Measurement Modeling of Predictors and Outcomes in Algorithmic Fairness

Elisabeth Kraus[1], Christoph Kern[2,3]

[1]*Department of Psychology, LMU Munich, Akademiestr. 7, 80799 München, Germany*
[2]*Institute of Statistics, LMU Munich, Ludwigstr. 33, 80809 München, Germany*
[3]*Munich Center for Machine Learning (MCML), Oettingenstraße 67, 80538 München, Germany*

Abstract

This contribution investigates structural equation modeling (SEM) as a pre-processing approach to mitigate measurement bias in algorithmic decision-making systems. We construct latent predictors and latent targets based on different measurement modeling strategies and evaluate their interplay in simulations and an application study. We systematically compare SEMs which preserve group-differences (group-overarching) to models which equalize group-differences (group-specific) in predictors and outcomes. In our simulations, we find that group-overarching models are a more effective strategy than group-specific models and lead to smaller subgroup prediction error and better calibrated risk scores. In the application study we apply SEM to a health risk prediction task and find support for the benefit of group-overarching models. We conclude that tackling fairness concerns by utilizing measurement models of both the predictors and the outcome can contribute to the fairness of ADM systems. Utilizing SEM during pre-processing allows to incorporate substantive knowledge about the prediction task into the model implementation.

Keywords

structural equation modeling, measurement models, bias mitigation

1. Introduction

The reliance on machine learning (ML) and prediction algorithms for decision-making, known as algorithmic decision-making (ADM), is becoming increasingly prevalent. Examples of such systems abound, from loan approval processes in finance [1] to profiling of the unemployed in the labor market context [2], and even risk assessment in the criminal justice system [3]. The AI Watch report [4] presents a total of 686 use cases in the public sector in Europe, many of which placed in domains with profound impacts on life chances. While these systems hold promise for more accurate and objective decision-making when appropriately designed, they can have adverse effects due to biased training data and inadequate model specification. Numerous examples exist of ADM systems exhibiting discriminatory behaviors [5, 6, 7].

Measurement bias has been identified as a key source of algorithmic unfairness [8, 9]. This type of bias can arise due to the use of biased proxy variables in the prediction models' specification step. A prominent example studied by Obermeyer et al. [10] is the use of healthcare costs as

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany

e.kraus@psy.lmu.de (E. Kraus); christoph.kern@stat.uni-muenchen.de (C. Kern)
0000-0001-7363-4299 (C. Kern)

CEUR Workshop Proceedings (CEUR-WS.org)

a proxy for actual health needs in the context of a risk prediction model deployed by health insurance companies in the U.S. As healthcare costs systematically differ by race, social biases can sneak into the model due to the use of inadequate proxy variables. Careful model specification and valid measurement thus is critical to mitigate potential fairness issues of risk prediction models downstream.

To address this need, we study the use of measurement models – implemented via structural equation modeling (SEM) – to investigate the impact of different pre-processing strategies of predictors and targets on a models' fairness outcomes. Drawing inspiration from the field of psychometrics [11, 12], we apply measurement modeling in the context of machine learning, focusing on its impact on algorithmic fairness. Despite calls for more rigorous operationalizations of latent constructs in prediction models [13], measurement modeling techniques are rarely studied in fair ML contexts. As a significant exception, Boeschoten et al. [14] draw on the health prediction example of [10] and show how the use of measurement models can mitigate biases in the targets of prediction models. We aim to extend these efforts and systematically test the ability of different SEM specifications to mitigate unfairness.

Our study contributes to the fair ML literature by expanding on the limited body of work on measurement modeling for algorithmic fairness. Specifically, we compare SEMs which preserve group-differences (group-overarching models) to models which equalize group-differences (group-specific models) both in predictors and targets. We additionally study the fairness implications of different operationalizations of latent constructs by comparing SEMs which use different indicators in specifying the measurement relationships.

We study the use of SEM as a pre-processing technique by deriving latent scores based on the specified measurement models. Latent scores can be constructed for predictors and/or targets and can then be used flexibly in any type of prediction modeling setup. However, different measurement model specifications result in different sets of latent scores. Understanding how such different pre-processing strategies interact with fairness is critical for effectively mitigating unintended outcomes. A systematic mapping of different measurement model specifications to fairness outcomes can help guide model design, as different approaches may achieve similar accuracy but differ in their fairness properties. Our focus on mitigating biases via measurement modeling is especially relevant whenever multiple potential proxy variables for the predictors and/or target are available and substantive knowledge about the meaning of these predictor and/or target variables and their measurement relationships can be derived.

In the following sections, we provide a brief overview of measurement modeling from a social scientific perspective (section 2) and outline structural equation modeling specifications to address biases in the development of machine learning systems (section 3). We report insights gained by this approach through simulations (section 4) and a case study (section 5) involving the prediction of individual health, following the use case of [10]. We conclude by discussing potentials and pitfalls of the use of SEM techniques for bias mitigation (section 6).

Related Work There is an extensive body of research on pre-processing techniques for algorithmic fairness, but few methods incorporate substantive knowledge about the variables used to build the prediction model. Data-driven procedures include suppression, massaging, reweighing, or resampling [15]. Other studies, like [16] propose to transform data on the basis of a distortion model, based on the conception that discrimination stems from past differential treatment. Another popular approach to fairness pre-processing is Learning Fair Representations [17]. In this method, the idea is to map the covariate space to prototypes that are independent of the sensitive attribute and to then use the prototypes to predict outcomes.

Another line of work particularly addresses label bias, i.e. proxy variables with systematic measurement error that are used as prediction outcomes. Potential mitigation strategies in this context range from careful model specification [18] and sensitivity analysis [19] to adapted estimation procedures [20]. [21] propose a framework to study and mitigate label bias based on counterfactual reasoning.

However, there is limited work on incorporating latent constructs and measurement relationships in the context of fair predictive modeling [13, 22]. Our study is motivated by the use of structural equation modeling for a prediction models' target variable as proposed by [14]. As identified by [14], there can be non-negligible measurement error in the outcome with respect to sensitive group membership. They show that achieving fair inference on one single proxy measure of the outcome is insufficient when measurement error is present, as it may result in unfairness with respect to other proxy measures of the outcome. Instead, their study proposes utilizing measurement models containing multiple (error-prone) proxies for the outcome, allowing for fair inference in each proxy simultaneously by accommodating measurement error differences across groups defined by sensitive attributes. Their study, however, focuses on measurement modeling of the outcome variable only and does not systematically compare different SEM specifications and their fairness implications.

2. Background

2.1. By-Proxy Indicators in ADM Models

When ADM systems are built based on risk predictions and models of human behavior, the predictors and targets used often represent measurements of social indicators, such as job tenure, the number of chronic diseases or place of residence and living conditions. Yet, these variables are seldom used for what they appear to measure, but are seen as indicators for underlying constructs that cannot be measured directly. For instance, job tenure may be used as an indicator for trustworthiness, the number of chronic diseases as an indicator for current health and insurance policies are optimized by using postal code as an indicator for the socio-economic living environment.

However, it is often overlooked that such by-proxy variables can be better indicators of the underlying latent concept for one demographic group than for another. For example, car insurances tend to be less expensive for tenured individuals, assuming that reliability differences between tenured and non-tenured individuals might be true on average [23]. But if Black people are less likely to be tenured [24], then they are less likely to get the reliability due to tenure bonus in the calculation, even though there is no reason to assume that Black people are in general less reliable than white people. As a result, job tenure can be an unfair proxy because it is more closely related to the underlying idea of reliability for white individuals than for Black individuals.

We suggest the use of structural equation modeling (SEM) to study and possibly mitigate the impact of unfair predictors and targets in ADM contexts. SEM methodology was initially proposed in psychology and is heavily used in the social sciences. The problem of single unfair indicators has quite some history and is discussed in the field of psychometrics.

2.2. Unfairness in Measurement

In psychometrics, unfair indicators are typically identified because they are used in combination with other similar indicators in multi-item scales [25]. These scales are then analyzed with psychometric models such as structural equation models (SEMs) [26]. Structural equation models use a set of values of observable indicators to infer values of unobservable constructs, so called latent variables [11, 12]. The idea is to extract the common variance between indicators and attribute it to the underlying latent variable. In most SEMs the contribution strength (called loading) may vary between the indicators. The higher a loading is estimated, the better an indicator is in measuring the underlying construct and the more of its variance is shared with the other indicators.

The loadings can vary between indicators but also between groups of individuals. Group-specific loadings can lead to one indicator having a high loading in the first group, but a low loading in the second group. This expresses that the indicator is a good indicator for one group, but a worse indicator for another group. This analysis is called multi-group-SEM [27] and fulfills the need to identify diverging indicator quality. In multi-group SEMs, the final latent scores for individuals can be computed based on group-specific measurement models. In estimating latent scores, the indicator values are combined by a weighted sum of the indicator values. Thereby, weak indicators with low loadings are down-weighted in the calculation. In consequence, group differences between latent scores based on group-specific measurement models are reduced or even eliminated, compared to group differences between latent scores estimated from a single SEM model.

Whenever the latent scores are to be used as a decision criterion, group-specific model scores are recommended [28]. Only with group-specific model scores, all groups have the same distribution of latent scores and no group has a higher average. While this issue is well studied in the context of diagnostic decision-making [29], the question arises how group-specific latent scores compare to group-overarching (single group) latent scores in the context of fair prediction modeling.

3. Analytical Strategy

3.1. Multi-Group Measurement Models

In our exploration of SEM for algorithmic fairness, we intend to group sets of variables that measure the same unobservable, latent construct. We do this based on theoretical considerations with respect to the substantive meaning of these variables. The group of variables that is assumed to measure the same latent construct can be modeled to estimate latent variable scores, which in turn then substitute the original set of variables in the prediction model. In the factor analytic model for continuous predictor variables [27], the measurement relationships may be represented as

$$\mathbf{x} = \tau^g + \lambda^g \xi^g + \delta^g$$

where $\mathbf{x}$ is the vector of predictor variables measuring the latent construct ξ, τ^g a vector of intercepts, λ^g a matrix of factor loadings, δ^g a vector of error variables, and groups g which may be defined by (sensitive) demographic attributes. A corresponding measurement model can be defined for continuous target variables

$$\mathbf{y} = \tau^g + \lambda^g \eta^g + \epsilon^g$$

where $\mathbf{y}$ is the vector of target variables measuring the latent construct η and ϵ^g a vector of error variables.

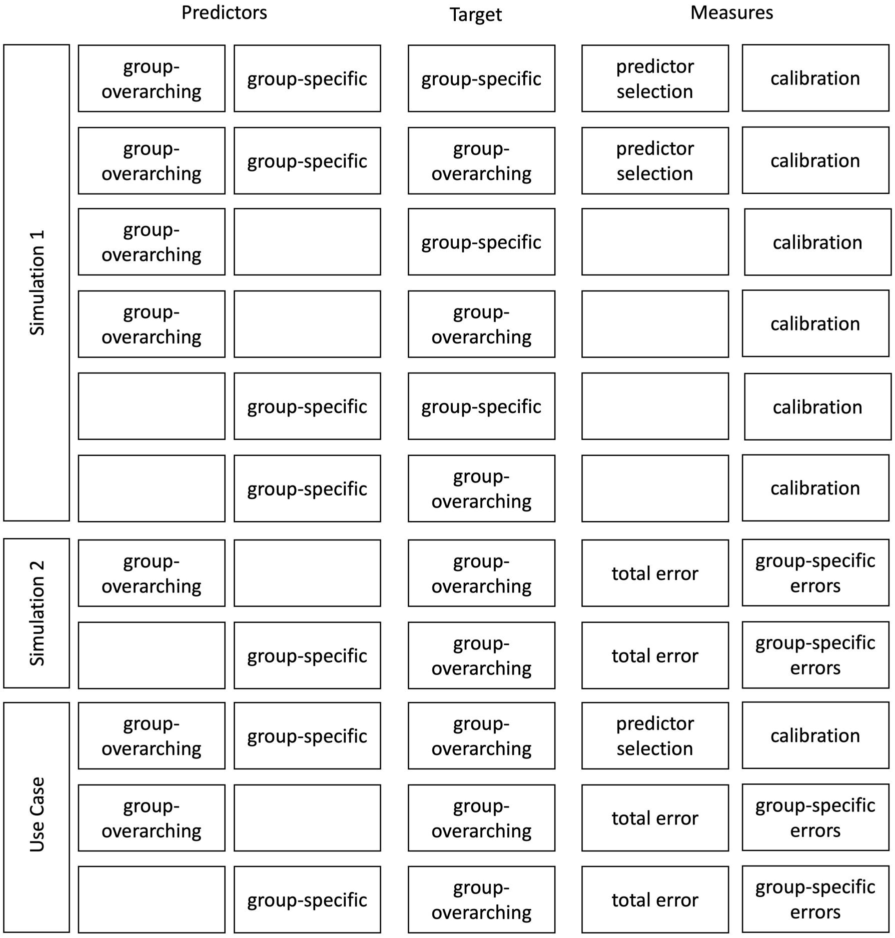
In both cases, the measurement model may be specified to map indicator values on the same scale by setting $\tau^g = \tau^{g'}$, $\lambda^g = \lambda^{g'}$ and $\delta^g = \delta^{g'}$ ($\epsilon^g = \epsilon^{g'}$) (group-overarching model). In contrast, group-specific parameters τ^g, λ^g and δ^g (ϵ^g) may be estimated to map indicator values on different scales for different groups (group-specific models). In the first case, group differences are preserved, whereas with group-specific models any initial group-level differences are equalized in the resulting latent variable scores. The decision on whether group-specific scales should be preferred is highly context-specific and, given its considerable fairness implications (section 4), should not only be based on SEM fit measures [30].

Measurement models can be similarly defined for categorical predictor and target variables by assuming the existence of a continuous response variable underlying each observed categorical variable and the specification of threshold relations linking both types of variables [12]. With continuous predictors or targets, the measurement models' parameters are typically estimated via Maximum Likelihood estimation [31], whereas with categorical variables Weighted Least Squares (WLS) estimation may be used [12]. While sample size requirements depend on the exact model specification [32], it is important to note that it needs to be carefully assessed whether estimating (multi-group) SEMs in contexts with small subgroups is a viable strategy [33].

3.2. Use Case

We study different measurement modeling strategies in the context of risk algorithms to guide health decisions [10]. In this setting, multiple health indicators such as the number of chronic diseases, blood sugar levels and high blood pressure are used to predict future health outcomes. In practice, the resulting risk score may be used as a decision criterion for assigning individuals to health support plans. It has been shown that the commercial approach to defining such risk scores led to a discrimination against Black patients [10].

We build on the work of Boeschoten et al. [14] and treat discrimination in health risk prediction as a measurement problem. The target variable, health status, may be measured using different types of indicators: based on both health and cost indicators or based on health indicators only. While health indicators such as blood sugar levels indicate a qualitative dimension of health, cost indicators such as costs for primary care may represent a quantitative dimension. However, using health cost as a prediction target can lead to unfair outcomes for Black patients as shown by [10]. While [14] construct a combined health-cost target with SEM,



	Predictors		Target	Measures	
Simulation 1	group-overarching	group-specific	group-specific	predictor selection	calibration
	group-overarching	group-specific	group-overarching	predictor selection	calibration
	group-overarching		group-specific		calibration
	group-overarching		group-overarching		calibration
		group-specific	group-specific		calibration
		group-specific	group-overarching		calibration
Simulation 2	group-overarching		group-overarching	total error	group-specific errors
		group-specific	group-overarching	total error	group-specific errors
Use Case	group-overarching	group-specific	group-overarching	predictor selection	calibration
	group-overarching		group-overarching	total error	group-specific errors
		group-specific	group-overarching	total error	group-specific errors

Figure 1: Our study design with different types of latent predictors, targets and evaluation measures used in two simulations and an application study.

we compare how the use of different indicators and measurement modeling strategies affect fairness outcomes.

We do not only focus on measurement error in the target variable but also in the predictor space. We investigate group-specific and group-overarching predictors when predicting group-specific and group-overarching targets. Specifically, we investigate if latent variable scores derived from single or multiple-group SEMs estimated during pre-processing have different effects on algorithmic fairness.

We conduct simulations and an application study motivated by our use case. Figure 1 presents our research design. The simulations focus on the interactions between group-overarching and group-specific predictors and both types of targets. In the application study, we focus only on the implications of different predictor types, given the same (group-overarching) target.

4. Simulation Study

In the simulation study we use SEM to model the target variable, health status, and one of the predictor variables, health costs. We simulate group differences in costs and estimate two sets of targets. The first set includes a group-overarching target, which represents latent health using health indicators and one cost indicator and a single measurement model for both race groups. Furthermore, we construct a group-specific target, which measures health based on health and cost indicators with a multi-group SEM. Both targets draw on health costs to investigate if the use of SEMs can circumvent fairness issues even when the measurement model incorporates problematic measurement paths. The second set of target variables draws only on health indicators to construct group-overarching and group-specific targets. Next to the latent targets, we construct group-overarching and group-specific measurement models for the predictor latent costs.

We conduct two simulations (see Figure 1). In *simulation 1*, we manipulate both the type of the latent target and of the latent predictor used in the prediction model and evaluate the calibration of the resulting risk scores. In *simulation 2*, we investigate overall mean squared errors (MSE) and group-specific error when predicting a group-overarching target and offering different versions of the latent predictors (group-overarching vs. group-specific).

4.1. Methods

4.1.1. Data Setup

We generate data for a model that includes latent health as the target variable and three health indicators and one latent cost variable as the predictor variables. The data generation setup is presented in Figure 2. We follow a consecutive strategy in simulating the dataset. First, we create standard normal variables to simulate the three health indicators (hi1-hi3) and the latent cost variable (costs). The latent health target variable (health) is then derived by a linear combination of health indicators, latent costs, and a residual. The residuals follows a normal distribution with $\mu = 0$ and $\sigma = 0.2$. Health indicators and latent costs are each weighted by 0.5, resulting in the linear generation equation: $health = 0.5 \cdot h1 + 0.5 \cdot h2 + 0.5 \cdot h3 + costs + res$. To produce the latency of the latent costs and latent health variables, indicator variables are generated by measurement equations $cost_1 = a_1 + \lambda \cdot costs + e_1, \ldots, cost_4 = a_4 + \lambda \cdot costs + e_4$ and $hi_4 = a_4 + \lambda \cdot health + e_4, \ldots, hi_6 = a_6 + \lambda \cdot health + e_6$, with $a \sim N(0, 0.5)$, $\lambda = 1$, and $e \sim N(0, 0.2)$. Four indicators are simulated for the latent cost variable, and three indicators are simulated for the latent health variable.

We then add a group variable (race), which is associated with latent costs and thus introduces group differences in the simulated structural relationships. We specify a logistic relationship and vary how strongly groups and costs are related by setting different values for our group-difference parameter α and the group proportions by different values for our proportion parameter h: $p(race = 1) = exp(h + \alpha \cdot costs)/(1 + exp(h + \alpha \cdot costs))$. The α-values range between 0 and 3, with a step size of 0.05, resulting in 61 different values. For the proportion of the disadvantaged group, h is set to be 0 for equally sized groups and 1.5 for a proportion of around 15% of the disadvantaged group.

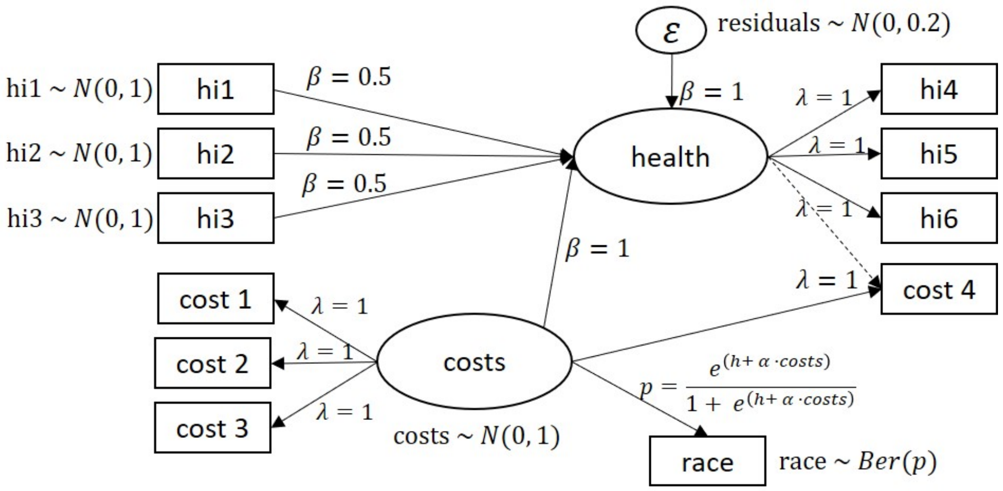


Figure 2: Data simulation setup in the simulation study.

The size of each simulated dataset is set to $n = 1000$. To use the dataset for the prediction task, we sort the health predictors (hi1-hi3) and the cost indicators (cost1-cost3) to the predictor side and the latent health indicators (hi4-hi6) to the the target side.

4.1.2. Estimation and Prediction

To generate different prediction setups, we then estimate latent variable scores for the target construct in two ways. For the group-overarching target, we estimate latent scores for health based on the three target-health indicators (hi4-hi6) and one cost indicator (cost4) with equal loadings for both race groups ($\lambda^{g1} = \lambda^{g2}$). For the group-specific target, we estimate latent scores based on the same indicators while allowing for group-specific parameters in the measurement model by race ($\lambda^{g1} \neq \lambda^{g2}$). We construct an additional set of latent scores by estimating group-overarching and group-specific models which draw on the three target-health indicators (hi4-hi6) only.

We also estimate two versions of the predictor latent health costs based on the cost indicators (cost1-cost3). For the first version, we map all observations on the same scale using the group-overarching model with equal loading parameters for both race groups, preserving group differences in costs ($\lambda^{g1} = \lambda^{g2}$). In the second approach, we map observations on group-specific scales for latent costs by allowing for group-specific parameters ($\lambda^{g1} \neq \lambda^{g2}$). By scaling each group separately, group differences in the resulting latent variable scores are lost.

Based on the simulated data, we eventually estimate different risk scores by predicting both targets with LASSO regression [34]. The predictor set comprises the health indicators h1-h3 and the estimated latent scores for costs. Latent health represents the target variable. Cross-validation is used to determine the optimal regularization parameter λ. The individual predictions of the LASSO model form the risk scores.

4.1.3. Evaluation

In simulation 1, we assess the calibration of the predicted risk scores with respect to the first health indicator across different combinations of latent targets and predictors. Specifically, we assess miscalibration by race by plotting the derived score against the values of single health indicators separately for racial groups [14, 10]. Congruent lines indicate fairness of the risk scores with respect to the health indicator. A downwards shift in the plotted relationships for positive health indicators indicates an underestimation of the risk for that group. We further evaluate how often group-overarching and group-specific cost predictors are chosen during variable selection in the LASSO regressions. We manipulate the group differences in costs (α-values) and hold the group size constant ($h = 0$), which results in a total of 61 datasets for the first simulation.

In simulation 2, we focus on the group-overarching target and evaluate how different predictor versions affect model performance by race. We evaluate the group-specific mean squared errors (MSE) in comparison to the overall MSE in different conditions. There is a total of 61 (α-values) x 2 (group proportions h) x 2 (predictor versions; only group specific, only group-overarching) = 244 datasets for the second simulation.

4.1.4. Software

All analyses are performed using R, version 4.3.2. [35]. For SEM analyses we use the package lavaan [36] and the package glmnet [37] is used for LASSO regression.

4.2. Results

In simulation 1, we first compare the effects of using different types of latent predictors and targets on calibration fairness. Figure 3 shows that unfairness in terms of group-specific miscalibration increases with group differences in health costs as varied by the simulation parameter α. However, the degree as to which this miscalibration can be mitigated critically depends on the measurement modeling strategy that is used for both predictors and target. Models with group-specific predictors result in unfair risk scores, whereas group-overarching predictors, particularly in combination with the group-overarching target (first column in Figure 3), largely reduce racial miscalibration. This pattern remains unchanged when the latent variable model of the target variable does not include cost as a target indicator (Figure 7 in Appendix A).

We additionally observe that group-overarching predictors are chosen more frequently in the LASSO models when predicting group-overarching targets and that group-specific predictors are chosen more often when predicting the group-specific target (see Table 1). Overall, the group-specific predictor is chosen less often compared to the group-overarching predictor.

In simulation 2, we focus on the group-overarching target variable, which yields the best results in terms of calibration fairness in simulation 1. Figure 4 shows how (group-specific) prediction error depends on the type of latent predictor used as well as on the simulated group differences and group size. We observe that models which use the group-overarching predictor yield a low MSE overall, while using the group-specific predictor leads to an increase in MSE with increases in group differences (α). Furthermore, the group-specific error depends on the

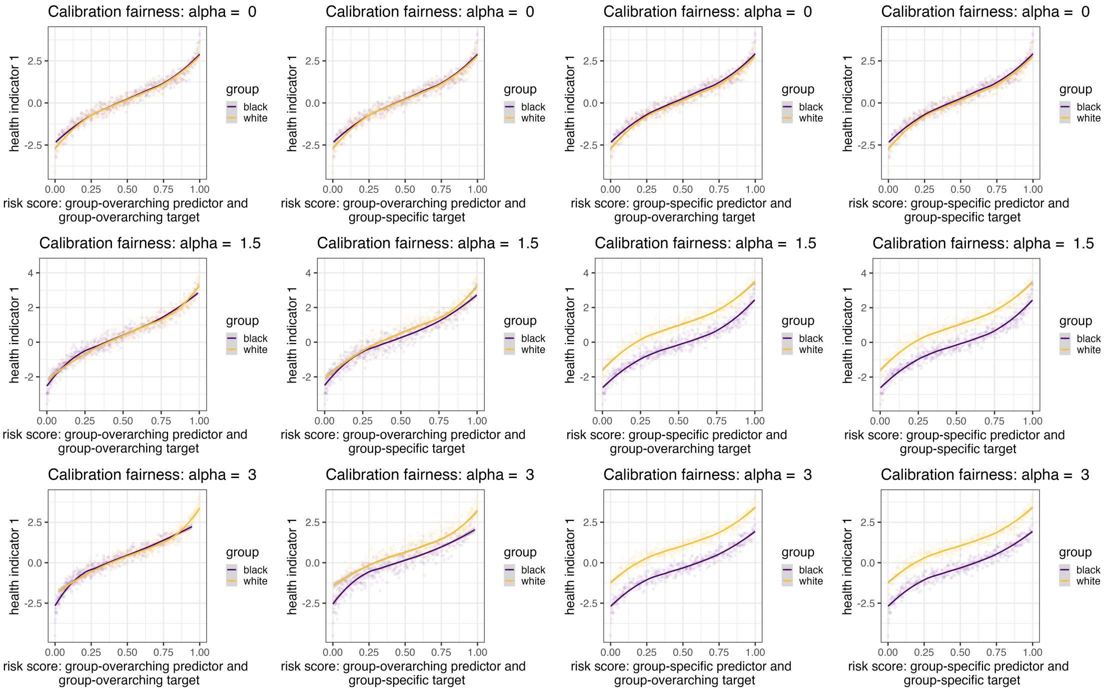


Figure 3: Calibration fairness of risk scores based on prediction models with different types of latent targets and predictors.

Predictor in LASSO regression	Overarching target	Group-specific target
Group-overarching predictor	61 (100%)	26 (42.6%)
Group-specific predictor	12 (19.7%)	61 (100%)

Table 1
Absolute frequencies and percentages of different predictors getting chosen during LASSO prediction.

proportion of the disadvantaged group in the sample. With imbalanced group sizes (right plot in Figure 4) and group-specific predictors, increasing group differences lead to an increase in prediction error particularly for members of the minority group, i.e. Black individuals (comparison between light blue and dark blue line). When the group-overarching scores are used as predictor, the MSE does not increase with group-differences for both groups (comparison between orange and red line).

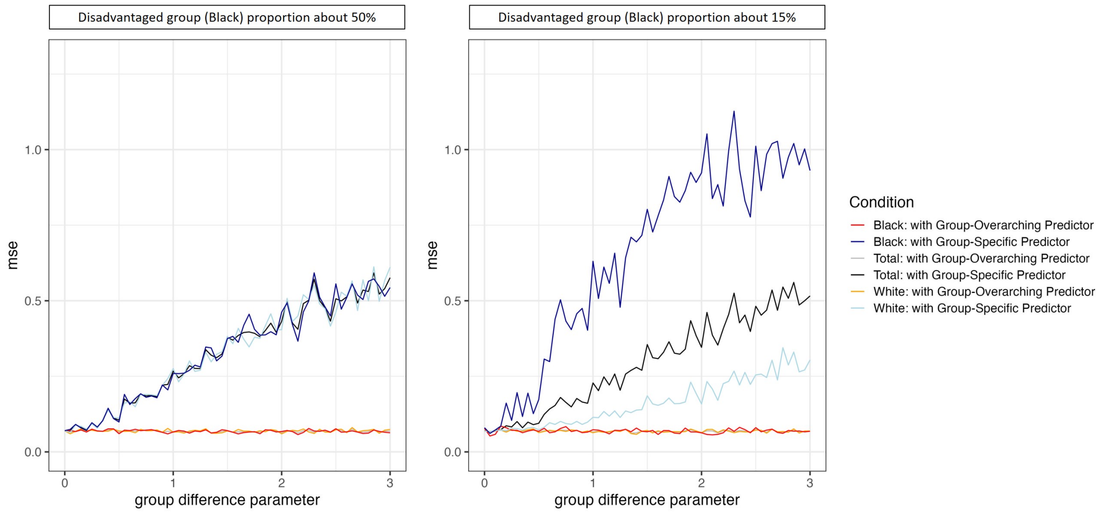


Figure 4: Overall and group-specific error of prediction models using different types of predictors (group-overarching target).

5. Application Study

In the application study, we apply SEM-based measurement and prediction modeling to a dataset which includes multiple health indicators, health cost indicators and demographic variables. We study calibration fairness of the risk score using single group SEMs for modeling the target variable and compare the effects of group-overarching versus group-specific cost predictors.

5.1. Methods

5.1.1. Data

We use the data provided by Obermeyer et al. [10] and similarly used by Boeschoten et al. [14]. It contains about 48,000 observations and 160 variables with health indicators and health costs measured at two different time points, the current timepoint (t) and at a timepoint one year earlier ($t-1$), as well as demographic information. This information can be used to train a risk prediction model, based on which individuals may be assigned to health support programs [10].

5.1.2. Measurement Models

We perform psychometric modeling of the prediction target, health status at timepoint t, and for the predictor health costs at timepoint $t-1$. The target is based on latent scores of a group-overarching measurement model. We use the following health indicators to define the latent target: number of chronic diseases, cholesterol, blood sugar, kidney function, blood pressure, and anemia.

We create two different versions of the predictor health costs. The latent scores of a group-overarching model mapping health costs on a common scale for white and Black individuals (recognizing the differences in health costs between the groups), and the latent scores of a group-specific model mapping health costs on two separate scales (equalizing differences in health costs between the groups). The indicators for the latent cost variables are costs in dollars for: dialysis, emergency care, home care, in patient medical care, in patient surgery, laboratory procedures, and out patient primary care, out patient specialists, out patient surgery, pharmaceutics, physical therapy, radiology and other costs. We shift the distributions of the latent scores by their minimum and log-transform the latent values to account for their skewness.

We construct risk scores by predicting the latent health scores derived from the latent variable model for health at t with latent predictors of costs at $t-1$. We use LASSO regression predictions to construct the risk scores.

5.1.3. Evaluation

We evaluate prediction performance and the (mis)calibration of the predicted risk score with respect to different health indicators. We further evaluate which predictors (group-specific or group-overarching latent scores) are chosen in the LASSO regression to predict health. Furthermore, we evaluate the risk scores and prediction errors when only one of the latent predictors is offered for the prediction task.

5.2. Results

The measurement models for constructing the latent variable scores for the target variable health at timepoint t fit the data well with CFI = .91 and RMSEA = .042. The two SEMs estimated for modeling health costs at $t-1$ have a similar model fit. The fit for both the group-overarching model (CFI = .83; RMSEA = .050) and for the group-specific model (CFI = .82; RMESA = .052) is acceptable.

To construct the risk score, we run a LASSO regression predicting latent health. The final model is chosen by cross-validation and results in a MSE of 0.62 and 72.9% of explained deviance. The predictors chosen in the LASSO model are: Age, Hypertension, Kidney function, Complications in diabetes, Number of chronic diseases and the latent scores for health costs of the group-overarching model. These results match the simulation results because again the group-overarching latent scores are chosen over the group-specific scores when predicting a group-overarching target.

We evaluate calibration fairness by plotting the predicted risk score against single health indicators in Figure 5. For three out of six indicators (number of chronic diseases, cholesterol (LDL), and kidney function), we observe congruent lines indicating no differences between Black and white individuals with equivalent risk scores. However, Black individuals have higher blood sugar and higher blood pressure compared to white individuals and white individuals have higher anemia compared to Black individuals with the same predicted risk. These results indicate that the group-overarching SEM strategy was able to mitigate most, but not all, miscalibration in the real data application.

In an additional LASSO regression, we delete the group-overarching predictor which leads to the group-specific predictor being chosen next to the predictors mentioned above. This model has a MSE of 0.62 and explains 72.9% of the deviance. The group-specific prediction errors of

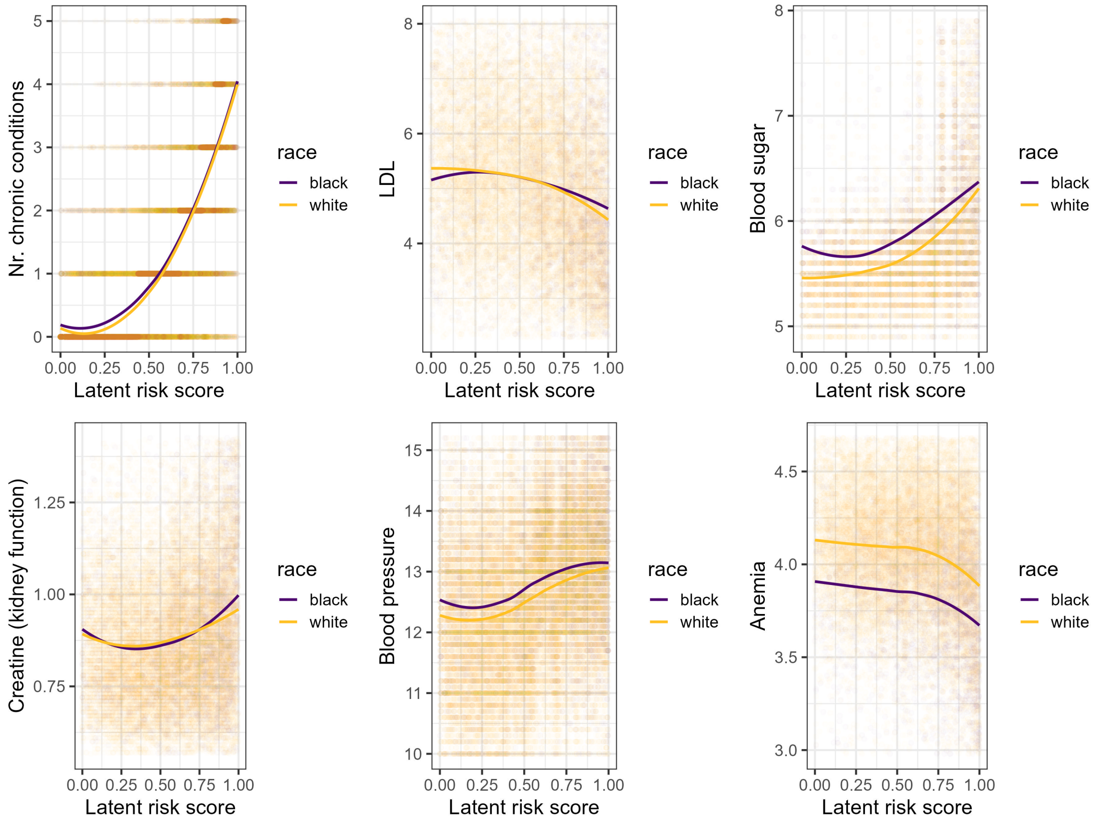


Figure 5: Relationship between latent risk score and health indicators in the application study.

both LASSO models differ only slightly. When the group-overarching predictor is offered, MSE for Black individuals is 1.005. When only the group-specific predictor is offered, the MSE for Blacks increases slightly to 1.008. In comparison, MSE for white individuals increases from 0.5751 to 0.5755. While the increase is higher for Black compared to white individuals, these differences are extremely small.

6. Discussion

We demonstrate the use of structural equation modeling (SEM) to construct latent predictors and targets in prediction modeling contexts. SEM can be used as a pre-processing technique for indicator variables that represent latent concepts. However, the specific type of measurement modeling strategy (single vs. multi-group SEM) that is employed can have considerable fairness implications downstream and thus needs to be chosen carefully.

Our results underscore that predictions can be severely miscalibrated when inadequate modeling strategies are employed. By applying measurement methodology in a use case of predicting individual health from past health and healthcare costs [10, 14], we demonstrate how different types of SEMs can improve calibration fairness and group-specific model error. Given structural group differences, we observe that using group-overarching measurement models are a more effective strategy than using group-specific models. That is, when the true relationship includes group differences, multi-group SEMs with group-specific parameters negate any such differences during pre-processing and thus the resultant latent variable scores can lead to misspecified prediction models. Employing SEMs which preserve group differences while drawing on multiple proxy variables, however, can effectively reduce fairness issues downstream. This finding is independent of the use of healthcare cost as a target indicator and supported by the second simulation which showed that the group-overarching predictor led to prediction models with lower group-specific error.

We highlight that integrating substantive knowledge both about the predictors and the target during model design is critical. We recommend considering SEM whenever predictor or target variables are by-proxy indicators and have questionable value with respect to different social groups. Additionally, however, we advocate for a comprehensive investigation of resulting latent scores under consideration of group-specific sample sizes. Final choices regarding the design of the prediction system should be made using a combination of empirical results and practical as well as substantive considerations in the context of the specific use case. Eventually, the main goal of utilizing SEM during pre-processing is to incorporate substantive knowledge about the prediction task into the model implementation.

We note that we explored only a single use case and thus advocate for more research on the potentials and limitations of different types of SEMs in various ADM contexts. Another limitation of this study is that we only examined risk scores built on the basis of LASSO regression. With many plausible alternative SEM designs and ML models which could have been studied additionally, there is a large space of decisions open for exploration.

In conclusion, we highlight that tackling fairness concerns by utilizing measurement models of both the predictors and the target can contribute to the fairness of ADM systems. The integration of techniques like SEM from psychometrics into machine learning workflows presents a potential avenue for refining the development of fair decision-making algorithms.

References

[1] A. Mukerjee, R. Biswas, K. Deb, A. P. Mathur, Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management, International Transactions in Operational Research 9 (2002) 583–597. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-3995.00375. doi:10.1111/1475-3995.00375, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-3995.00375.

[2] S. Desiere, K. Langenbucher, L. Struyven, Statistical profiling in public employment services, Technical Report 224, 2019. URL: https://www.oecd-ilibrary.org/content/paper/b5e5f16e-en. doi:https://doi.org/https://doi.org/10.1787/b5e5f16e-en.

[3] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, ProPublica (2016) 254–264. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] L. Tangi, C. Van Noordt, M. Combetto, D. Gattwinkel, F. Pignatelli, AI Watch. European

landscape on the use of Artificial Intelligence by the Public Sector, Scientific analysis or review, Policy assessment, Other KJ-NA-31088-EN-N (online), Luxembourg (Luxembourg), 2022. doi:10.2760/39336(online).

[5] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, Machine learning: The high interest credit card of technical debt (2014).

[6] M. J. Kusner, J. R. Loftus, The long road to fairer algorithms, Nature 578 (2020) 34–36.

[7] L. Henriques-Gomes, Robodebt: five years of lies, mistakes and failures that caused a $1.8bn scandal, The Guardian (2023). URL: https://www.theguardian.com/australia-news/2023/mar/11/robodebt-five-years-of-lies-mistakes-and-failures-that-caused-a-18bn-scandal.

[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM computing surveys (CSUR) 54 (2021) 1–35.

[9] K. T. Rodolfa, P. Saleiro, R. Ghani, Bias and Fairness, 2 ed., Chapman and Hall/CRC, 2020. Num Pages: 32.

[10] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (2019) 447–453. URL: https://www.science.org/doi/abs/10.1126/science.aax2342. doi:10.1126/science.aax2342.

[11] K. G. Jöreskog, Structural analysis of covariance and correlation matrices, Psychometrika 43 (1978) 443–477.

[12] B. Muthén, A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, Psychometrika 49 (1984) 115–132.

[13] A. Z. Jacobs, H. Wallach, Measurement and fairness, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 375–385. URL: https://doi.org/10.1145/3442188.3445901. doi:10.1145/3442188.3445901.

[14] L. Boeschoten, E.-J. van Kesteren, A. Bagheri, D. L. Oberski, Achieving fair inference using error-prone outcomes, International Journal of Interactive Multimedia and Artificial Intelligence 6 (2021) 9–15. doi:10.9781/ijimai.2021.02.007.

[15] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and information systems 33 (2012) 1–33.

[16] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, K. R. Varshney, Optimized pre-processing for discrimination prevention, Advances in neural information processing systems 30 (2017).

[17] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: International conference on machine learning, PMLR, 2013, pp. 325–333.

[18] M. Zanger-Tishler, J. Nyarko, S. Goel, Risk scores, label bias, and everything but the kitchen sink, Science Advances 10 (2024) eadi8411. URL: https://www.science.org/doi/abs/10.1126/sciadv.adi8411. doi:10.1126/sciadv.adi8411.

[19] R. Fogliato, A. Chouldechova, M. G'Sell, Fairness evaluation in presence of biased noisy labels, in: International conference on artificial intelligence and statistics, PMLR, 2020, pp. 2325–2336.

[20] J. Wang, Y. Liu, C. Levy, Fair classification with group-dependent label noise, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 526–536. URL: https://doi.org/10.1145/3442188.3445915. doi:10.1145/3442188.3445915.

[21] L. Guerdan, A. Coston, K. Holstein, Z. S. Wu, Counterfactual prediction under outcome measurement error, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1584–1598. URL: https://doi.org/10.1145/3593013.3594101. doi:10.1145/3593013.3594101.

[22] S. Milli, L. Belli, M. Hardt, From optimizing engagement to measuring value, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 714–722. URL: https://doi.org/10.1145/3442188.3445933. doi:10.1145/3442188.3445933.

[23] J. Lemaire, Automobile insurance: actuarial models, volume 4, Springer Science & Business Media, 2013.

[24] L. W. Perna, Sex and race differences in faculty tenure and promotion, Research in higher education 42 (2001) 541–567.

[25] D. E. Thissen, H. E. Wainer, Test scoring., Mahwah: Lawrence Erlbaum Associates Publishers, 2001.

[26] R. O. Mueller, G. R. Hancock, Structural equation modeling, in: The reviewer's guide to quantitative methods in the social sciences, Routledge, 2018, pp. 445–456.

[27] K. G. Jöreskog, Simultaneous factor analysis in several populations, Psychometrika 36 (1971) 409–426.

[28] N. J. Dorans, L. L. Cook, Fairness in educational assessment and measurement, Taylor & Francis, 2016.

[29] P. W. Holland, H. Wainer, Differential item functioning, Routledge, 2012.

[30] L. Hu, P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Structural Equation Modeling: A Multidisciplinary Journal 6 (1999) 1–55. URL: https://doi.org/10.1080/10705519909540118. doi:10.1080/10705519909540118. arXiv:https://doi.org/10.1080/10705519909540118.

[31] K. G. Jöreskog, A general approach to confirmatory maximum likelihood factor analysis, Psychometrika 34 (1969) 183–202.

[32] D. L. Jackson, The effect of the number of observations per parameter in misspecified confirmatory factor analytic models, Structural Equation Modeling: A Multidisciplinary Journal 14 (2007) 48–76. URL: https://doi.org/10.1080/10705510709336736. doi:10.1080/10705510709336736.

[33] T. A. Schmitt, Current methodological considerations in exploratory and confirmatory factor analysis, Journal of Psychoeducational Assessment 29 (2011) 304–321. URL: https://doi.org/10.1177/0734282911406653. doi:10.1177/0734282911406653.

[34] 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 67 (2005) 91–108.

[35] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2024. URL: https://www.R-project.org/.

[36] Y. Rosseel, lavaan: An R package for structural equation modeling, Journal of Statistical Software 48 (2012) 1–36. doi:10.18637/jss.v048.i02.

[37] J. K. Tay, B. Narasimhan, T. Hastie, Elastic net regularization paths for all generalized linear models, Journal of Statistical Software 106 (2023) 1–31. doi:10.18637/jss.v106.i01.

A. Appendix

We conducted simulations with an additional set of target variables. While the commercial risk score [10] and also the analyses by Boeschoten et al. [14] included health costs as an indicator for the prediction target, we consider this measurement path problematic. Therefore, we repeated our analyses without using the cost indicator for constructing the latent target (Figure 6). The calibration fairness results, however, remained unchanged, highlighting that the main cause of variation is the use of single- vs. multi-group SEMs when constructing latent scores (Figure 7).

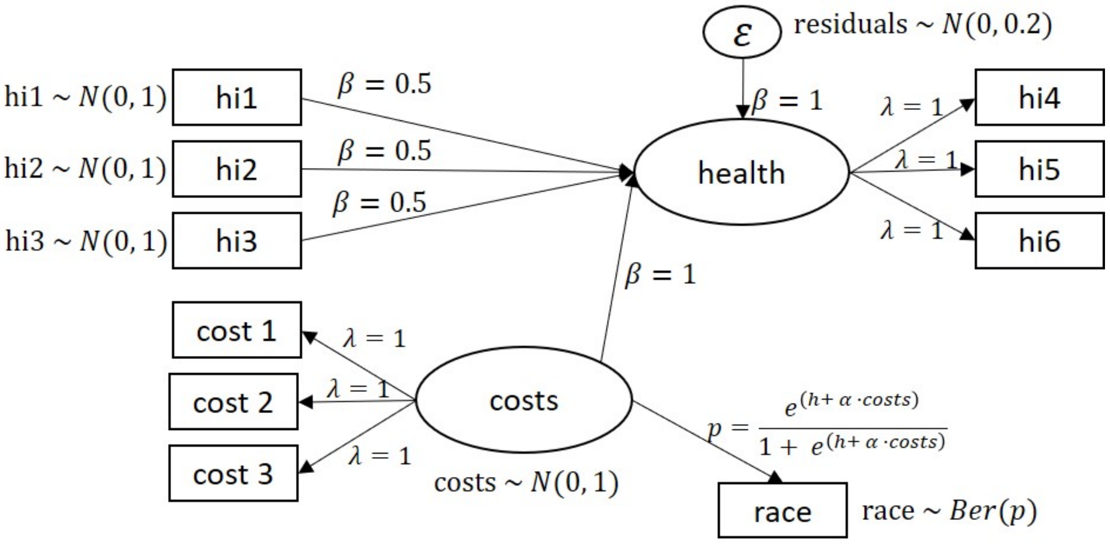


Figure 6: Data simulation setup for additional set of target variables (without cost as target indicator).

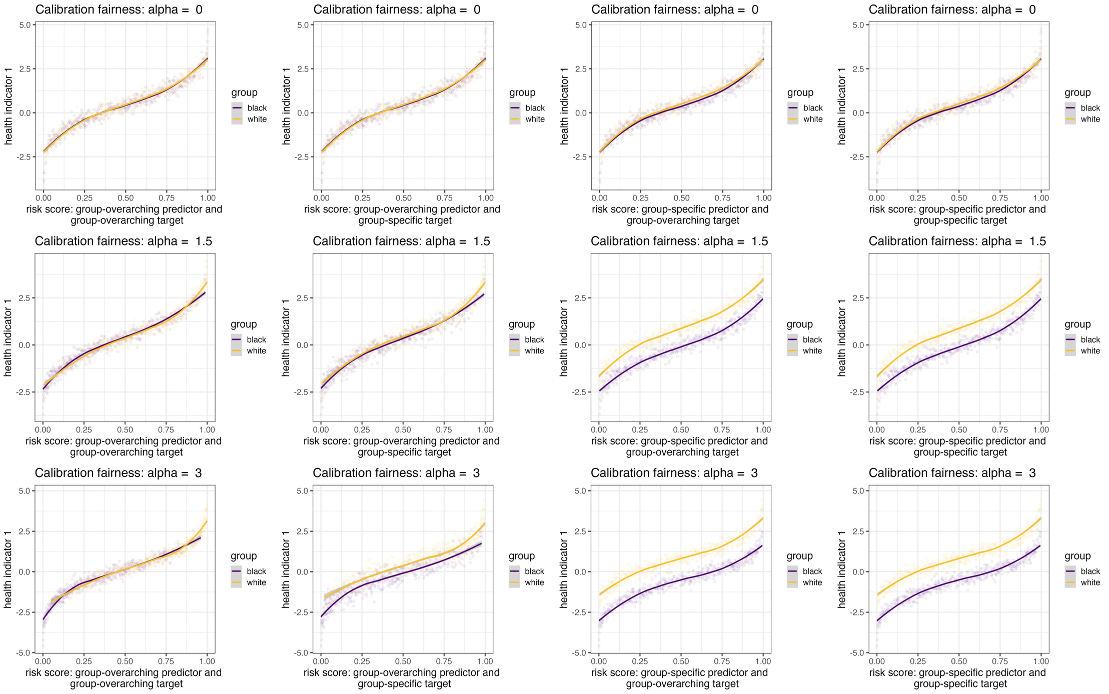


Figure 7: Calibration fairness of risk scores based on prediction models with different types of latent targets (without cost as target indicator) and predictors.