

Hierarchical Bayesian models to mitigate systematic disparities in prediction with proxy outcomes

Jonas M. Mikhaeil¹ , Andrew Gelman^{1,2} and Philip Greengard¹

¹Department of Statistics, Columbia University, New York 10027, USA

²Department of Political Science, Columbia University, New York, USA

Address for correspondence: Jonas M. Mikhaeil, Department of Statistics, Columbia University, New York, NY 10027, USA.
Email: j.mikhaeil@columbia.edu

Abstract

Label bias occurs when the outcome of interest is not directly observable and instead, modelling is performed with proxy labels. When the difference between the true outcome and the proxy label is correlated with predictors, this can yield systematic disparities in predictions for different groups of interest. We propose Bayesian hierarchical measurement models to address these issues. When strong prior information about the measurement process is available, our approach improves accuracy and helps with algorithmic fairness. If prior knowledge is limited, our approach allows assessment of the sensitivity of predictions to the unknown specifications of the measurement process. This can help practitioners gauge if enough substantive information is available to guarantee the desired accuracy and avoid disparate predictions when using proxy outcomes. We demonstrate our approach through practical examples.

Keywords: algorithmic fairness, Bayesian hierarchical models, label bias, measurement models, prediction

1 Introduction

In the social sciences, measurement is often indirect, and researchers use proxy outcomes (Adcock & Collier, 2001; Knox et al., 2022). Even seemingly objective outcomes such as suicide rates can be systematically distorted (Douglas, 1967). Sociological accounts of the processes with which data are collected highlight the unavoidable imperfections of data more broadly (Starr, 1987). The use of imperfect proxies for the outcome can reduce the accuracy of predictions that are relevant to downstream decisions, possibly underserving specific subgroups of the population (Fogliato et al., 2020; Mullainathan & Obermeyer, 2021; Obermeyer et al., 2019). We propose to mitigate these problems by modelling the relationship between proxy and true outcomes with Bayesian measurement models.

Consider the example of building a statistical model to predict diabetes risk using demographic and health information from survey data. The goal of building such a model is to be able to cheaply identify patients who are at risk of diabetes and who should undergo more costly and time-consuming testing. The model should be accurate and calibrated. If the model underpredicts the risk for certain groups of people, then decisions based on it can lead to these groups being underserved.

One challenge in this example is that we are only given the diagnosis, not true underlying disease status. There are several potential sources of error (usually referred to as *label bias*) that this proxy outcome may introduce into a model. If the measurement error—the difference between the proxy outcome (survey response) and the true outcome (being diabetic)—is correlated with a predictor, then prediction errors can be correlated with that predictor. We demonstrate with a simple example in a linear regression setting in Section 2.1 and return to the example of diabetes risk in Section 4.

Received: March 1, 2024. Revised: August 2, 2024. Accepted: November 23, 2024

© The Royal Statistical Society 2024. All rights reserved. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

There are various ways of dealing with label bias in specific contexts (Jiang & Nachum, 2020; Knox et al., 2022; Wang et al., 2021). Label bias often degrades prediction accuracy and, when the measurement errors are correlated with the covariates, leads to systematic errors in prediction. In the context of predicting risk, these systematic disparities in prediction are referred to as miscalibration (Rothblum & Yona, 2023) and have been shown to negatively impact the utility of downstream decisions (Parastouei et al., 2021; Van Calster & Vickers, 2015). Label bias is especially problematic when measurement errors are correlated with membership to a protected group, which is often the case in social science applications (Biderman & Reiss, 1967; Fang et al., 2022; Zanger-Tishler et al., 2024) or the healthcare sector (Basu, 2023; Cerdeña et al., 2020; Diao et al., 2021; Eneanya et al., 2019). In this situation, decisions based on these predictions can lead to some communities being under-served on average, thus violating certain conceptions of algorithmic fairness (Corbett-Davies et al., 2023; Dwork et al., 2012; Hardt et al., 2016).

Zanger-Tishler et al. (2024) show that in the presence of label bias, the addition of features may deteriorate prediction accuracy on the true labels of interest. In particular, if a feature's correlation with the true outcome and proxy outcome, conditional on the other covariates, have different signs, then including that feature in a regression will deteriorate predictive accuracy. This can occur when a feature is only weakly related with the true outcome but both this feature and the outcome are causally constitutive of the remaining features. Zanger-Tishler et al. (2024) demonstrate this situation with the relationship between criminal behaviour (the outcome of interest), arrests (the proxy outcome), and the level of policing in a neighbourhood; we continue studying this example in Section 2.2.

In the present paper, we demonstrate that, in the setting where dropping a predictor would increase prediction accuracy, we can increase prediction accuracy even further using a measurement model and that, with sufficient knowledge about the data-generating process, measurement models can mitigate systematic disparities in prediction. Our work highlights the benefits of making explicit assumptions about measurement errors, even in purely predictive settings. Measurement models are a way to make these assumptions transparent and allow users to critically question if enough domain knowledge is at hand to make the proxies valid and to ensure that downstream decisions based on them do not underserve specific groups of interest. While measurement models, in principle, allow researchers to adjust predictions to mitigate disparities and achieve decisions that improve outcomes for particular groups, the inclusion of membership information to protected groups may be problematic in itself (Goel et al., 2017) and violate the legal doctrine of 'no disparate treatment'. We do not address this tension here; in any application with label bias of this sort, both societal and legal considerations will be crucial.

Building measurement models tailored to specific applications has been made easier by recent advances in probabilistic programming languages such as Stan (Stan Development Team, 2023), where reasonably general Bayesian models can be set up in simple, user-friendly language, allowing researchers to represent prior knowledge, including uncertainty, about the measurement process and any discrepancy between the proxy and the true outcome in a statistical model.¹

In Section 2, we introduce hierarchical Bayesian measurement models and discuss general methodological considerations. We go on to discuss pitfalls of correlated measurement error in the simple case of linear regression, where label bias can be studied analytically (see Section 2.1). After presenting our proposed methodology, we demonstrate the use of Bayesian measurement models in two applications. In Section 3, we study the simulated criminal justice model considered in Zanger-Tishler et al. (2024). Next, in Section 4, we consider the problem of predicting diabetes risk based on diagnosis information. We use public health research on diabetes prevalence to adjust for the fact that among diabetics, diagnoses are more likely to be made in those with healthcare access. By adjusting predictions for healthcare status, we achieve more accurate and equitable predictions than possible with regression on the proxy labels. While the examples are chosen to resemble real-world applications, they are not supposed to be case studies. Rather, they are chosen to showcase how our proposed methodology—Bayesian measurement models—might improve on classical techniques dealing with label bias.

¹ All models and code to reproduce our results are available under <https://github.com/JonasMikhaeil/HierarchicalBayesianMeasurementModels>.

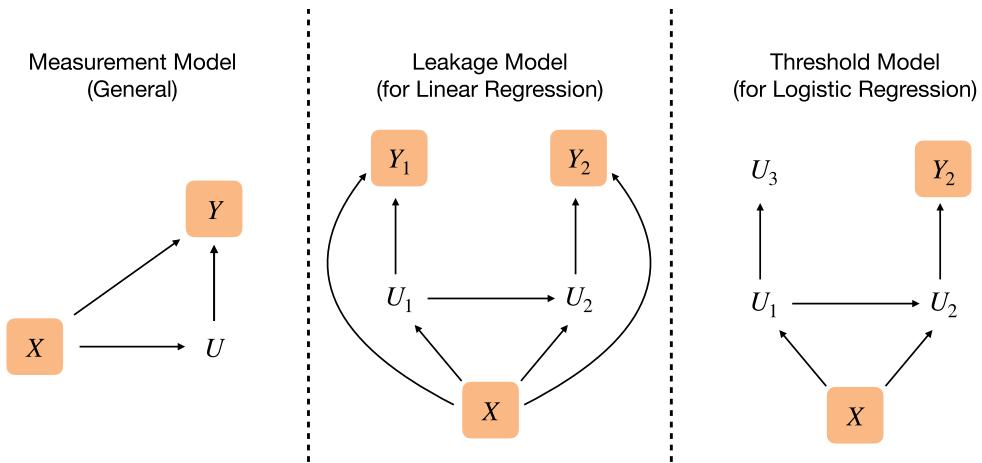


Figure 1. Some measurement models for label bias.

2 Measurement models for label bias

In situations in which only a noisy proxy y of the desired outcome of interest u is available, some model, explicit or implicit, of the measurement process is necessary for accurate and reliable prediction. The classical approach of using regression $\mathbb{E}(y | X)$ on the proxy labels to predict the true outcome u implicitly equates the outcome of interest and the observed proxy outcome. In the case of linear regression, this yields accurate inference if the measurement error is mean independent of the covariates X , see Section 2.1. Often, there is good reason to believe this is not the case. Measurement models in general, and Bayesian hierarchical models, in particular, are a useful tool to model more complicated measuring processes and account for noise that is not independent of the covariates.

The general idea behind measurement models (see Figure 1) is to introduce the true outcome u as a latent (unobserved) quantity. Prior knowledge about the application is then used to model the relationship between the covariates X , the latent outcomes of interest u , and the observed proxies y . Because parts of the variables remain unobserved, some of the model parameters are not (or only partially) identified (Gustafson, 2015). Measurement models thus rely on domain knowledge in two ways: The measurement process has to be sufficiently understood to supply a model structure (which includes distributional assumptions about the latent outcomes) as well as reasonable values of the nonidentified parameters of the model. We give guidance on how to determine which parameters require strong priors in [online supplementary material Appendix C](#).

For the identified part of the model, classical advice about Bayesian workflow applies. In particular, posterior predictive checks (Gelman et al., 1996; Rubin, 1984) can be used to assess model fit. If parametric assumptions are too rigid, nonparametric components (such as Gaussian processes or splines) can be used. Another way of adding flexibility and moving beyond the limitations of parametric models is to add unit-specific error terms (such as in the threshold model of Section 2.3).

When only limited prior knowledge is available, nonidentified parameters should be treated as sensitivity parameters in a sensitivity analysis (Richardson et al., 2011; Depaoli et al., 2020; Kallioinen et al., 2024). Such an analysis is performed in Section 3.3, which details the impact of misspecification of the parameters in a stylized example where the data-generating process is known. Gelman and Hennig (2017) discuss the use of informative priors in Bayesian practice more generally and the value of transparency in scientific endeavours.

Measurement models are flexible and can be tailored to the application of interest. Here, we present two models, a leakage model for linear regression, which we will use to model a stylized example of arrests and crime (see Section 3), and a threshold model for logistic regression, which we will apply to estimate diabetes risk based on diagnosis data (see Section 4). Before we do so, we will illustrate the pitfalls of dependent label bias explicitly in the case of linear regression.

2.1 Simple illustration: label bias in linear regression

In this section, we use the simple case of linear regression to analytically demonstrate issues that can arise when using regression on proxy outcomes to predict true outcomes. The validity of this classical approach rests on the assumption that the measurement error is uncorrelated with the covariates. We demonstrate that if this assumption is inaccurate, predictions can be systematically inaccurate. Throughout this section, we treat the covariates X as random (Buja et al., 2016, 2019; Rosset & Tibshirani, 2020) allowing them to be correlated with the measurement errors.

We provide three main formulas. First, in Proposition 1, we provide a formula for the error of the linear regression solution when fitting on a proxy as opposed to the true outcome. While our proposition is focused on regression on proxies, it is similar to the well-known omitted variable bias (Wooldridge, 2010). Proposition 2 demonstrates that when the proxy is correlated with predictors, then the prediction error is also correlated with the predictors. Finally, Proposition 3 provides a lower bound on the prediction error when using a proxy outcome in terms of the prediction error when using the true outcome. The primary purpose of these propositions is to demonstrate the systematic errors that can arise when using proxy labels in a simple setting that can be studied analytically. Proofs can be found in [online supplementary material Appendix A](#).

We start by assuming that some true outcome, u , and a proxy outcome, y , are n -dimensional random vectors. We also assume that X is an $n \times m$ random matrix of centred covariates with a leading column of ones such that $\mathbb{E}(X^T X)$ is full rank, i.e. the covariates are not multicollinear. We assume $(X, u, y) \sim P$, where P is some probability distribution over the covariates, true outcome, and proxy outcome.² We define β to be the expected solution to linear regression with covariates X and data u . That is,

$$\beta = \arg \min_w \mathbb{E}(\|Xw - u\|^2). \quad (1)$$

The expected solution to the linear regression changes when using the proxy outcome y and the same covariates X . The expected solution with a proxy outcome is given by the following proposition.

Proposition 1 ((Proxy outcome regression solution)). Let $(X, u, y) \sim P$. Then,

$$\arg \min_w \mathbb{E}(\|Xw - y\|^2) = (1 + \gamma)\beta + \alpha, \quad (2)$$

where the vector $[\alpha \gamma] \in \mathbb{R}^{m+1}$ is the expected solution to the linear regression with outcome $u - y$ (the measurement error) and $n \times (m + 1)$ matrix of covariates $[X \ u]$. That is,

$$[\alpha \gamma] = \arg \min_w \mathbb{E}(\|Mw - e\|^2), \quad (3)$$

where e is the measurement error defined by $e = u - y$ and where M is defined to be the $n \times (m + 1)$ random matrix $[X \ u]$.

That is, if the measurement error e is uncorrelated with the covariates and the outcome, then, in expectation, β is recovered from the proxies. On the other hand, correlation between the measurement error and the covariates or the outcome will introduce error in the approximation of β . That error, $\gamma\beta + \alpha$, is obtained from combining (1) and (2). This can pose problems in causal investigations (Knox et al., 2022) and even in predictive settings. The right panels of [Figure 2](#) provide an illustrative example of error introduced by the use of a proxy outcome in the linear regression setting. We demonstrate the case where $m = 2$, i.e. X consists of an intercept and one predictor.

When label bias introduces error into the solution to a linear regression, the predictions made using that linear regression will be systematically distorted. We define the predictions as $\hat{E}(y | X) = X(X^T X)^{-1}X^T y$. The following proposition provides a formula for the covariance between the covariates, X , and prediction error, $u - \hat{E}(y | X)$.

² We assume that the expectations taken with respect to P in the proofs of Section A all exist.

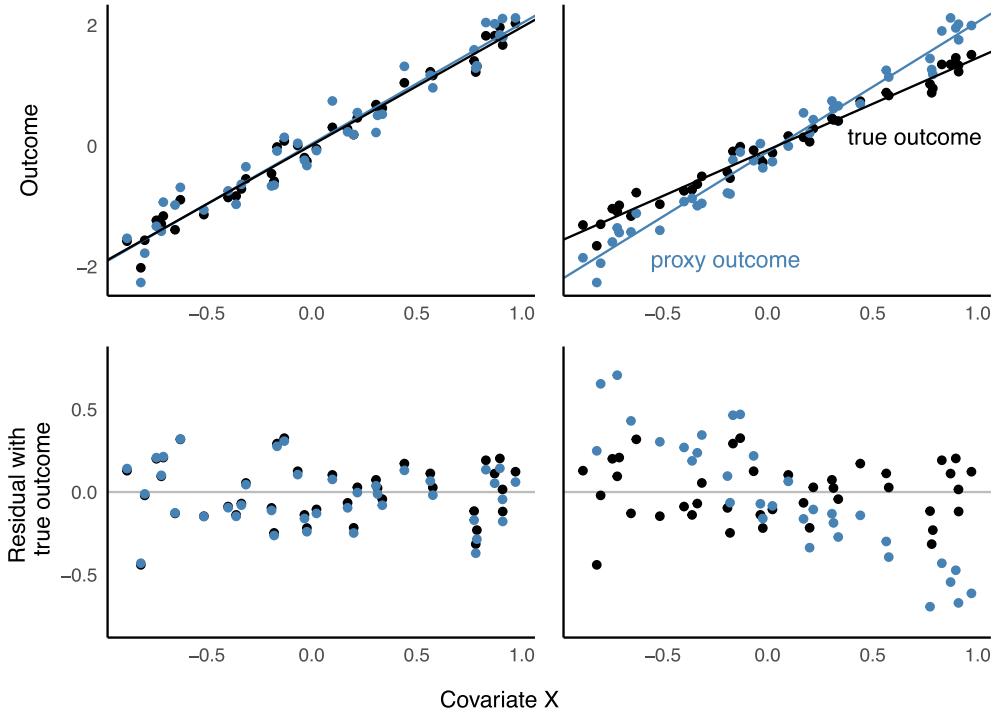


Figure 2. Illustration of label bias in linear regression. (Left) If the measurement errors are uncorrelated with the covariate, regression yields unbiased and consistent estimates. (Right) In the case of dependent measurement errors, regression estimates are biased and inconsistent. Prediction accuracy is degraded.

Proposition 2 ((Covariance of covariates and prediction error)). Let $(X, u, y) \sim P$. Then,

$$\mathbb{E}[(u - \hat{E}(y | X))^T X] = -(\gamma\beta + \alpha)^T \mathbb{E}(X^T X), \quad (4)$$

where β is defined in (1), and α, γ are defined in (3).

That is, if there is a correlation between the covariates and the measurement error $u - y$, then the prediction error will also be correlated with the covariates. This shows that the use of proxy labels may introduce *systematic* disparities in predictions. These disparities are liable to negatively affect downstream decisions based on them (Parastouei et al., 2021; Van Calster & Vickers, 2015) and may lead to protected groups being underserved, thus violating certain conceptions of algorithmic fairness (Corbett-Davies et al., 2023; Dwork et al., 2012; Hardt et al., 2016).

In our last proposition, we compare prediction error when fitting with the true outcome to prediction error when using a proxy. In particular, we provide a lower bound for the mean squared error (MSE) in the true outcome using linear regression predictions trained on a proxy in terms of the MSE in the true outcome using linear regression trained on the true outcome. We show that label bias degrades prediction accuracy when using linear regression because of the systematic disparities in prediction caused by the correlation between the measurement error and the outcome u and covariates X .

Proposition 3 ((Prediction error with true outcome versus proxy)). Let $(X, u, y) \sim P$. Then we have

$$\text{MSE}(u, \hat{E}(y | X)) \geq \text{MSE}(u, \hat{E}(u | X)) + (\gamma\beta + \alpha)^T \mathbb{E}(X^T X)(\gamma\beta + \alpha),$$

where β is defined in (1), and α, γ are defined in (3).

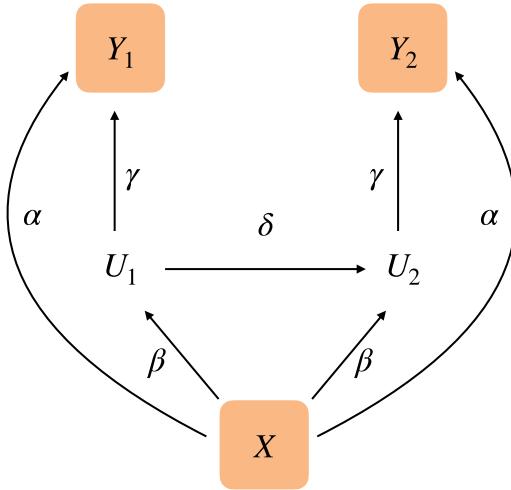


Figure 3. Data-generating process for Zanger-Tishler et al. (2024) stylized example of criminal behaviour (true outcome) and arrest (proxy outcome). Observed variables are marked with a box.

In Section 3 and 4, we will see that given sufficient domain knowledge these systematic disparities in prediction can be mitigated, improving both overall prediction accuracy and reducing the risk of exacerbating disparate outcomes of downstream decisions.

2.2 Leakage model for linear regression

Measurement models are tailored to specific applications and depend on both knowledge about the structure and the parameters of the measurement process. In this section, we describe a *leakage model* for linear regression based on the stylized criminal justice example we will study in Section 3. Suppose we observe a proxy label y_t at two different time points $t \in \{1, 2\}$. These proxies depend both on the observed covariates X and the true outcomes u_t . In the criminal justice example, arrests are proxies y_t for the true outcome u_t of crime. Not all crime leads to arrests, so there is a degree of leakage between proxies and latent outcomes of interest. We assume that the proxies do not influence each other; that is, the entire temporal relationships in the model are driven by the dependence of u_2 on u_1 .³ We are interested in learning this relationship and inferring u_t based on y_t and X . This assumption is based on our knowledge of the data-generating process for the example we are studying here (see Figure 3). In other situations, we might assume that the proxies at time $t = 1$ influence the outcome at time $t = 2$, for example, arrests might deter future crime. Measurement models are flexible enough to allow for this and our model is easily extended to this case.

This situation studied here is illustrated in the centre panel of Figure 1 and can be modelled by the following Bayesian hierarchical model:

$$\begin{aligned}
 y_1 | u_1, \alpha, \gamma, \sigma_y &\sim \text{normal}(X\alpha + \gamma u_1, \sigma_y) \\
 y_2 | u_2, \alpha, \gamma, \sigma_y &\sim \text{normal}(X\alpha + \gamma u_2, \sigma_y) \\
 u_1 | \beta \sigma_u &\sim \text{normal}(X\beta, \sigma_u) \\
 u_2 | u_1, \beta, \eta, \sigma_u &\sim \text{normal}(X\beta + \eta(u_1 - X\beta), \sigma_u \sqrt{1 - \eta^2}),
 \end{aligned} \tag{5}$$

with appropriate priors on all parameters. Because the true outcomes (u_1, u_2) remain unobserved, this model is only partially identified (Gustafson, 2015). We give guidance on identifying parameters that require strong priors in [online supplementary material Appendix C](#). In this example,

³ The centre panel of Figure 1 has an arrow from u_1 to u_2 , implying a causal relationship if the figure is understood as a directed acyclic graph. Our model, however, does not rest on this assumption and is still applicable if u_1 and u_2 are just assumed to be correlated.

weak priors suffice for (α, η, σ_y) when we use strong priors on $(\beta, \gamma, \sigma_u)$. We will use this model for a stylized example of criminal behaviour and arrests in Section 3.

2.3 Threshold model for logistic regression

Here, we develop a *threshold model* for logistic regression. We deploy this model for diabetes prediction in Section 4.

Suppose we observe binary proxy labels $y \in \{0, 1\}$ instead of a binary outcome of interest u_3 . The proxies are indicative of the true outcome but they are not fully reliable, that is there are cases of u_3 that y does not indicate. In our diabetes example, u_3 indicates diabetes disease status. Not everyone with diabetes is diagnosed, however, so diagnosis y is not a fully reliable proxy.

This situation can be modelled by introducing two (continuous) latent characteristics u_1 and u_2 that cause u_3 and y , respectively, by crossing a threshold:

$$\begin{aligned} y &= \begin{cases} 1 & \text{if } u_2 \geq 0 \\ 0 & \text{else} \end{cases} \\ u_1 | \beta &\sim \text{logistic}(X\beta, 1) \\ u_2 &= u_1 - t(X) - e \\ u_3 &= \begin{cases} 1 & \text{if } u_1 \geq 0 \\ 0 & \text{else} \end{cases} \\ e &\sim \text{normal}^+(0, 0.1). \end{aligned} \tag{6}$$

The thresholds $t(X)$ can depend on covariates X allowing for disparities in how accurate the proxies are for different subpopulations. The structure of the model is illustrated in the right panel of Figure 1.

In the diabetes example of Section 4, the latent variable u_1 can be understood as quantifying the severity of diabetes. We assume that for uninsured people symptoms have to be more severe to be diagnosed. This is modelled by introducing insurance-dependent thresholds $t(\text{health insurance})$ that offset the latent characteristic u_2 that determines diagnosis. By introducing $e > 0$, we allow for idiosyncratic behaviour that impacts the proxy, e.g. patient's personal propensity to visit a doctor.

We assume that there are no false positives, that is, $u_3 \geq y$. In essence, this assumes that people are not mistakenly diagnosed with diabetes and that their response about their diagnosis is truthful. If we have reasons to believe this to be false, we could allow $t(\text{health insurance})$ to be random, leading to false positive diagnoses for a fraction of the population.

The latent characteristic u_1 depends linearly on the covariates, so the threshold model closely resembles ordinary logistic regression (Gelman et al., 2014) but allows for discrepancies between the outcomes of interest u_3 and the observed labels y .

In Section 4, we use this model to predict diabetes risk based on diabetes diagnosis with varying thresholds based on health insurance status.

3 Stylized example: criminal behaviour and arrests

Figure 3 portrays the data-generating process for a stylized example of label bias in (Zanger-Tishler et al., 2024). This model simulates individual-level behaviour (u_0 and u_1) and arrest outcomes (y_0 and y_1) at two time points. Arrests depend both on an individual's behaviour and the individual's neighbourhood (X). This is a linear structural equation model,

$$\begin{aligned} X &\sim \text{normal}(0, \sigma_X) \\ \begin{bmatrix} u_0 \\ u_1 \end{bmatrix} | X &\sim \text{MVN}\left(\begin{bmatrix} \beta X \\ \beta X \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \delta \\ \delta & \sigma_u^2 \end{bmatrix}\right) \\ y_0 | X, u_0 &\sim \text{normal}(\alpha X + \gamma u_0, \sigma_y) \\ y_1 | X, u_1 &\sim \text{normal}(\alpha X + \gamma u_1, \sigma_y). \end{aligned} \tag{7}$$

Zanger-Tishler et al. (2024) show how to set the variances of the exogenous variables such that the remaining variables (X, u_0, u_1, y_0 , and y_1) are standardized and can be interpreted as the extent to which an individual differs from the population average. For example, u_0 is interpreted as how

criminal an individual is compared with the population, and X as the level of police enforcement in a neighbourhood.

The label bias in this problem arises because only arrests (y_0 and y_1) and neighbourhood (X) are observable. Criminal behaviour, (u_0, u_1) , which is the true outcome of interest, is not observable and therefore arrests are used as a proxy for criminal behaviour. We have two regression models, a simple one $\mathbb{E}(y_1 | y_0)$ and a complex one $\mathbb{E}(y_1 | y_0, X)$. [Zanger-Tishler et al. \(2024\)](#) show (see Corollary 1) that it is preferable (in terms of expected squared difference between true and predicted outcome) to not include an additional feature if the correlation of that feature with the true and proxy outcome conditional on other covariates have differing signs. For the stylized example here, they show that this is the case for the inclusion of neighbourhood, X , in a model for predicting criminal behaviour based on arrests y_0 when the correlation between neighbourhood and criminal behaviour, β , is small.

This simple example illustrates the theoretical insight of [Zanger-Tishler et al. \(2024\)](#), that the inclusion of additional features can degrade the predictive accuracy of regression models in the presence of label bias.

3.1 Bayesian measurement model

Here, we propose to use the leakage model for regression introduced in Section 2.2, which improves on both the simple and complex estimators of Section 3. In this stylized example, the data-generating process is known (see [Figure 3](#)) and corresponds to the model structure of our measurement model; compare Equations (5) and (7). Making an informed decision about the model structure is the first step when using a measurement model. The second step is the choice of priors. While for some parameters weak priors are sufficient, the model is only partially identified ([Gustafson, 2015](#)) necessitating strong priors on the nonidentified part of the model (see [online supplementary material Appendix C](#)).⁴

$$\begin{aligned}\sigma_y &\sim \text{normal}^+(0, 1) \\ \alpha &\sim \text{normal}(0, 1) \\ \eta &\sim \text{normal}(0, 0.2) \\ \beta &\sim \text{normal}(\beta_{\text{true}}, 0.1) \\ \gamma &\sim \text{normal}(\gamma_{\text{true}}, 0.1).\end{aligned}$$

The strong priors are informed by our knowledge of the data-generating process. In general, when prior knowledge is limited, the nonidentified parameters of the model should be treated as sensitivity parameters in a sensitivity analysis. Section 3.3 performs such a sensitivity analysis and investigates the impact of misspecification of the nonidentifiable parameters.

Here, we have implicitly switched from a prediction setting, in which we are only interested in $\mathbb{E}(u_1 | y_0, X)$, to an inference setting where we are modelling the joint distribution $P(u_0, u_1, y_0, y_1)$. If we require predictions on a set of new variables for which only the features are observed, we can do so by incorporating the unobserved outcomes as missing variables.

[Figure 4](#) shows that the error model performs better than both of the regression models with its performance being comparable with a regression on the true labels.

In practice, the measurement process might be more complicated than in this stylized example. For example, we might have a multitude of correlated covariates each impacting the measurement error to varying degrees. Our approach is flexible enough to cover such cases, however, with increasing complexity, it may become less likely that sufficient domain knowledge is available to tightly constrain all nonidentifiable parameters. [Gelman and Hennig \(2017\)](#) discuss the value of transparency and the use of informative priors in practice more generally. While this may limit the efficacy of our method, it also limits the applicability of classical methods. If the measurement process cannot be accounted for, prediction accuracy can be arbitrarily degraded (see [Proposition 3](#)). Domain knowledge is crucial for predictions with label bias. If the structure of the measurement process is known but the values of its nonidentifiable parameters are not, our method offers two advantages

⁴ We assume σ_u is known.

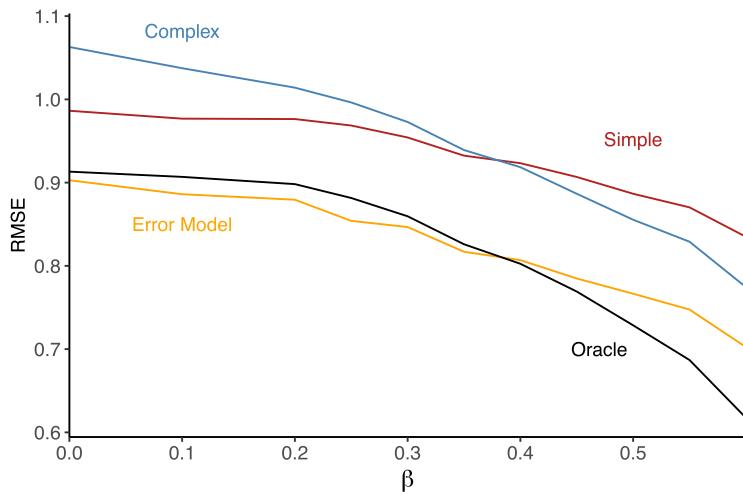


Figure 4. Root mean squared error for simple and complex regression models trained on proxy outcomes in comparison with a Bayesian error model. The prediction accuracy of the error model is superior to both regression models for all β and comparable to an oracle model (trained on the true outcomes).

over classical methods: For one, using wide priors, we can propagate our uncertainty about the measurement process to the predictions. Secondly, we can vary these parameters systematically to check the sensitivity of the predictions to them; see Section 3.3. Neither of these can easily be done for classical methods, risking practitioners to be overly confident in their predictions under label bias.

3.2 Disparate predictions

In many applications systematic disparities in prediction between different subgroups can negatively affect downstream decisions (Parastouei et al., 2021; Van Calster & Vickers, 2015), and, depending on our decision process, lead to decreased fairness (Corbett-Davies et al., 2023; Dwork et al., 2012; Hardt et al., 2016). For both the simple and complex regression models studied in (Zanger-Tishler et al., 2024), prediction errors are correlated with the degree of policing in a neighbourhood X , i.e. they systematically under- or overpredict crime rates based on neighbourhood. This correlation strongly depends on the relationship between neighbourhood and behaviour as well as arrests. On the other hand, prediction errors of the Bayesian measurement model are uncorrelated with neighbourhood X as long as the neighbourhood is accounted for in the model and priors are specified correctly (see Section 3.3). Removing neighbourhood information from the model slightly decreases predictive performance but introduces dependence between prediction errors and neighbourhood X . We plot correlations between prediction errors and neighbourhood in Figure 5.

This shows that modelling the measurement process is key for both overall accurate predictions as well as minimizing systematic disparities in prediction.

3.3 The impact of misspecification

The reliability and accuracy of predictions based on proxies crucially depend on the validity of assumptions we make about the measurement process. In the previous section, we have explored the benefits of measurement models when we can correctly account for the measurement process. While there are real-world examples in which this can be done (see Section 4), this may be unrealistic in the case of predicting crime rates. Despite various proxies being available (for example data on self-reported criminal offending Bureau of Labor Statistics, 2019), the true crime rate is empirically inaccessible. Predictions of the true crime rates thus hinge on untestable assumptions that are often obscured by being stated only implicitly, as is often the case when using regression trained on proxy labels.

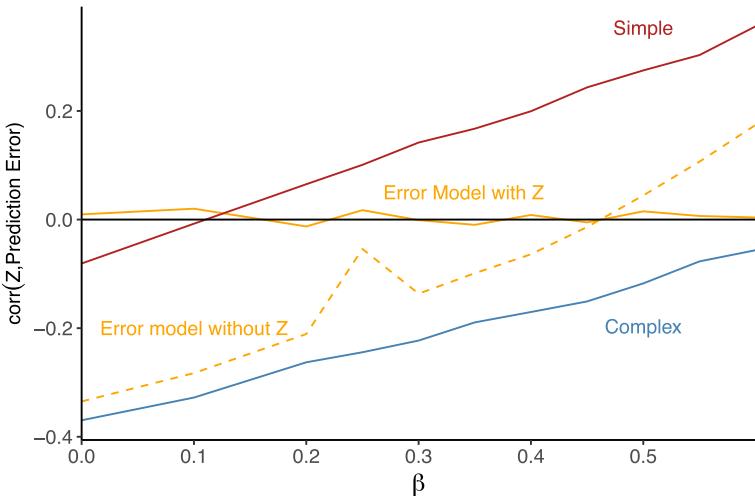


Figure 5. Correlation between prediction error $u_1 - \hat{u}_1$ and neighbourhood police enforcement X for simple and complex regression models trained on proxy outcomes in comparison with a Bayesian measurement model. While both regression models produce predictions that are systematically biased based on neighbourhood, the measurement model has prediction errors uncorrelated with neighbourhood. The dashed line shows that without using neighbourhood information, the measurement model also produces biased predictions.

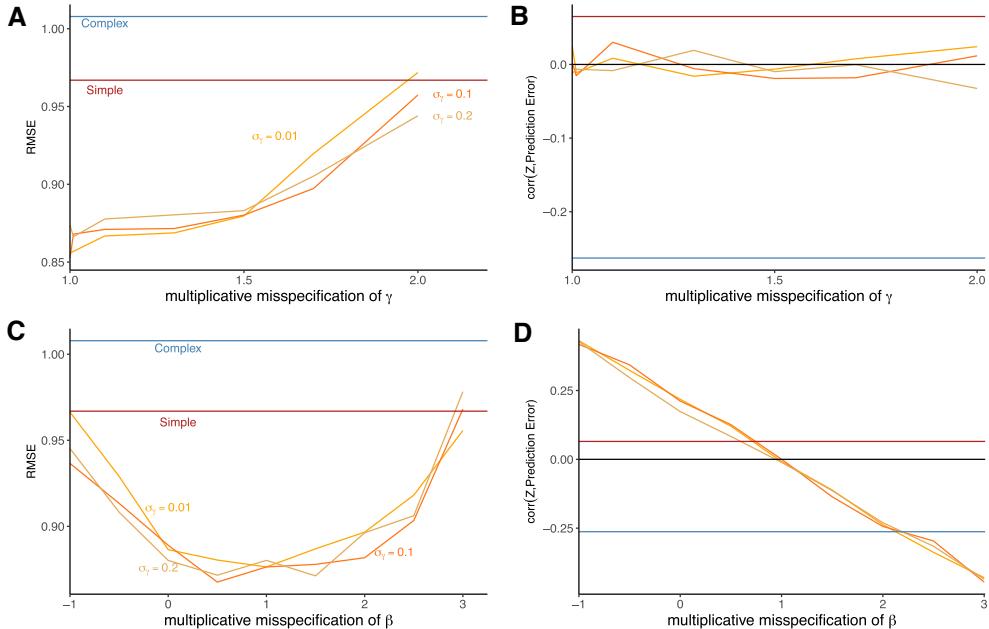


Figure 6. Impact of multiplicative misspecification (\times true parameter) of γ (upper row, $\gamma_{\text{true}} = 0.4$) and β (lower row, $\beta_{\text{true}} = 0.2$) on predictive performance and disparity in prediction accuracy with respect to neighbourhood police enforcement.

Measurement models, however, force us to make our assumptions transparent and allow to test the prediction's sensitivity to them. Figure 6 shows the impact of misspecifying the (nonidentifiable) parameters β and γ in our measurement model (5). While misspecification of either leads to degraded prediction accuracy, correctly specifying β —the relationship between neighbourhood policing and criminal behaviour—is paramount to mitigate systematic disparities in prediction. In

the case of regression models that use proxies to predict crime, these assumptions are often only implicit (and cannot be easily varied), with [Zanger-Tishler et al. \(2024\)](#) criterion being a step in the direction of transparency. Assumptions being made only implicitly, however, neither implies the results to be agnostic or robust with regards to the underlying measurement process. [Figures 4](#) and [5](#) show that the accuracy and systematic disparities in prediction of both the simple and complex model (as well as the decision which one to choose for prediction) depend on the underlying relationship of crime and neighbourhood policing as well.

Given that criminal behaviour, let alone its relationship with policing, is virtually impossible to quantify, predicting crime based on arrests is always skewed by our prior assumptions about crime ([Biderman & Reiss, 1967](#); [Hinton, 2016](#)).

4 Empirical example: health insurance and diabetes

It is estimated that more than 10% of the US population has some form of diabetes ([Centers for Disease Control and Prevention, 2021](#)). While early identification of diabetes is crucial as behavioural counselling, dietary interventions, increased physical activity, or pharmacologic therapy may improve future health outcome ([Davidson et al., 2021](#)), testing for diabetes also comes with monetary and personal costs. In practice, this necessitates risk-based screening decisions ([Duan et al., 2021](#)). In the case that diagnosis information is used to infer the model, predictions will suffer because of label bias. Due to a variety of factors, many people with diabetes have never been diagnosed, making diagnosis an imperfect proxy. For example, it has been estimated that 29% of American diabetics without health insurance remain undiagnosed compared with only 16% with some kind of health insurance ([Fang et al., 2022](#)), a difference that can easily be explained by impeded access to healthcare services. Our analysis is based on publicly available data from the National Health and Nutrition Examination Survey ([Centers for Disease Control and Prevention, 2022](#)), which provides information about both diagnosed (self-reported information of having been diagnosed with diabetes in the past) and undiagnosed diabetes based on measured blood sugar levels. A ready-to-analyse version of this dataset is provided by [Coots et al. \(2024\)](#). This offers an empirically realistic situation of label bias with the necessary ground-truth data to evaluate the advantages and disadvantages of a measurement model, when compared with simple regression or the approach recommended by [Zanger-Tishler et al. \(2024\)](#) to drop a predictor.

The left and centre panels of [Figure 7](#) show that this situation suffers from the phenomenon described in [Zanger-Tishler et al. \(2024\)](#): inclusion of information on the insurance status degrades predictive power when using regression on the proxy labels. For both models, label bias leads to underestimation of diabetes risk. If this bias is not taken into account, decisions based on an optimal treatment threshold are liable to be harmful ([Rothblum & Yona, 2023](#)) because people who would benefit from treatment will not receive it. When insurance status is included as a covariate, disparate predictions are prone to lead to decisions that further under-serve the uninsured

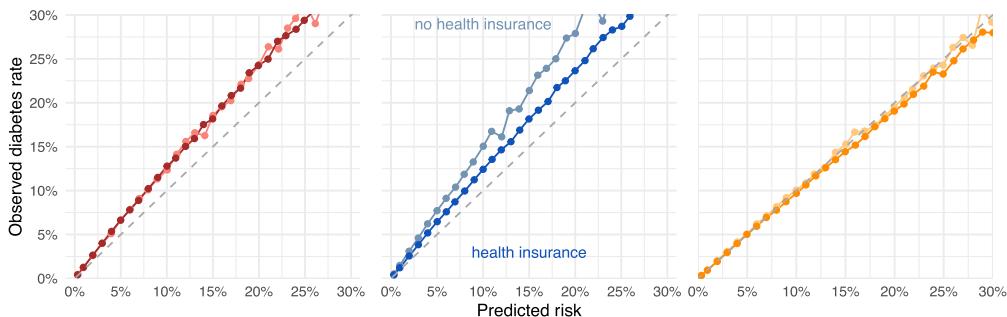


Figure 7. Predicted diabetes risk against diabetes rate against observed diabetes rate estimated with logistic regression on the true outcomes for a simple (left), complex (middle), and measurement model (right) by health insurance status (darker shade: insured, lighter shade: uninsured). For both insured and uninsured people, our measurement model performs better than both regression models closely matching predictions of logistic regression on the true outcomes (dashed grey line).

population, violating conceptions of algorithmic fairness (Corbett-Davies et al., 2023; Dwork et al., 2012; Hardt et al., 2016).

We model this situation with the measurement model presented in Section 2.3. Here, y are binary indicator for diabetes diagnosis (proxy labels) and u_3 indicates diabetes (true outcomes, assumed to not be observed). u_1 is a latent variable that can be understood as the underlying severity of diabetes. We assume that for uninsured people, the severity of symptoms has to be higher to be diagnosed. To account for that, we introduce health insurance-dependent thresholds $t(\text{health insurance})$ that offset the latent characteristic u_2 that determines if a patient is diagnosed.

This measurement model critically depends on the thresholds $t(\text{health insurance})$, which cannot be inferred from diagnosis data alone. We can, however, use prior knowledge, as in (Fang et al., 2022), to inform our choice. In [online supplementary material Appendix B.1](#), we discuss in detail how we determine the thresholds. The right panel of [Figure 7](#) shows that this measurement model based on diabetes diagnosis correcting for impeded access to health care services is well calibrated and predicts diabetes risk better than either a simple or complex regression model. [Table 1 in online supplementary material Appendix B](#) shows improved prediction quality across a range of metrics for classification.

5 Conclusion and discussion

The use of imperfect proxies as dependent variable is ubiquitous in quantitative research in the social sciences. These analyses suffer from label bias, which is often assumed to be a minor problem. If the measurement error is correlated with covariates, label bias can have detrimental effects even in purely predictive settings. In these situations, predictions will suffer from systematic disparities—that is, we will over- or underpredict the outcome systematically based on the covariates. If the measurement errors are correlated with membership in a protected group, these systematic disparities in prediction will not only lead to degraded prediction accuracy but may also be a concern from an algorithmic fairness perspective. In our diabetes example, see Section 4, label bias leads classical predictions of the diabetes risk to systematically underpredict true risk, and more so for uninsured people. Decisions based on these estimates will consequently under-serve uninsured people.

In this paper, we advocate the use of Bayesian measurement models to mitigate these problems. We show that measurement models are preferable to classical regression models in two examples: a stylized criminal justice example, in which the data-generating process is known (see Section 3), and a real-world example where we estimate diabetes risk based on diagnosis information (see Section 4).

We find that when sufficient knowledge about the measurement process is available, these models can mitigate systematic disparities in prediction allowing for more accurate and fairer downstream decisions. Our method explicitly requires the user to model the measurement process. This highlights the importance of assumptions about the relationship between measurement error with covariates for reliable, equitable, and accurate predictions. While these assumptions often remain implicit in classical regression methods, our measurement model helps users to make them more transparent. With this transparency also comes the benefit of being able to test the sensitivity of the predictions to the assumed measurement process. This kind of sensitivity analysis is not easily available for classical methods. Overall, this can allow users to better question if enough domain knowledge is at hand to judge if the proxies are useful and to ensure the fairness of downstream decisions based on them.

While we advocate for modelling the measurement process to mitigate systematic disparities in prediction to achieve fairer downstream decisions, we need to firmly state that this cannot be taken as general advice. Using information necessary in the modelling of proxies, such as protected class status, may be in itself problematic and violate the legal doctrine of ‘no disparate treatment’ (for example the Equal Protection Clause of the U.S. Constitution’s Fourteenth Amendment). This is a fundamental tension and cannot be resolved in general. Any application based on data that is skewed by societal injustices will require careful political, social, and legal consideration. Our paper should, however, be a general warning against the uncritical uses of classical regression methods when faced with this kind of data: in these situations, predictions can suffer from systematic disparities, and decisions based on them can further exacerbate the social injustice that skewed the data.

Acknowledgments

We thank Sharad Goel and Michael Zanger-Tishler for their recommendation to use the diabetes example and helpful discussions about results. We also extend our gratitude to Reviewer 2 for their thorough feedback including on technical details. We thank the U.S. Office of Naval Research for partial support of this work.

Conflicts of interest: None declared.

Funding

We thank the Office of Naval Research for grant N000142212648.

Data availability

The data used is publicly available from the National Health and Nutrition Examination Survey ([Centers for Disease Control and Prevention, 2022](#)). Code to reproduce our results is available under <https://github.com/JonasMikhaeil/HierarchicalBayesianMeasurementModels>.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Adcock R., & Collier D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3), 529–546. <https://doi.org/10.1017/S0003055401003100>

Basu A. (2023). Use of race in clinical algorithms. *Science Advances*, 9(21), eadd2704. <https://doi.org/10.1126/sciadv.add2704>

Biderman A. D., & Reiss A. J. (1967). On exploring the ‘dark figure’ of crime. *The Annals of the American Academy of Political and Social Science*, 374(1), 1–15. <https://doi.org/10.1177/000271626737400102>

Buja A., Berk R., Brown L., George E., Pitkin E., Traskin M., Zhao L., & Zhang K. (2016). Models as approximations, part i: A conspiracy of nonlinearity and random regressors in linear regression.

Buja A., Brown L., Kuchibhotla A. K., Berk R., George E., & Zhao L. (2019). Models as approximations ii: A model-free theory of parametric regression. <https://doi.org/10.48550/arXiv.1612.03257>

Bureau of Labor Statistics (2019). National Longitudinal Survey of Youth 1979 cohort, 1979–2016 (rounds 1–27).

Centers for Disease Control and Prevention (2021). National Diabetes Statistics Report, <https://www.cdc.gov/diabetes/data/statistics-report/index.html>.

Centers for Disease Control and Prevention (2022). National Health and Nutrition Examination Survey, <https://www.cdc.gov/nchs/nhanes/index.html>.

Cerdeña J. P., Plaisime M. V., & Tsai J. (2020). From race-based to race-conscious medicine: How anti-racist uprisings call us to act. *The Lancet*, 396(10257), 1125–1128. [https://doi.org/10.1016/S0140-6736\(20\)32076-6](https://doi.org/10.1016/S0140-6736(20)32076-6)

Coots M., Saghafian S., Kent D., & Goel S. (2024). A framework for considering the value of race and ethnicity in estimating disease risk. *Annals of Internal Medicine*. <https://doi.org/10.7326/M23-3166>

Corbett-Davies S., Gaebler J. D., Nilforoshan H., Shroff R., & Goel S. (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312), 1–117.

Davidson K. W., Barry M. J., Mangione C. M., Cabana M., Caughey A. B., Davis E. M., Donahue K. E., Doubeni C. A., Krist A. H., Kubik M., Li L., Ogedegbe G., Owens D. K., Pbert L., Silverstein M., Stevermer J., Tseng C. -W., & Wong J. B. (2021). Screening for prediabetes and type 2 diabetes: US preventive services task force recommendation statement. *Journal of the American Medical Association*, 326(8), 736–743. <https://doi.org/10.1001/jama.2021.12531>

Depaoli S., Winter S. D., & Visser M. (2020). The importance of prior sensitivity analysis in Bayesian statistics: Demonstrations using an interactive shiny app. *Frontiers in Psychology*, 11. <https://doi.org/10.3389/fpsyg.2020.608045>

Diao J. A., Wu G. J., Taylor H. A., Tucker J. K., Powe N. R., Kohane I. S., & Manrai A. K. (2021). Clinical implications of removing race from estimates of kidney function. *JAMA*, 325(2), 184–186. <https://doi.org/10.1001/jama.2020.22124>

Douglas J. D. (1967). *Social meanings of suicide*. Princeton University Press.

Duan D., Kengne A. P., & Echouffo-Tcheugui J. B. (2021). Screening for diabetes and prediabetes and their prediction. *Endocrinology and Metabolism Clinics of North America*, 50(3), 369–385. <https://doi.org/10.1016/j.ecl.2021.05.002>

Dwork C., Hardt M., Pitassi T., Reingold O., & Zemel R. (2012). Fairness through awareness. *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. <https://doi.org/10.1145/2090236.2090255>

Eneanya N. D., Yang W., & Reese P. P. (2019). Reconsidering the consequences of using race to estimate kidney function. *Journal of the American Medical Association*, 322(2), 113–114. <https://doi.org/10.1001/jama.2019.5774>

Fang M., Wang D., Coresh J., & Selvin E. (2022). Undiagnosed diabetes in U.S. adults: Prevalence and trends. *Diabetes Care*, 45(9), 1994–2002. <https://doi.org/10.2337/dc22-0242>

Fogliato R., Chouldechova A., & G'Sell M. (2020). Fairness evaluation in presence of biased noisy labels. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*.

Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., & Rubin D. B. (2014). *Bayesian data analysis* (3rd ed.). CRC Press.

Gelman A., & Hennig C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 180(4), 967–1033. <https://doi.org/10.1111/rssa.12276>

Gelman A., Meng X.-L., & Stern H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4), 733–760. <http://www.jstor.org/stable/24306036>

Goel S., Perelman M., Shroff R., & Sklansky D. A. (2017). Combatting police discrimination in the age of big data. *New Criminal Law Review: An International and Interdisciplinary Journal*, 20(2), 181–232. <https://doi.org/10.1525/nclr.2017.20.2.181>

Gustafson P. (2015). *Bayesian inference for partially identified models: Exploring the limits of limited data*. Chapman and Hall/CRC.

Hardt M., Price E., Price E., & Srebro N. (2016). Equality of opportunity in supervised learning. *30th Conference on Neural Information Processing Systems (NIPS 2016)*.

Hinton E. K. (2016). *From the war on poverty to the war on crime: The making of mass incarceration in America*. Harvard University Press.

Jiang H., & Nachum O. (2020). Identifying and correcting label bias in machine learning. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Kallioinen N., Paananen T., Bürkner P.-C., & Vehtari A. (2024). Detecting and diagnosing prior and likelihood sensitivity with power-scaling. *Statistics and Computing*, 34(1). <https://doi.org/10.1007/s11222-023-10366-5>

Knox D., Lucas C., & Cho W. K. T. (2022). Testing causal theories with learned proxies. *Annual Review of Political Science*, 25(1), 419–441. <https://doi.org/10.1146/polisci.2022.25.issue-1>

Mullainathan S., & Obermeyer Z. (2021). On the inequity of predicting a while hoping for b. *AEA Papers and Proceedings*, 111, 37–42. <https://doi.org/10.1257/pandp.20211078>

Obermeyer Z., Powers B., Vogeli C., & Mullainathan S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>

Parastouei K., Sepandi M., & Eskandari E. (2021). Predicting the 10-year risk of cardiovascular diseases and its relation to healthy diet indicator in Iranian military personnel. *BMC Cardiovascular Disorders*, 21(1), 419. <https://doi.org/10.1186/s12872-021-02231-y>

Richardson T. S., Evans R. J., & Robins J. M. (2011). Transparent parametrizations of models for potential outcomes. *Bayesian Statistics*, 9, 569–610. <https://doi.org/10.1093/acprof:oso/9780199694587.003.0019>

Rosset S., & Tibshirani R. J. (2020). From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 115(529), 138–151. <https://doi.org/10.1080/01621459.2018.1424632>

Rothblum G. N., & Yona G. (2023). Decision-making under miscalibration. *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*. <https://doi.org/10.4230/LIPIcs.ITCS.2023.92>

Rubin D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4), 1151–1172. <https://doi.org/10.1214/aos/1176346785>

Stan Development Team (2023). Stan Modeling Language Users Guide and Reference Manual, version 2.33.

Starr P. (1987). The sociology of official statistics. In A. Alonso, P. Starr (Eds.), *The Politics of Numbers*. Russell Sage Foundation.

Van Calster B., & Vickers A. J. (2015). Calibration of risk prediction models: Impact on decision-analytic performance. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 35(2), 162–169. <https://doi.org/10.1177/0272989X14547233>

Wang J., Liu Y., & Levy C. (2021). Fair classification with group-dependent label noise. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3442188.344591>

Wooldridge J. M. (2010). *Econometric analysis of cross section and panel data*. MIT Press.

Zanger-Tishler M., Nyarko J., & Goel S. (2024). Risk scores, label bias, and everything but the kitchen sink. *Science Advances*, 10(13), eadi8411. <https://doi.org/10.1126/sciadv.adi8411>