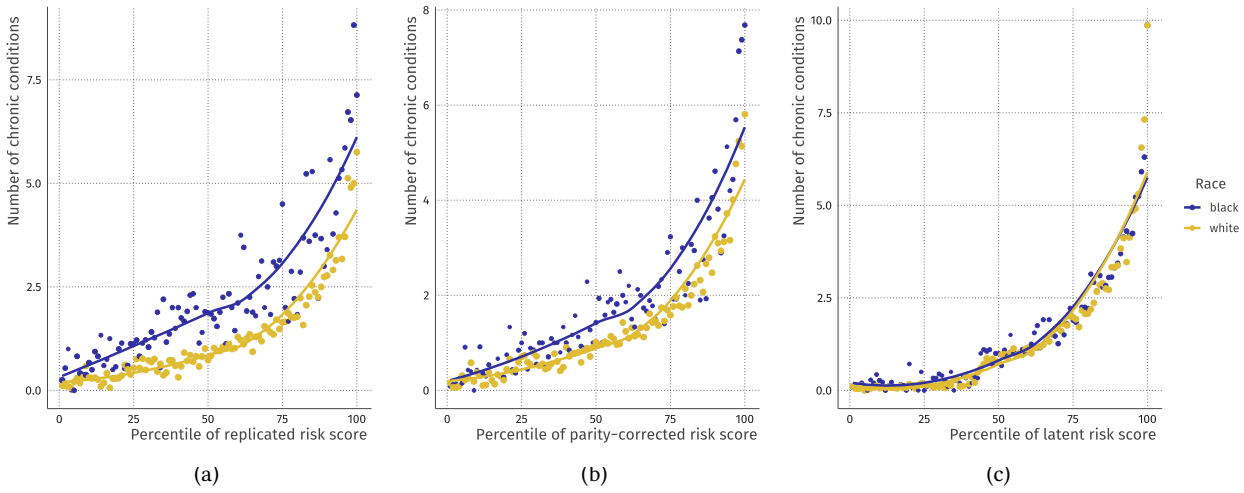

Fig. 4. Effect of including a measurement model in constructing risk scores. The first panel shows the uncorrected risk score based on healthcare cost, the middle panel shows the same risk score but corrected for the sensitive feature, and the third panel shows the corrected risk score based on the latent health outcome using a measurement model.

reason for this is explained in Fig. 3. Pane A illustrates that statistical parity is present when plotting the risk score against healthcare costs, meaning that for a given risk score, the healthcare costs for both races are approximately equal. However, Pane B illustrates that when the number of chronic conditions are plotted against healthcare costs, there are differences between the two race groups, meaning that for a given amount of chronic conditions, white patients cost more than black patients.

As a result, standard statistical parity correction on healthcare cost does not remove the disparity in chronic conditions. This becomes visible when comparing Pane B of Fig. 3 with Pane A of Fig. 4. In addition, from Pane B of Fig. 4 it can be seen that the results improve compared to not including race at all (Pane A of Fig. 4), yet race differences remain for the chronic conditions proxy. As a consequence, individuals belonging to $S=b$ will still have a lower health status when being selected for intervention.

C. Fair Inference on Latent Health

A cause for the fact that conditional statistical parity is not met when following Pane A of Fig. 1 can be that $\hat{Z}$ is a (bad) proxy. Instead of using one bad proxy, it is better to use multiple (bad) proxies as indicators of an unobserved latent variable measuring 'true health'. How such a model can be specified is illustrated in Pane B of Fig. 1. Such a model can be applied in practice by following the steps in the framework described in Section C. Similarly to [6], the sensitive feature is excluded during prediction.

Fig. 4 shows the effect of including a measurement model in constructing risk scores. The figure illustrates that using a measurement model with multiple imperfect measurements of health as indicators for 'true health' substantially improves conditional statistical parity, when compared to either the uncorrected risk score on a proxy, or a parity-corrected risk score on the proxy. Additionally, Table I shows a numerical summary which corroborates this finding. Here, we created a prediction model for the number of chronic conditions using both risk score and race. The parameter for race then indicates whether a race difference exists for health, conditional on the risk score. This conditional dependence becomes close to 0 when using the latent risk score (95% CI = [´0.113, 0.012]). Thus, by using this measurement model, the problem that individuals belonging to $S=b$ had a lower health status when being selected for intervention is minimised.

TABLE I. Estimated Conditional Parity on the Number of Chronic Conditions for Different Risk Scores. β Parameters Are Linear Regression Parameters, Indicating the Deviation of White Patients From Black Patients in the Number of Chronic Conditions, Conditional on Risk Score. For Example, a Value of -0.963 Means that White Patients Have on Average a 0.963 Fewer Chronic Conditions for the Same Risk Score

Risk score	β	2.5%	97.5%
Replicated	-0.963	-1.063	-0.864
Parity-corrected	-0.577	-0.677	-0.478
Latent	-0.051	-0.113	0.012

D. Investigating Unfairness in Proxies

When using a measurement model with multiple imperfect measurements of health as indicators of 'true health', differences in measurement error over the different groups of the sensitive feature can still be present. Panel C of Fig. 1 illustrates how differences over the sensitive feature groups in the error prone indicator variables can be incorporated directly when estimating 'true health'. For example, differences in measurement error of healthcare cost can be present for the different groups of race.

Including a DIF parameter δ on the healthcare cost variable yields a model which fits significantly better on the test set than the model without the DIF parameter ($\chi^2(1) = 50,\ p < 0.001$). The value of the DIF parameter on cost is estimated as $\delta = 0.198$ (95% CI = [0.172, 0.225]). This means that for the same level of health, the log-healthcare costs of the white race class in this data set is estimated to be 0.198 higher. This means that the cost of healthcare for white patients is $(e^{0.198} - 1) \cdot 100\%$ = 21.9% higher than that for black patients, given an equal level of health as measured by the measurement model (95% CI = [18.7, 25.2]).

Applying the same procedure to the other indicators leads to estimates of DIF for those indicators. The results are shown in Table II. This table shows that some proxies have stronger DIF than others, meaning some proxies are more unfair than other proxies. Notable, the avoidable healthcare cost and the renal failure items have low levels of DIF for race, whereas the healthcare cost and the number of active chronic conditions have strong DIF.

TABLE II. Estimated Differential Item Functioning Parameters for Each Indicator (Proxy) of Health. δ Parameters Should Be Interpreted as the Mean Deviation of the Black Patients Compared to the White Patients Given Health.

Indicator	δ	2.5%	97.5%
No. active chronic conditions	0.453	0.364	0.541
Mean blood pressure	-0.262	-0.320	-0.204
Diabetes severity (HbA1c)	-0.343	-0.391	-0.296
Anemia severity (hematocrit)	0.250	0.231	0.268
Renal failure (creatinine)	-0.019	-0.025	-0.014
Cholesterol (mean LDL)	-0.235	-0.317	-0.153
Healthcare cost (log)	0.198	0.172	0.225
Avoidable healthcare cost (log)	-0.052	-0.096	-0.008

VI. Conclusion

In this paper, we have argued that when measurement error is at play, performing fair inference on a proxy measure of the outcome is insufficient to achieve a fair inference on the true outcome. This manifests itself, as shown in [7], as unfairness in other proxy measures of the outcome of interest. Alternatively, in this study we proposed to make use of existing measurement models containing multiple error-prone proxies for the outcome of interest. In addition, fair inference can be accounted for in each of these proxies simultaneously if needed by allowing for measurement error in proxies to differ over groups defined by differing values of a sensitive feature. We provided a framework to perform these estimations and applied this framework to the exemplary data set provided by [7]. Here, it was concluded that fair inference was accounted for when multiple proxies were used in a measurement model instead of a single proxy. Additionally accounting for differences in measurement error over race groups was not needed to further improve fairness in predicted risk scores, although substantive group differences were found for some proxies.

References

[1] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, p. 0049124118782533, 2018.

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

[3] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[4] J. Kleinberg, S. Mullainathan, M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[5] M. J. Kusner, J. Loftus, C. Russell, R. Silva, "Counterfactual Fairness," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett Eds., Curran Associates, Inc., 2017, pp. 4066–4076.

[6] S. Verma, J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, FairWare '18, Gothenburg, Sweden, May 2018, pp. 1–7, As-sociation for Computing Machinery.

[7] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.

[8] R. Nabi, I. Shpitser, "Fair Inference on Outcomes," in T*hirty-Second AAAI Conference on Artificial Intelligence*, Apr. 2018.

[9] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[10] A. Z. Jacobs, H. Wallach, "Measurement and fairness," *arXiv preprint arXiv:1912.05511*, 2019.

[11] A. P. Dawid, A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[12] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, L. Moy, "Learning from crowds.," *Journal of Machine Learning Research*, vol. 11, no. 4, 2010.

[13] D. Borsboom, "When does measurement invariance matter?," *Medical care*, vol. 44, no. 11, pp. S176–S181, 2006.

[14] J. Pearl, *Causality models, reasoning, and inference*. Cambridge: Cambridge University Press, 2013. OCLC: 956314447.

[15] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[16] T. B. Brakenhoff, M. Mitroiu, R. H. Keogh, K. G. Moons, R. H. Groenwold, M. van Smeden, "Measurement error is often neglected in medical literature: a systematic review," *Journal of clinical epidemiology*, vol. 98, pp. 89–97, 2018.

[17] D. Borsboom, "Latent variable theory," *Measurement: Interdisciplinary Research and Perspectives*, vol. 6, no. 1-2, pp. 25–53, 2008, doi: 10.1080/15366360802035497.

[18] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Advances in Neural Information Processing Systems*, 2017, pp. 656–666.

[19] W. A. Fuller, *Measurement error models*, vol. 305. John Wiley & Sons, 2009.

[20] H. M. Blalock, A. B. Blalock, "Methodology in social research," 1968.

[21] A. L. McCutcheon, *Latent class analysis*. No. 64, Sage, 1987.

[22] G. Rasch, *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

[23] G. J. McLachlan, K. E. Basford, *Mixture models: Inference and applications to clustering*, vol. 38. M. Dekker New York, 1988.

[24] F. M. Lord, *Applications of item response theory to practical testing problems.* Routledge, 2012.

[25] K. A. Bollen, *Structural equations with latent variables.* Wiley series in probability and mathematical statistics Applied probability and statistics, New York, NY Chichester Brisbane Toronto Singapore: Wiley, 1989. OCLC: 18834634.

[26] A. Skrondal, S. Rabe-Hesketh, *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* CRC Press, 2004.

[27] K. G. Jöreskog, "Testing structural equation models," *Sage focus editions*, vol. 154, pp. 294–294, 1993.

[28] G. J. Mellenbergh, "Item bias and item response theory," *International Journal of Educational Research*, vol. 13, pp. 127–143, Jan. 1989, doi: 10.1016/0883-0355(89)90002-5.

[29] P. W. Holland, H. Wainer, *Differential Item Functioning*. New York: Routledge, 1993.

[30] N. Schmitt, G. Kuljanin, "Measurement invariance: Review of practice and implications," *Human resource management review*, vol. 18, no. 4, pp. 210–222, 2008.

[31] P. Flore, "Stereotype threat and differential item functioning: A critical assessment," 2018.

[32] J.-B. E. Steenkamp, H. Baumgartner, "Assessing measure-ment invariance in cross-national consumer research," *Journal of consumer research*, vol. 25, no. 1, pp. 78–90, 1998.

[33] R. J. Vandenberg, C. E. Lance, "A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research," *Organizational research methods*, vol. 3, no. 1, pp. 4–70, 2000.

[34] B. M. Byrne, *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. routledge, 2013.

[35] Y. Rosseel, "Lavaan: An r package for structural equation modeling and more. version 0.5–12 (beta)," *Journal of statistical software*, vol. 48, no. 2, pp. 1–36, 2012.

[36] K. G. Jöreskog, A. S. Goldberger, "Estimation of a model with multiple indicators and multiple causes of a single latent variable," *Journal of the American Statistical Association*, vol. 70, no. 351a, pp. 631–639, 1975.

[37] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.

[38] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media, 2009.

Laura Boeschoten

Laura Boeschoten is a postdoctoral researcher at the Methodology and Statistics department of Utrecht University. She is particularly interested in measurement error, missing data and latent variable models. Her PhD was on a joint project between Tilburg University and Statistics Netherlands where she developed a new methodology that estimates and corrects for measurement error in combined survey-register data-sets (Tilburg University, The Netherlands, 2019) and she currently continues this line of research by further investigating how this methodology can be implemented in the production if various official statistics. Other lines of research focus on developing an infrastructure that enables the use of so-called 'Data Download Packages' for scientific research and on investigating the validity and reliability of measurements obtained from these 'Data Download Packages'.

Erik-Jan van Kesteren

Erik-Jan van Kesteren is an assistant professor of human data science at Utrecht University, the Netherlands. His educational background is mainly in social science and statistics. In between studies he worked in a data management team at a large company in the Netherlands. His PhD research was on extending latent variable models for modern data problems, with a focus on regularization, optimization, and software (Utrecht University, The Netherlands, 2021). Currently, Erik-Jan works with a small social data science team at ODISSEI (odissei-data.nl), helping social scientists with their computational and data problems.

Ayoub Bagheri

Ayoub Bagheri is an assistant professor in applied data science at the Methodology and Statistics department of Utrecht University. The focus of his academic career has been to develop intelligent systems for improving health, education, and social sciences by mining big data, especially text data. His current research in terests include machine learning, text mining, and natural language processing. As part of the Human Data Science group, he works on several projects in the domain of applied data science for health and social sciences. He is also part of the organization team of the special interest group Text Mining of the focus area applied data science in Utrecht University.

Daniel Oberski

Daniel Oberski holds a joint appointment as associate professor of data science methodology at Utrecht University, department of Methodology Statistics, and at the University Medical Center Utrecht (UMCU), department of Biostatistics. His work focuses on latent variable modeling and data science applications in the social, behavioral, and biomedical sciences. He leads the social data science team at the national research infrastructure for the social sciences in the Netherlands, ODISSEI. He is also lead data scientist of UMCU's "digital health" program, which works to implement data science in clinical care at the hospital.