Risk Scores, Label Bias, and Everything but the Kitchen Sink

Michael Zanger-Tishler
Harvard University

Julian Nyarko
Stanford University

Sharad Goel
Harvard University

In designing risk assessment algorithms, many scholars promote a "kitchen sink" approach, reasoning that more information yields more accurate predictions. We show, however, that this rationale often fails when algorithms are trained to predict a proxy of the true outcome, as is typically the case. With such "label bias", one should exclude a feature if its correlation with the proxy and its correlation with the true outcome have opposite signs, conditional on the other model features. This criterion is often satisfied when a feature is weakly correlated with the true outcome, and, additionally, that feature and the true outcome are both direct causes of the remaining features. For example, due to patterns of police deployment, criminal behavior and geography may be weakly correlated and direct causes of one's criminal record, suggesting one should exclude geography in criminal risk assessments trained to predict arrest as a proxy for behavior.

1 Introduction

Risk assessments are central to the allocation of resources and the imposition of sanctions. In medicine, estimated health risks guide treatment decisions (*1*); in banking, default risk determines whether an applicant should be granted a loan (*2*); in education, the risk of non-completion is an important factor for college admissions decisions (*3*); and in criminal justice, recidivism risk helps judges decide whether to detain or release a defendant while their cases proceed (*4–6*). Increasingly, the risk of such adverse events is estimated with the help of statistical algorithms. In training these algorithms, there is a widely shared view that the investigator should use as much data as is available to them (*7–9*). This view rests on the intuition that more information leads to (weakly) better predictions: If the added data are informative in estimating risk, then they will improve the performance of the algorithm, and if the added data do not contain a helpful signal, then they will be discarded without hurting performance. Proponents of this view stress that feature importance in the predictive context neither requires nor implies a causal link between algorithmic inputs and predicted outcomes (*8*). Absent the constraints of rigorous causal identification, it is argued that investigators can remain entirely atheoretical and simply hand all available data over to the predictive algorithm.

In this paper, we show how "label bias", present in virtually all real-world scenarios in which algorithms are deployed today, can invalidate this common rationale. Label bias occurs when the outcome of

interest is not observed directly, but is instead observed with measurement error. For instance, although criminal risk assessment tools seek to estimate the risk of future criminal *behavior*, we typically only observe whether individuals are *arrested* or *convicted* of a crime. Similarly, tools to estimate health risk often seek to divert resources to the patients with the most significant medical *needs*, but our observations are often limited to medical *expenditures*. The inclusion of additional features will in general improve an algorithm's prediction of the proxy label (e.g., arrest or medical expenditures), but in the presence of label bias, the additional information can decrease the quality of predictions for the true label (e.g, criminal behavior or medical need). Below, we formally demonstrate and empirically illustrate conditions under which the inclusion of additional features hurts the predictive performance on the true outcome of interest. Because researchers rarely have access to the true label, whether or not to include a particular feature often rests on unverifiable assumptions about the relationships that gave rise to the proxy label. The findings highlight that most predictive contexts require investigators to spend significant time and care in developing a theoretical model of the underlying data generating process, thus removing one of the most important differentiators between prediction and causal inference.

Our study contributes to a burgeoning literature examining the use of algorithmic risk prediction in a variety of domains. These algorithms are frequently used to predict the risk of adverse events such as future criminal offending and failure to appear in court (*10*), the risk of child abuse (*11–13*), money laundering (*14*), students lagging behind in their learning (*15*), and the risk of non-payment of loans (*2*). They are also used in situations where organizations or governments are deciding how to allocate scarce resources such as providing building permits (*16*), assigning students to schools (*17*), assigning high-risk patients to programs providing them more care (*18*), and determining who will receive kidney transplants (*19*). Further, corporations are currently using these tools to inform decisions about who receives information about housing advertisements (*20*) and employment opportunities (*21*). Algorithmic risk assessment tools can be better than humans at determining risk (*22*). However, scholars continue to critique these algorithms and study whether and under what conditions they can fairly and effectively be deployed in society (*23–26*).

In addition, our analysis builds on and contributes to a substantial body of literature examining the impact of label bias in statistical analyses. Prior work in the social sciences has long focused on the importance of measurement error for causal studies. Within this literature, a main focus has traditionally been on examining the importance of measurement error in the *independent* variable, which can, at best, attenuate the causal estimates (*27, pp. 320–323*), and at worst, bias the coefficients in ways that are difficult to predict (*28*). Less attention has been given to label bias (i.e., measurement error in the dependent variable), perhaps because it is often assumed that proxy labels differ from the true labels by random noise, in which case one can still obtain unbiased causal estimates (*27, pp. 318–320*). Existing research, however, suggests that there is a non-random relationship between the true and proxy labels across a variety of contexts, such as in the case of arrest and offending (*29*). More recent contributions have considered the impact of such systematic errors in the labels. For example, Knox et al. (*30*) examine the potential for biases to arise in causal estimates when latent concepts that cannot be directly measured—like political "ideology" and "democracy"—are approximated by proxy variables constructed from statistical models. Complementary work in computer science has examined the impact of label bias in a predictive setting. For instance, although predictive models may perform well on the proxy label, research has shown they are not guaranteed to be accurate on the true label if the measurement error between the true and proxy label is non-random (*31*). Similarly, label bias can also reduce the fairness of these algorithms on the true label (*32*). When feasible, training predictions on the true label

rather than a proxy has been shown to reduce racial inequalities in algorithmic prediction and increase accuracy (*18, 33, 34*).

We build on these contributions by explicitly examining how the performance decrease from label bias interacts with the inclusion of additional predictors into the model. To establish our results, we begin, in Section 2, by deriving analytic conditions for when excluding factors in a model trained to predict a proxy label is guaranteed to improve predictions of the true outcome of interest. We demonstrate and build intuition for these analytic results using a stylized example of estimating recidivism risk in the presence of label bias, where reoffense is the true label of interest and rearrest is the observed proxy. Then, in Section 3, we turn to two case studies. First, we consider partially synthetic recidivism data with real rearrest outcomes (the proxy label) and simulated reoffense outcomes (the true label). This setting resembles one that many researchers face in practice, where data on the true label are often prohibitively difficult or impossible to obtain. We show how different assumptions about how the true label relates to the observed proxy affect decisions about what predictors to include in the risk assessment model. Second, we consider a dataset from the health sciences. In targeting patients for high-risk care management programs, we rely on data by Obermeyer et al. (*18*) which contain, among other items, information on both the true label (healthcare need) and a proxy (healthcare spending). Using this dataset, we estimate the welfare costs of using a kitchen-sink predictive model instead of more judiciously selecting a model that accounts for label bias. We conclude in Section 4 with a discussion of our findings and point out potential paths forward.

2 A Statistical Condition for Excluding Features

To build intuition for how label bias impacts the choice of features in predictive models, we start with a simplified motivating example from the criminal justice context. In the United States, after an arrest, a judge will often decide whether or not to detain the arrested individual based on their estimated risk to public safety. In practice, this risk is commonly estimated using statistical risk assessments. The underlying risk models are trained using information about future arrests and convictions. However, arrests and convictions are not direct measures of public safety risks. Instead, they merely act as proxies, making these risk assessment tools susceptible to label bias.

In Figure 1, we sketch the data-generating process for a stylized, linear structural equation model (SEM) (*35*) of arrests and behavior, where we treat arrests as the observed proxy for unobserved behavior, our true outcome of interest. The model produces synthetic data on individual-level behavior (B_0 and B_1) and arrest (A_0 and A_1) outcomes at two time periods ($t = 0$ and $t = 1$), as well as the neighborhood (Z) in which the individual resides. Importantly, arrests depend both on behavior and on neighborhood, reflecting the fact that people who engage in the same behavior may be arrested at different rates depending on where they live. For example, Beckett et al. (*36*) found that the geographic concentration of police resources in Seattle led to higher arrest rates for Black individuals delivering drugs compared to white individuals delivering drugs—where the true racial distribution of those delivering drugs was estimated from survey data and ethnographic observations. Similarly, Cai et al. (*37*) found that the issuance of speeding tickets varied across neighborhoods even after adjusting for the true, underlying incidence of speeding, as estimated by the movement of mobile phones.

In this SEM, all of the variables are normally distributed, with mean 0 and variance 1. We can thus interpret their values as representing the extent to which individuals differ from the population averages.

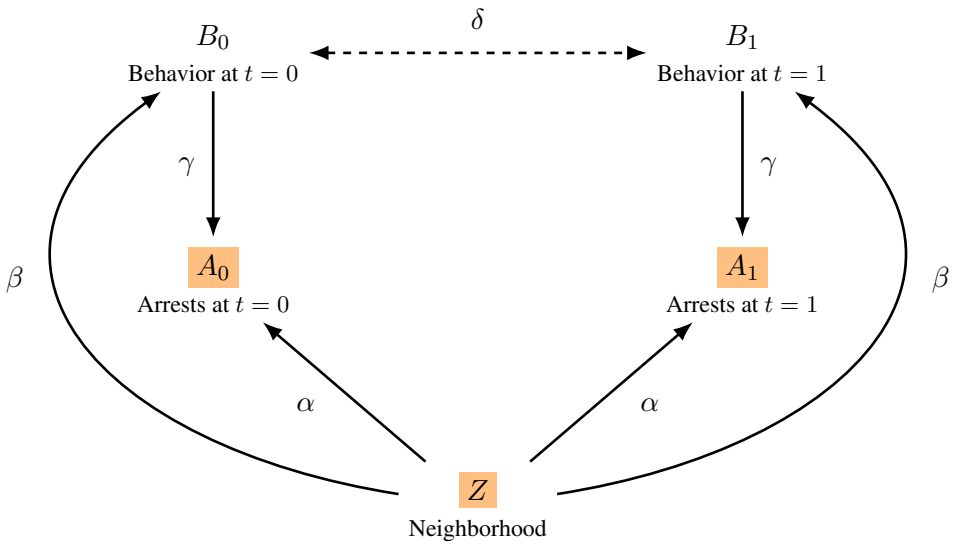

Figure 1: *The data-generating process for our stylized example of criminal behavior (true label) and arrest (proxy label), with observed variables highlighted in orange.*

In the case of neighborhood (Z), we can think of its value as denoting the level of police enforcement in that area. Further details about the model are provided in the Appendix.

Using synthetic data generated with this SEM, we train a "complex", kitchen-sink model to predict arrests at time $t = 1$ (A_1) based on arrests at time $t = 0$ (A_0) and neighborhood (Z). The more parsimonious, "simple" model bases its predictions only on arrests at time $t = 0$, omitting neighborhood. We now examine how the performance of the complex and simple models vary for different values of β, the parameter that describes the relationship between neighborhood and behavior, holding the other parameters fixed.[1] Across values of β, the left-hand panel of Figure 2 shows that the complex model outperforms the simple model—in terms of root mean squared error (RMSE)—when evaluated on the proxy label. As expected, including more information reduces error when evaluated on the label used to train the models, a pattern that has traditionally motivated the inclusion of more features in predictive models. However, moving to the right-hand panel of Figure 2, we see that the simple model outperforms the complex model on the *true* label for some values of β. In particular, the simple model outperforms the complex one for small values of β, corresponding to a weak relationship between neighborhood and behavior.

Our SEM illustrates a scenario in which simple models outperform more complex models due to the presence of label bias. To understand this result, imagine two individuals, both of whom have the same prior arrest record, but with only one of them living in a heavily policed neighborhood. Further assume that where one lives has little impact on criminal behavior (corresponding to small β), but that heavier policing increases the chance of being arrested for an offense. In this case, we can infer that the

[1]For this simulation, we set $\alpha = \gamma = \delta = 0.4$, though the general pattern is largely invariant to this choice, as we describe in more detail below.

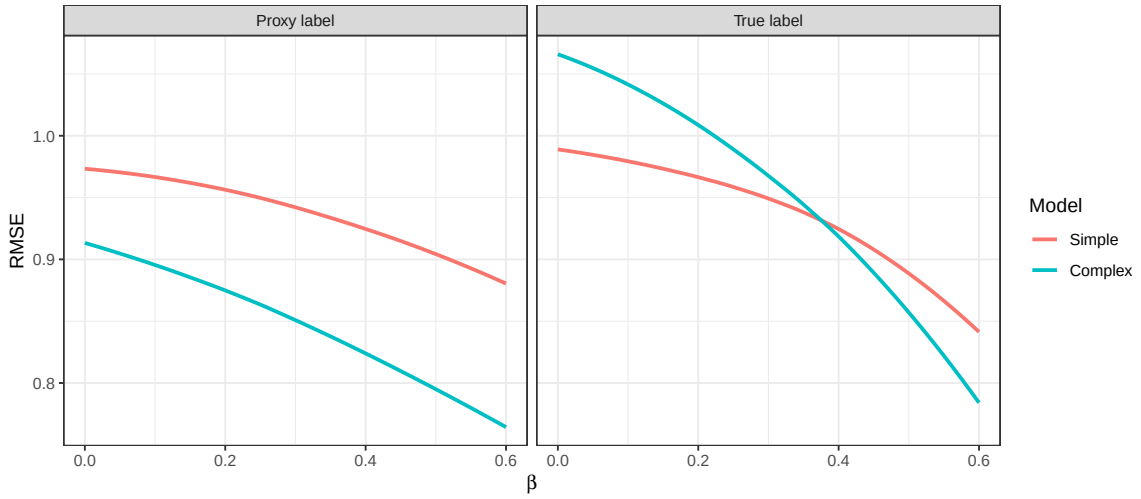

Figure 2: *Performance of simple and complex models trained to predict a proxy label, when evaluated on the proxy label (left) and the true label (right) for a range of β values. Whereas the complex model outperforms the simple model on the proxy label, the simple model outperforms the complex model on the true label for certain values of β.*

individual living in the heavily policed neighborhood engaged in past criminal activity less frequently than the individual living in the less heavily policed neighborhood. This is because fewer actual offenses are required to build a given arrest record in areas of high enforcement. Extrapolating from their past behavior, we would accordingly expect the individual in the heavily policed area to be less likely to engage in future criminal behavior. Thus, using information about one's neighborhood to predict future arrests (the proxy label) correctly tells us that the individual living in the heavily policed neighborhood is more likely to be rearrested, but it incorrectly suggests that individual is also more likely to engage in future criminal behavior (the true label). So, when predicting arrests as a proxy for behavior, it is better in this case to exclude information on one's neighborhood.

The SEM depicts a specific data-generating process, but the phenomenon we identify is generalizable. Theorem 1 and Corollary 1 below establish formal conditions under which this pattern is guaranteed to occur. Proofs for both results are in the Appendix.

Theorem 1 *Suppose Y and Y' are two arbitrary random variables with finite variance, where Y is the "true" outcome of interest and Y' is a proxy. For a random vector $X = (X_1, \ldots, X_k)$ and a random vector $Z = (Z_1, \ldots, Z_\ell)$, consider the estimators*

$$\hat{Y}_{X,Z} = \mathbb{E}[Y' \mid X, Z], \text{ and}$$
$$\hat{Y}_X = \mathbb{E}[Y' \mid X],$$

where $\hat{Y}_{X,Z}$ is the "complex" estimator that uses all available features, and $\hat{Y}_X$ is the "simple" estimator that omits Z. If $\hat{Y}_{X,Z} \neq \hat{Y}_X$ and

$$\mathbb{E}\left[\text{Cov}\left(\hat{Y}_{X,Z}, Y \mid X\right)\right] \leq 0,$$

then $\mathbb{E}\left[\left(\hat{Y}_X - Y\right)^2\right] < \mathbb{E}\left[\left(\hat{Y}_{X,Z} - Y\right)^2\right]$*, meaning the simple estimator outperforms the complex estimator.*

In the setting of Theorem 1, one seeks to estimate a true outcome of interest Y, and is choosing between two different estimators designed to predict the proxy label Y'. The first, "complex" estimator ($\hat{Y}_{X,Z}$) uses both X and Z to predict Y', whereas the second ($\hat{Y}_X$) uses only X. The theorem shows that if, conditional on X, the true label (Y) is negatively correlated with the complex estimator ($\hat{Y}_{X,Z}$), then the simple model generally outperforms the complex estimator on the true outcome of interest. Intuitively, this result holds because the condition of the theorem means that the complex estimator goes in the "wrong" direction relative to the true outcome of interest.

If, alternatively, the true and proxy labels differ only by additive, independent noise, then Proposition 1 in the Appendix shows that including more information when predicting the proxy label will in general improve predictive performance on the true label. Thus, in the absence of systematic measurement error—including the case where there is no measurement error—that result confirms the conventional wisdom that more information is better.

To build further insight into this result, we consider the case where $\ell = 1$ (i.e., Z is a single random variable) and the complex estimator $\hat{Y}_{X,Z}$ is linear in Z. In this setting, Corollary 1 establishes a simpler condition under which performance increases by omitting information. Specifically, if, conditional on X, Z is positively correlated with true label Y but negatively correlated with the proxy label Y' (or vice versa), then omitting Z when predicting the proxy label will in general improve performance on the true outcome of interest.

Corollary 1 *Consider the setting of Theorem 1 with* $\ell = 1$*. Suppose additionally that* Z *has finite variance and* $\hat{Y}_{X,Z}$ *is linear in* Z*, i.e.,* $\hat{Y}_{X,Z} = f(X) + cZ$ *for some function* f *and a constant* $c \in \mathbb{R}$*. If* $\hat{Y}_{X,Z} \neq \hat{Y}_X$ *and either* $\mathbb{E}\left[\mathrm{Cov}\left(Y, Z \mid X\right)\right] = 0$ *or*

$$\mathrm{sign}\left(\mathbb{E}\left[\mathrm{Cov}\left(Y, Z \mid X\right)\right]\right) = -\mathrm{sign}\left(\mathbb{E}\left[\mathrm{Cov}\left(Y', Z \mid X\right)\right]\right),$$

then $\mathbb{E}\left[\left(\hat{Y}_X - Y\right)^2\right] < \mathbb{E}\left[\left(\hat{Y}_{X,Z} - Y\right)^2\right]$.

The linearity assumption of Corollary 1 holds in a variety of settings. In particular, as described in the Appendix, it holds when Y', X, and Z are jointly multivariate normal, as is the case in our SEM above. To apply the corollary, one needs information on the correlations of Y and Z and of Y' and Z, conditional on X. The former involves directly observed quantities—the proxy label and the potential predictors—and so, in practice, can be computed from data. For our stylized SEM, we show in the Appendix that this correlation is positive for all (non-degenerate) parameter choices, meaning that neighborhood (Z) is positively correlated with future arrests (A_1), conditional on past arrests (A_0). The second conditional correlation we must consider when applying Corollary 1—the correlation between Y and Z, conditional on X—is not typically directly observed, as it depends on the true label Y. Understanding its sign thus involves assumptions about how the true label is related to the predictors Z and X. For our SEM, we show in the Appendix that this correlation is negative for small values of β. That is, when β is small, neighborhood (Z) and future behavior (B_1) are negatively correlated conditional on past arrests (A_0). Intuitively, this is because A_0 is a collider, and so when we fix its value, increasing Z requires decreasing B_0, which in turn decreases B_1. Thus, for small values of β, omitting neighborhood when predicting the proxy label improves performance on the true label, as shown in Figure 3.

3 Case Studies

To better understand the practical implications of our results, we now turn to two real-world datasets. The first allows us to further consider criminal risk assessments, adding additional realism to our stylized SEM above; the second dataset comes from the medical domain, where the goal of the risk assessment we consider is to identify patients with complex healthcare needs.

3.1 Criminal risk assessments

Continuing with our running example studying arrest and criminal behavior, we use data on individuals from a major U.S. county who were arrested for a felony offense between 2013 and 2019. For simplicity, we limit the sample to the 25,918 cases where the individuals' race was identified as either Black or non-Hispanic white. The dataset includes further details on each case, including information on the charges, the location, date and time of the incident, and the criminal history of the arrested individual. In addition, the dataset contains information on future rearrests, which we use as our proxy label for future offenses. Using these data, we fit simple and complex models trained on the proxy label (future arrests). We then examine model performance on the true label (future criminal offenses, which we simulate, as described below, since they are not directly observed). Our "complex" model includes three features: the age of the arrested individual; the number of times the individual was previously arrested; and whether or not the arrest occurred in a "high policing" area (i.e., a police district accounting for disproportionately high numbers of arrests). Our "simple" model includes age and number of past arrests, but not location information—similar to many commonly used criminal risk assessment tools.

This example mirrors many instances of label bias in the real world, as it is difficult, and perhaps impossible, to directly estimate the risk of "true" offending (*38*). This is in part because criminal behavior that is not reported to the police will not be included in administrative records. We thus simulate offending outcomes under a range of data-generating processes, and then examine how assumptions about criminal behavior affect model performance after including or omitting location information. In particular, for a fixed value of $\rho \in \mathbb{R}$, describing the impact of neighborhood on criminal behavior, we assume that each individual in our dataset commits a future offense with the following probability:

$$\Pr(B_1 = 1) = \text{logit}^{-1}\left(-1 - \frac{1}{100}X_{\text{age}} + \frac{1}{2}A_0 + \rho Z\right),$$

where B_1 indicates future criminal behavior (our true label), X_{age} is the arrested individual's age, A_0 is the number of times they were previously arrested, and Z indicates whether the arrest took place in a high-policing area. The intercept and the coefficients for A_0 and X_{age} were selected to approximate the coefficients from a regression of future arrests on age and past arrests in our data.

Based on the data-generating process described above, we now evaluate the ability of our simple and complex risk assessment models to predict the synthetic true label, future criminal behavior. We evaluate model performance in terms of AUC, as the outcome is binary.[2] Figure 3 shows that the simple model

[2] AUC is a common measure of performance in the machine learning community when considering binary outcomes. Given a random individual who engaged in future criminal activity and a random individual who did not, the AUC of a risk assessment model is the probability that the model correctly identifies the individual in the pair who engaged in criminal activity. Our formal theoretical results are stated in terms of RMSE, but this example and our subsequent example show that the general pattern and intuition extend to other popular evaluation metrics.

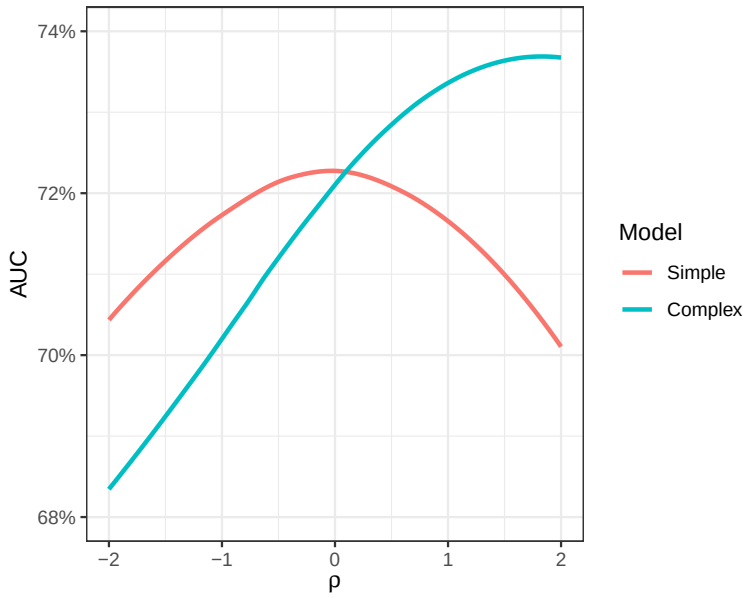

Figure 3: *Performance of simple (age and past arrests) and complex (age, past arrests, and neighborhood) models trained to predict future arrests (the proxy label), evaluated on future criminal behavior (the true label). Because the future criminal behavior is not directly observable, the plot shows results for synthetic outcomes generated under a range of data-generating processes parameterized by ρ, the hypothesized relationship between neighborhood and future criminal behavior.*

outperforms the complex model on the true label when ρ is negative, and the complex model outperforms the simple model when ρ is sufficiently positive. Given two arrested individuals who are the same age and have the same number of past arrests, negative values of ρ indicate that the individual who was arrested in the high-policing area is the less likely of the pair to engage in future criminal behavior. That pattern is akin to what we saw in our stylized SEM depicted in Figure 1. Accordingly, to the extent that one believes the hypothesized data-generating process with negative ρ is a sufficiently accurate description of criminal behavior, it is better to exclude neighborhood information when training risk assessment tools on the proxy label, future arrests.

3.2 Identifying high-needs patients

We continue by applying our results to a well-known case of label bias in the literature, that of a commercial risk assessment tool that health systems rely on to target patients for "high-risk care management" programs (*18*). These programs seek to enroll patients with complex medical needs, and subsequently provide them with a higher level of care. Because these programs are capacity constrained, the role of statistical risk assessments in this case is to accurately identify patients who would benefit the most from the additional care. In practice, though, the risk assessment algorithms are often designed to predict future medical expenditures, a proxy for medical need as the true outcome of interest. Analyzing these algorithms, Obermeyer et al. (*18*) conclude that, due to label bias, Black patients are less likely to be enrolled in the program than white patients with the same level of medical need. This is because un-

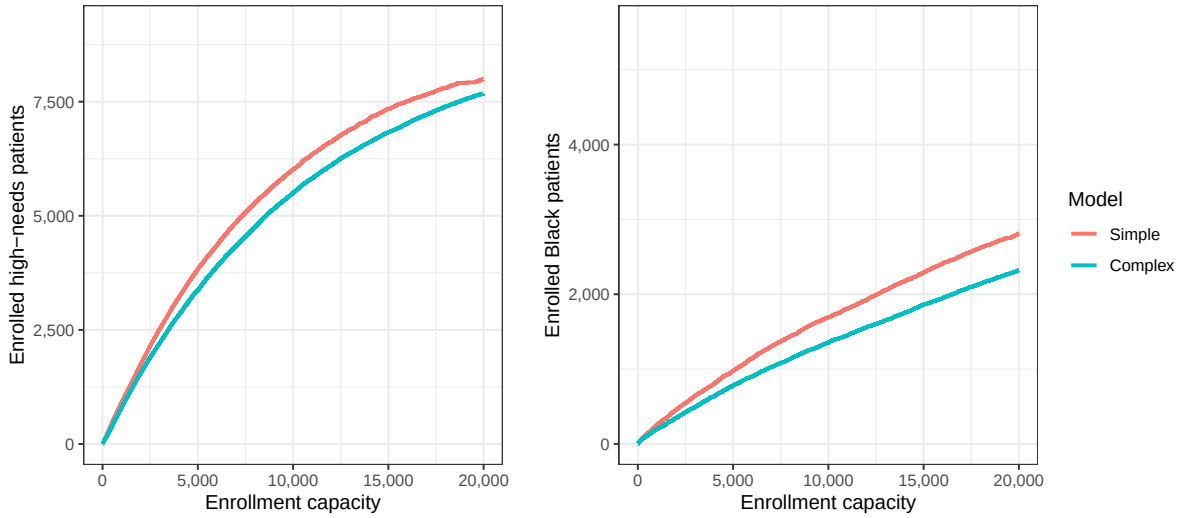

Figure 4: *Enrollment of high needs patients (left) and demographic composition of enrolled patients (right) under the simple and complex models for a range of program capacities.*

equal access to healthcare means that white individuals are more likely to seek medical treatment—and accordingly incur higher medical costs—than equally sick racial minorities.

Obermeyer et al. (*18*) highlight the importance of appropriately selecting the target of prediction, and illustrate the accuracy and equity gains one can achieve by switching from predicting expenditures to a more direct measure of medical need. Here we revisit the problem, and investigate how the choice of risk factors used to identify patients impacts enrollment decisions. To do so, we start with the data released by Obermeyer et al. (*18*), which include detailed information on patient demographics (sex, race, and age), current and future health, and past and future medical expenditures.[3] We then train simple and complex models on the proxy label, future medical costs. Our complex model includes all information available at the time of the enrollment decision (i.e., patient demographics, current health, and past medical expenditures); our simple model includes only current health, excluding past medical costs and demographic variables. In the end, the complex model includes 150 predictors, and the simple model includes 128 predictors.

Finally, we evaluate both models on their ability to predict whether a patient, in the subsequent year, is found to suffer from at least three chronic diseases—a measure of future health need identified by Obermeyer et al. (*18*). The left-hand panel of Figure 4 shows the number of high-needs patients enrolled under the simple and complex models at different enrollment capacities, where the patients with highest estimated risk under the respective models are enrolled in the program. At each capacity level, the simple model outperforms the complex model in identifying more high-needs patients. Additionally, as shown in the right-hand panel of Figure 4, the simple model enrolls more Black patients than the complex model at every capacity level. This pattern stems from the simple model prioritizing patients with high expected medical needs over patients with high expected medical expenditures—the latter population

[3]Obermeyer et al. (*18*) released a synthetic dataset, with variables having the same conditional distributions as those in the original dataset, using the `synthpop` package in R (the original data cannot be released in order to protect patient privacy). The data are available at: https://gitlab.com/labsysmed/dissecting-bias).

being disproportionately white. Thus, if one only has access to a proxy label, systematically excluding predictors in a risk assessment tool can improve both the accuracy and equity of the instrument.

4 Conclusion

In building predictive models, the traditional guidance is to include all available information to maximize performance. But, as we have shown, a more judicious selection of features can lead to better model performance in the presence of label bias. Because the true label of interest is often not readily available, it raises the question of what examiners should and can do to mitigate the negative consequences from taking a kitchen-sink approach to prediction. The examples we have discussed highlight several approaches that vary in their appropriateness based on data availability and understanding of the underlying data-generating process.

Most directly, Obermeyer et al. (*18*) illustrate how some instances of label bias can be addressed simply by making a more concentrated effort to collect data on the true label of interest. If such an effort is generally possible, but prohibitively costly, investigators should consider whether the true label of interest can be obtained for a smaller subset of the population. This subset, even if it is not sufficiently large to train models predicting the true label, might still be used to explore how the selection of features affects model performance on the true label. If obtaining the true label is impossible, but investigators have access to a wealth of other features, one may simulate the true label of interest. In doing so, researchers should use their domain-specific knowledge to make reasonable assumptions about the relationship between the true label of interest and the features in question. We illustrated this process using felony offense data. Importantly, investigators need not constrain themselves to one particular relationship between the true label and the features, but can instead assess the sensitivity of feature selection to label bias across a wide range of plausible assumptions. Finally, investigators can make additional theoretical assumptions about the data-generating process in order to determine how label bias affects the choice of risk factors in a specific application—as we did in our healthcare example. As shown in that example, caution is particularly warranted for features that do not appear to be directly risk relevant. These features often yield little improvement on the true outcome of interest, and raise the likelihood that performance may decrease or that their inclusion may exacerbate disparities.

More generally, our findings suggest, in contrast to conventional wisdom, that one cannot entirely divorce the predictive enterprise from theoretical considerations. Instead, a successful deployment of predictive tools often rests on the plausibility of the assumptions about the underlying processes that give rise to the observed data, highlighting the continued utility of domain-specific expertise in the predictive context.

References

1. Sendhil Mullainathan and Ziad Obermeyer. Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics*, 137(2):679–727, 2022.

2. Martin Leo, Suneel Sharma, and Koilakuntla Maddulety. Machine learning in banking risk management: A literature review. *Risks*, 7(1):29, 2019.

3. Lovenoor Aulck, Dev Nambi, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. Mining university registrar records to predict first-year undergraduate attrition. *International Educational Data Mining Society*, 2019.

4. Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

5. Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018.

6. Zhiyuan Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. The limits of human predictions of recidivism. *Science Advances*, 6(7):eaaz0652, 2020.

7. Richard Berk. *Machine Learning Risk Assessments in Criminal Justice Settings*. Springer, 2019.

8. Charles F Manski. Patient-centered appraisal of race-free clinical risk assessment. *Health Economics*, 31(10):2109–2114, 2022.

9. Charles F Manski, John Mullahy, and Atheendar Venkataramani. Using measures of race to make clinical predictions: Decision making, patient health, and fairness. Technical report, National Bureau of Economic Research, 2022.

10. Kosuke Imai, Zhichao Jiang, D James Greiner, Ryan Halen, and Sooahn Shin. Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *arXiv preprint arXiv:2012.02845*, 2022.

11. Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

12. Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

13. Ravi Shroff. Predictive analytics for city agencies: Lessons from children's services. *Big Data*, 5(3):189–196, 2017.

14. Yan Zhang and Peter Trubey. Machine learning and sampling scheme: An empirical study of money laundering detection. *Computational Economics*, 54:1043–1063, 2019.

15. Lindsay Cattell and Julie Bruch. Identifying students at risk using prior performance versus a machine learning algorithm. Technical report, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic, 2021.

16. Viktor Mayer-Schönberger and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt, 2013.

17. Maxwell Allman, Itai Ashlagi, Irene Lo, Juliette Love, Katherine Mentzer, Lulabel Ruiz-Setz, and Henry O'Connell. Designing school choice for diversity in the San Francisco Unified School District. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 290–291, 2022.

18. Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

19. John J Friedewald, Ciara J Samana, Bertram L Kasiske, Ajay K Israni, Darren Stewart, Wida Cherikh, and Richard N Formica. The kidney allocation system. *Surgical Clinics*, 93(6):1395–1406, 2013.

20. Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. Potential for discrimination in online targeted advertising. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 5–19. PMLR, 23–24 Feb 2018.

21. Anja Lambrecht and Catherine Tucker. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 65(7):2966–2981, 2019.

22. Sharad Goel, Ravi Shroff, Jennifer Skeem, and Christopher Slobogin. The accuracy, equity, and jurisprudence of criminal risk assessment. In *Research Handbook on Big Data Law*, pages 9–28. Edward Elgar Publishing, 2021.

23. Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *Available at SSRN*, 2022.

24. Sam Corbett-Davies, Johann Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *arXiv preprint arXiv:1808.00023*, 2023.

25. Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

26. Alex Chohlas-Wood, Madison Coots, Sharad Goel, and Julian Nyarko. Designing equitable algorithms. *arXiv preprint arXiv:2302.09157*, 2023.

27. Jeffrey M Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, 2015.

28. Aaron Chalfin and Justin McCrary. Are U.S. cities underpoliced? Theory and evidence. *Review of Economics and Statistics*, 100(1):167–186, 2018.

29. Riccardo Fogliato, Alice Xiang, Zachary Lipton, Daniel Nagin, and Alexandra Chouldechova. On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 100–111, 2021.

30. Dean Knox, Christopher Lucas, and Wendy K Tam Cho. Testing causal theories with learned proxies. *Annual Review of Political Science*, 25:419–441, 2022.

31. Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparency*, pages 526–536, 2021.

32. Riccardo Fogliato, Alexandra Chouldechova, and Max G'Sell. Fairness evaluation in presence of biased noisy labels. In *International Conference on Artificial Intelligence and Statistics*, pages 2325–2336. PMLR, 2020.

33. Emma Pierson, David M Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1):136–140, 2021.

34. Sendhil Mullainathan and Ziad Obermeyer. On the inequity of predicting a while hoping for b. In *AEA Papers and Proceedings*, volume 111, pages 37–42, 2021.

35. Judea Pearl. Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference*, 1(1):155–170, 2013.

36. Katherine Beckett, Kris Nyrop, and Lori Pfingst. Race, drugs, and policing: Understanding disparities in drug delivery arrests. *Criminology*, 44(1):105–137, 2006.

37. William Cai, Johann Gaebler, Justin Kaashoek, Lisa Pinals, Samuel Madden, and Sharad Goel. Measuring racial and ethnic disparities in traffic enforcement with large-scale telematics data. *PNAS Nexus*, 1(4):pgac144, 2022.

38. Albert D Biderman and Albert J Reiss Jr. On exploring the "dark figure" of crime. *The Annals of the American Academy of Political and Social Science*, 374(1):1–15, 1967.

39. Sewall Wright. Systems of mating. I. The biometric relations between parent and offspring. *Genetics*, 6(2):111, 1921.

40. Morris L Eaton. *Multivariate Statistics: A Vector Space Approach.* John Wiley & Sons, Inc., 1983.

Acknowledgements

We thank Avi Feller, Johann Gaebler, Talia Gillis, and Josh Grossman for helpful feedback and conversations.

A Proof of Theorem 1

We prove a generalization of the result, namely that,

$$\mathbb{E}\left[\left(\hat{Y}_{X,Z}-Y\right)^2\right]-\mathbb{E}\left[\left(\hat{Y}_X-Y\right)^2\right]=\mathrm{Var}\left(\hat{Y}_{X,Z}\right)-\mathrm{Var}\left(\hat{Y}_X\right)-2\mathbb{E}\left[\mathrm{Cov}\left(\hat{Y}_{X,Z},Y\mid X\right)\right], \quad (1)$$

and, assuming $\hat{Y}_{X,Z}\neq\hat{Y}_X$,

$$\mathrm{Var}\left(\hat{Y}_{X,Z}\right)-\mathrm{Var}\left(\hat{Y}_X\right)>0. \quad (2)$$

Together, Eqs. (1) and (2) immediately imply the statement of the theorem.

To start, note that for any square-integrable random variable $\hat{Y}$,

$$\mathbb{E}\left[\left(\hat{Y}-Y\right)^2\right]=\mathbb{E}\left[Y^2\right]+\mathbb{E}\left[\hat{Y}^2\right]-2\mathbb{E}\left[Y\cdot\hat{Y}\right].$$

Since Y' is square-integrable by assumption, so are $\hat{Y}_{X,Z}$ and $\hat{Y}_X$ (by the law of total variance), and so,

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{Y}_{X,Z}-Y\right)^2\right]-\mathbb{E}\left[\left(\hat{Y}_X-Y\right)^2\right]\\
&\qquad=\mathbb{E}\left[\hat{Y}_{X,Z}^2\right]-\mathbb{E}\left[\hat{Y}_X^2\right]+2\Big(\mathbb{E}\left[Y\cdot\hat{Y}_X\right]-\mathbb{E}\left[Y\cdot\hat{Y}_{X,Z}\right]\Big)\\
&\qquad=\mathrm{Var}\left(\hat{Y}_{X,Z}\right)-\mathrm{Var}\left(\hat{Y}_X\right)+2\Big(\mathbb{E}\left[Y\cdot\hat{Y}_X\right]-\mathbb{E}\left[Y\cdot\hat{Y}_{X,Z}\right]\Big),
\end{aligned} \quad (3)$$

where the last line follows from the fact that $\mathbb{E}\left[\hat{Y}_{X,Z}\right]=\mathbb{E}\left[\hat{Y}_X\right]=\mathbb{E}[Y']$. Now,

$$\begin{aligned}
\mathbb{E}\left[Y\cdot\hat{Y}_X\right]-\mathbb{E}\left[Y\cdot\hat{Y}_{X,Z}\right]&=\mathbb{E}\left[\mathbb{E}\left[Y\cdot\hat{Y}_X\mid X\right]-\mathbb{E}\left[Y\cdot\hat{Y}_{X,Z}\mid X\right]\right]\\
&=\mathbb{E}\left[\hat{Y}_X\cdot\mathbb{E}\left[Y\mid X\right]-\mathbb{E}\left[Y\cdot\hat{Y}_{X,Z}\mid X\right]\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[\hat{Y}_{X,Z}\mid X\right]\cdot\mathbb{E}\left[Y\mid X\right]-\mathbb{E}\left[Y\cdot\hat{Y}_{X,Z}\mid X\right]\right]\\
&=-\mathbb{E}\left[\mathrm{Cov}\left(\hat{Y}_{X,Z},Y\mid X\right)\right],
\end{aligned} \quad (4)$$

where we repeatedly applied the law of iterated expectations, and used the fact that $\hat{Y}_X$ is measurable with respect to X in the second equality. Eqs. (3) and (4) together establish Eq. (1).

Next we prove Eq. (2). First note that

$$\begin{aligned}
\mathbb{E}\left[\hat{Y}_{X,Z}^2\mid X\right]&\geq\left(\mathbb{E}\left[\hat{Y}_{X,Z}\mid X\right]\right)^2\\
&=\left(\mathbb{E}\left[\mathbb{E}[Y'\mid X,Z]\mid X\right]\right)^2\\
&=\left(\mathbb{E}\left[Y'\mid X\right]\right)^2\\
&=\hat{Y}_X^2,
\end{aligned} \quad (5)$$

where the first line follows from Jensen's inequality, and the second equality follows from the law of iterated expectations. As a result, by another application of the law of iterated expectations,

$$\begin{aligned}\mathbb{E}\left[\hat{Y}_{X,Z}^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\hat{Y}_{X,Z}^2 \mid X\right]\right] \\ &\geq \mathbb{E}\left[\hat{Y}_X^2\right],\end{aligned} \tag{6}$$

which in turn implies,

$$\mathrm{Var}\left(\hat{Y}_{X,Z}\right) \geq \mathrm{Var}\left(\hat{Y}_X\right). \tag{7}$$

Further, the inequality in Eq. (5) is strict on the set where $\mathrm{Var}\left(\hat{Y}_{X,Z} \mid X\right) \neq 0$. Suppose, toward a contradiction, that $\mathrm{Var}\left(\hat{Y}_{X,Z} \mid X\right) = 0$ a.s. Then,

$$\begin{aligned}0 &= \mathbb{E}\left[\mathrm{Var}\left(\hat{Y}_{X,Z} \mid X\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\hat{Y}_{X,Z} - \mathbb{E}\left[\hat{Y}_{X,Z} \mid X\right]\right)^2 \mid X\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\hat{Y}_{X,Z} - \hat{Y}_X\right)^2 \mid X\right]\right] \\ &= \mathbb{E}\left[\left(\hat{Y}_{X,Z} - \hat{Y}_X\right)^2\right],\end{aligned}$$

in which case $\hat{Y}_{X,Z} = \hat{Y}_X$, contradicting the assumption of the theorem. Consequently, it must be that $\mathrm{Var}\left(\hat{Y}_{X,Z} \mid X\right) \neq 0$ on a set of positive measure, implying the inequality in Eq. (5) is strict on a set of positive measure. Strict inequality in Eq. (5) implies strict inequality in Eqs. (6) and (7), establishing Eq. (2) and concluding the proof.

□

B Proof of Corollary 1

By Theorem 1, it is sufficient to show that $\mathbb{E}\left[\mathrm{Cov}\left(\hat{Y}_{X,Z}, Y \mid X\right)\right] \leq 0$. We start by noting that

$$\begin{aligned}\mathbb{E}\left[\mathrm{Cov}\left(\hat{Y}_{X,Z}, Y \mid X\right)\right] &= \mathbb{E}\left[\mathrm{Cov}\left(f(X) + cZ, Y \mid X\right)\right] \\ &= c \cdot \mathbb{E}\left[\mathrm{Cov}(Y, Z \mid X)\right].\end{aligned}$$

Now, if $\mathbb{E}[\mathrm{Cov}(Y, Z \mid X)] = 0$, then the result follows immediately. If $\mathbb{E}[\mathrm{Cov}(Y, Z \mid X)] \neq 0$, then by the assumption of the theorem,

$$\mathrm{sign}\left(\mathbb{E}\left[\mathrm{Cov}\left(\hat{Y}_{X,Z}, Y \mid X\right)\right]\right) = -\mathrm{sign}\left(c \cdot \mathbb{E}\left[\mathrm{Cov}\left(Y', Z \mid X\right)\right]\right). \tag{8}$$

Now, by repeatedly applying the law of iterated expectations, we have

$$\begin{aligned}\mathbb{E}\left[Z \cdot Y' \mid X\right] &= \mathbb{E}\left[\mathbb{E}\left[Z \cdot Y' \mid X, Z\right] \mid X\right] \\ &= \mathbb{E}\left[Z \cdot \mathbb{E}\left[Y' \mid X, Z\right] \mid X\right] \\ &= \mathbb{E}\left[Z \cdot \hat{Y}_{X,Z} \mid X\right] \\ &= f(X) \cdot \mathbb{E}[Z \mid X] + c \cdot \mathbb{E}\left[Z^2 \mid X\right].\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbb{E}[Y' \mid X] &= \mathbb{E}[\mathbb{E}[Y' \mid X, Z] \mid X] \\ &= \mathbb{E}[\hat{Y}_{X,Z} \mid X] \\ &= f(X) + c \cdot \mathbb{E}[Z \mid X].\end{aligned}$$

Putting the above together, we get

$$\begin{aligned}\mathrm{Cov}\left(Y', Z \mid X\right) &= \mathbb{E}\left[Z \cdot Y' \mid X\right] - \mathbb{E}[Y' \mid X] \cdot \mathbb{E}[Z \mid X] \\ &= c \cdot \left(\mathbb{E}\left[Z^2 \mid X\right] - \mathbb{E}\left[Z \mid X\right]^2\right) \\ &= c \cdot \mathrm{Var}(Z \mid X).\end{aligned}$$

Finally, by Eq. (8),

$$\begin{aligned}\mathrm{sign}\left(\mathrm{Cov}\left(\hat{Y}_{X,Z}, Y \mid X\right)\right) &= -\mathrm{sign}\left(c^2 \cdot \mathrm{Var}(Z \mid X)\right) \\ &\leq 0,\end{aligned}$$

establishing the result.

□

C Kitchen-Sink Models and Independent Noise

When the proxy label Y' and the true label Y simply differ by additive, independent noise, then it is advantageous to use all available information when constructing risk scores. The following proposition formalizes this statement.

Proposition 1 *In the setting of Theorem 1, suppose $Y' = Y + S$ where $S \perp\!\!\!\perp X, Z$. Then*

$$\mathbb{E}\left[\left(\hat{Y}_{X,Z} - Y\right)^2\right] \leq \mathbb{E}\left[\left(\hat{Y}_X - Y\right)^2\right].$$

Proof. First note that

$$\begin{aligned}\hat{Y}_{X,Z} &= \mathbb{E}[Y \mid X, Z] + \mathbb{E}[S \mid X, Z] \\ &= \mathbb{E}[Y \mid X, Z] + \mathbb{E}[S],\end{aligned}$$

where the second equality uses the independence assumption. Similarly,

$$\begin{aligned}\hat{Y}_X &= \mathbb{E}[Y \mid X] + \mathbb{E}[S \mid X] \\ &= \mathbb{E}[Y \mid X] + \mathbb{E}[S].\end{aligned}$$

Now, using the notation $Y_{X,Z} = \mathbb{E}[Y \mid X, Z]$ and $Y_X = \mathbb{E}[Y \mid X]$, we have

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{Y}_{X,Z} - Y\right)^2\right] - \mathbb{E}\left[\left(\hat{Y}_X - Y\right)^2\right] \\ &\quad= \mathbb{E}\left[\left(Y_{X,Z} - Y + \mathbb{E}[S]\right)^2\right] - \mathbb{E}\left[\left(Y_X - Y + \mathbb{E}[S]\right)^2\right] \\ &\quad= \mathbb{E}\left[\left(Y_{X,Z} - Y\right)^2\right] - \mathbb{E}\left[\left(Y_X - Y\right)^2\right] + 2\mathbb{E}[S]\big(\mathbb{E}[Y_{X,Z} - Y] - \mathbb{E}[Y_X - Y]\big) \\ &\quad= \mathbb{E}\left[\left(Y_{X,Z} - Y\right)^2\right] - \mathbb{E}\left[\left(Y_X - Y\right)^2\right] \\ &\quad= \mathbb{E}\left[\mathbb{E}\left[\left(Y_{X,Z} - Y\right)^2 \mid X, Z\right]\right] - \mathbb{E}\left[\mathbb{E}\left[\left(Y_X - Y\right)^2 \mid X, Z\right]\right],\end{aligned}$$

where the third equality follows from the fact that $\mathbb{E}[Y_{X,Z}] = \mathbb{E}[Y_X] = \mathbb{E}[Y]$, and the last equality follows from the law of iterated expectations. Finally, since

$$\arg\min_c \mathbb{E}\left[\left(c - Y\right)^2 \mid X, Z\right] = Y_{X,Z},$$

we have that

$$\mathbb{E}\left[\left(Y_{X,Z} - Y\right)^2 \mid X, Z\right] - \mathbb{E}\left[\left(Y_X - Y\right)^2 \mid X, Z\right] \leq 0,$$

establishing the result.

□

D A Stylized Model of Arrest and Behavior

We formally describe and analyze the SEM depicted in Figure 1. Our model has three independent exogenous variables $U_Z = \mathrm{N}(0, \sigma_Z^2)$, $U_{A_0} = \mathrm{N}(0, \sigma_A^2)$, and $U_{A_1} = \mathrm{N}(0, \sigma_A^2)$. We additionally have two correlated exogenous variables $U_{B_0} = \mathrm{N}(0, \sigma_B^2)$ and $U_{B_1} = \mathrm{N}(0, \sigma_B^2)$ that are independent of the first three, with $\mathrm{Cov}(U_{B_0}, U_{B_1}) = \delta \geq 0$. Now, for non-negative constants α, β, and γ, the key variables in the model are generated by the following linear structural equations:

$$\begin{aligned}Z &= U_Z, \\ B_0 &= \beta Z + U_{B_0}, \\ B_1 &= \beta Z + U_{B_1}, \\ A_0 &= \alpha Z + \gamma B_0 + U_{A_0}, \\ A_1 &= \alpha Z + \gamma B_1 + U_{A_1}.\end{aligned} \tag{9}$$

We set the variances of the exogenous variables (σ_Z^2, σ_A^2, and σ_B^2) in a manner that ensures that the remaining variables (Z, B_0, B_1, A_0, and A_1) are standardized, meaning they have mean 0 and variance

1—we show how to do this below. We can thus interpret their values as representing the extent to which individuals differ from the population averages. In the case of neighborhood (Z), we can think of its value as denoting the level of police enforcement in an area.

To start, we set $\sigma_Z^2 = 1$, which ensures $\mathrm{Var}(Z) = 1$. Now, since $Z \perp\!\!\!\perp U_{B_0}$, we have that $\mathrm{Var}(B_0) = \beta^2 + \sigma_B^2$. Consequently, setting $\sigma_B^2 = 1 - \beta^2$ ensures that $\mathrm{Var}(B_0) = 1$ (and, similarly, that $\mathrm{Var}(B_1) = 1$). Finally, as above, $\mathrm{Var}(A_0) = \alpha^2 + \gamma^2 + \sigma_A^2 + 2\alpha\gamma\mathrm{Cov}(Z, B_0)$. One especially nice aspect of linear graphical models is that the covariance between any two variables can be immediately computed from the edge weights via the the Wright rules (*35, 39*). Specifically, when the nodes are standardized to have variance 1, then the covariance between any two variables in the graph is the sum, over all d-connected paths between the variables, of the product of the edge weights along the path. A path is d-connected if it does not pass through any colliders (i.e., nodes with head-to-head arrows along the path). To compute $\mathrm{Cov}(Z, B_0)$, observe that the only d-connected path between Z and B_0 is the direct path from Z to B_0, having edge weight β. As a result, $\mathrm{Cov}(Z, B_0) = \beta$, meaning that setting $\sigma_A^2 = 1 - \alpha^2 - \gamma^2 - 2\alpha\beta\gamma$ ensures that A_0 (and, analogously, A_1) have unit variance. Recapping, we have

$$\begin{aligned} \sigma_Z^2 &= 1, \\ \sigma_B^2 &= 1 - \beta^2, \\ \sigma_A^2 &= 1 - \alpha^2 - \gamma^2 - 2\alpha\beta\gamma. \end{aligned} \tag{10}$$

Our model is thus described by the four non-negative parameters α, β, γ, and δ, depicted as edge weights in Figure 1, with the constraint that the quantities in Eq. (10) are non-negative. Those constraints in turn imply that the parameters are each less than or equal to 1.

Our theoretical results in Theorem 1 and Corollary 1 require understanding the conditional distributions of model features. For multivariate normal random variables, these conditional distributions can be computed analytically (*40*), allowing us to examine properties of our motivating SEM in more depth. Specifically, suppose $\mathbf{W}$ is a k-dimensional multivariate normal random variable with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, which we partition into into its first q components and its remaining $k - q$ components: $\mathbf{W} = [\mathbf{W_1}, \mathbf{W_2}]$. Further suppose we accordingly partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ into its components:

$$\begin{aligned} \boldsymbol{\mu} &= \begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (k-q) \times 1 \end{bmatrix}, \\ \boldsymbol{\Sigma} &= \begin{bmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times q & q \times (k-q) \\ (k-q) \times q & (k-q) \times (k-q) \end{bmatrix}. \end{aligned}$$

Then the distribution of $\mathbf{W_1}$ conditional on $\mathbf{W_2}$ is multivariate normal with mean

$$\boldsymbol{\mu_1} + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1}(\mathbf{W_2} - \boldsymbol{\mu_2})$$

and covariance

$$\boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1}\boldsymbol{\Sigma_{21}}.$$

As a result, the linearity assumption of Corollary 1 is satisfied for multivariate normal random variables. In particular, in our motivating example, the conditional distribution of A_1 given A_0 and Z is

normal, with

$$\begin{aligned}
\mathbb{E}[A_1 \mid A_0, Z] &= \begin{bmatrix}\sigma_{A_1A_0} & \sigma_{A_1Z}\end{bmatrix} \begin{bmatrix}1 & \sigma_{A_0Z} \\ \sigma_{A_0Z} & 1\end{bmatrix}^{-1} \begin{bmatrix}A_0 \\ Z\end{bmatrix} \\
&= \frac{1}{1-\sigma_{A_0Z}^2} \begin{bmatrix}\sigma_{A_1A_0} & \sigma_{A_1Z}\end{bmatrix} \begin{bmatrix}1 & -\sigma_{A_0Z} \\ -\sigma_{A_0Z} & 1\end{bmatrix} \begin{bmatrix}A_0 \\ Z\end{bmatrix} \\
&= \frac{\sigma_{A_1A_0} - \sigma_{A_1Z} \cdot \sigma_{A_0Z}}{1-\sigma_{A_0Z}^2} A_0 + \frac{\sigma_{A_1Z} - \sigma_{A_1A_0} \cdot \sigma_{A_0Z}}{1-\sigma_{A_0Z}^2} Z,
\end{aligned}$$

where the σ notation denotes the covariance of the indexed random variables.

Further, the conditional distribution of (A_1, Z) given A_0 is likewise multivariate normal, with covariance matrix

$$\begin{aligned}
\begin{bmatrix}1 & \sigma_{A_1Z} \\ \sigma_{A_1Z} & 1\end{bmatrix} - \begin{bmatrix}\sigma_{A_1A_0} \\ \sigma_{A_0Z}\end{bmatrix} \begin{bmatrix}\sigma_{A_1A_0} & \sigma_{A_0Z}\end{bmatrix} &= \begin{bmatrix}1 & \sigma_{A_1Z} \\ \sigma_{A_1Z} & 1\end{bmatrix} - \begin{bmatrix}\sigma_{A_1A_0}^2 & \sigma_{A_1A_0} \cdot \sigma_{A_0Z} \\ \sigma_{A_1A_0} \cdot \sigma_{A_0Z} & \sigma_{A_0Z}^2\end{bmatrix} \\
&= \begin{bmatrix}1-\sigma_{A_1A_0}^2 & \sigma_{A_1Z} - \sigma_{A_1A_0} \cdot \sigma_{A_0Z} \\ \sigma_{A_1Z} - \sigma_{A_1A_0} \cdot \sigma_{A_0Z} & 1-\sigma_{A_0Z}^2\end{bmatrix}.
\end{aligned}$$

Consequently,

$$\mathrm{Cov}(A_1, Z \mid A_0) = \sigma_{A_1Z} - \sigma_{A_1A_0} \cdot \sigma_{A_0Z}, \tag{11}$$

and, analogously, we have that

$$\mathrm{Cov}(B_1, Z \mid A_0) = \sigma_{B_1Z} - \sigma_{B_1A_0} \cdot \sigma_{A_0Z}. \tag{12}$$

As above, we can compute the covariances in Eqs. (11) and (12) via the Wright rules. For example, as seen in Figure 1, there are two d-connected paths between Z and A_0: the direct connection with edge weight α; and the path through B_0, with product of edge weights $\beta\gamma$. Consequently, $\mathrm{Cov}(Z, A_0) = \alpha+\beta\gamma$. This procedure allows us to compute all of the terms appearing on the right-hand side of Eqs. (11) and (12), yielding:

$$\begin{aligned}
\sigma_{A_0Z} &= \alpha + \beta\gamma \\
\sigma_{A_1Z} &= \alpha + \beta\gamma \\
\sigma_{B_1Z} &= \beta \\
\sigma_{A_1A_0} &= \alpha^2 + 2\alpha\beta\gamma + \beta^2\gamma^2 + \gamma^2\delta \\
\sigma_{B_1A_0} &= \alpha\beta + \beta^2\gamma + \gamma\delta.
\end{aligned} \tag{13}$$

Leveraging the above, we now show that $\mathrm{Cov}(A_1, Z \mid A_0) \geq 0$, meaning that neighborhood is positively correlated with future arrests, conditional on past arrests. To see this, first note that

$$\begin{aligned}
\delta &= \mathrm{Cov}(U_{B_0}, U_{B_1}) \\
&\leq \sigma_B^2 \\
&= 1 - \beta^2,
\end{aligned}$$

and so $\beta^2 + \delta \leq 1$. Now,

$$\begin{aligned}
\mathrm{Cov}(A_1, Z \mid A_0) &= \sigma_{A_1 Z} - \sigma_{A_1 A_0} \cdot \sigma_{A_0 Z} \\
&= \alpha + \beta\gamma - (\alpha + \beta\gamma) \cdot (\alpha^2 + 2\alpha\beta\gamma + \beta^2\gamma^2 + \gamma^2\delta) \\
&= (\alpha + \beta\gamma) \cdot (1 - \alpha^2 - 2\alpha\beta\gamma - \beta^2\gamma^2 - \gamma^2\delta) \\
&= (\alpha + \beta\gamma) \cdot (1 - \alpha^2 - 2\alpha\beta\gamma - \gamma^2(\beta^2 + \delta)) \\
&\geq (\alpha + \beta\gamma) \cdot (1 - \alpha^2 - 2\alpha\beta\gamma - \gamma^2) \\
&= (\alpha + \beta\gamma) \cdot \sigma_A^2 \\
&\geq 0,
\end{aligned}$$

where the first inequality follows from the fact that $\beta^2 + \delta \leq 1$.

Next we consider $\mathrm{Cov}(B_1, Z \mid A_0)$, and note that

$$\begin{aligned}
\mathrm{Cov}(B_1, Z \mid A_0) &= \sigma_{B_1 Z} - \sigma_{B_1 A_0} \cdot \sigma_{A_0 Z} \\
&= \beta - (\alpha\beta + \beta^2\gamma + \gamma\delta) \cdot (\alpha + \beta\gamma).
\end{aligned}$$

In particular, when $\beta = 0$, meaning that neighborhood does not impact behavior, then

$$\mathrm{Cov}(B_1, Z \mid A_0) = -\alpha\gamma\delta.$$

In other words, when neighborhood does not impact behavior (i.e., when $\beta = 0$), neighborhood is negatively correlated with future behavior conditional on past arrests. (And, by the above, neighborhood is always positively correlated with future arrests conditional on past arrests.) By Corollary 1, it is thus better in this case to base predictions of future behavior solely on past arrests, excluding neighborhood, as we see in Figure 2.