Principled Machine Learning for Societally Consequential Decision Making

Amanda Coston

May 10 2023

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Machine Learning and Public Policy

Heinz College & Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA

Accepted by the Dissertation Committee and Approved by the Dean:

Alexandra Chouldechova, *Co-chair* Alexandra Date: 5/10/2023

Edward H. Kennedy, *Co-chair* E Kennedy Date: 5/8/2023

Hoda Heidari Heidari Date: 5/8/2023

Sendhil Mullainathan Sendhil M Date: 5/9/2023

Roni Rosenfeld Roni Rosenfeld Date: 5/10/2023
Department Head, Machine Learning Department

Ramayya Krishnan R Krishnan Date: 05/08/2023
Dean, Heinz College

Principled Machine Learning for Societally Consequential Decision Making

Amanda Coston

May 10 2023

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Machine Learning and Public Policy

Heinz College & Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA

Dissertation Committee
Alexandra Chouldechova, *Co-chair*
Edward H. Kennedy, *Co-chair*
Hoda Heidari
Sendhil Mullainathan

Keywords: Responsible AI, Machine Learning, Causal Inference, Doubly-Robust Estimation, Decision Support, Validity, Equity, Governance, Algorithmic Fairness

Funding & Acknowledgments: We are grateful to our data partners at Commonwealth Bank of Australia and the Allegheny County Department of Human Services. This research was supported in part by the Block Center for Technology and Society, the Tata Consultancy Services (TCS) Presidential Fellowship, the K&L Gates Presidential Fellowship, the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745016, and the Meta Research PhD Fellowship.

Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are solely those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

For my grandmothers

Abstract

Machine learning algorithms are widely used for decision-making in societally high-stakes settings such as child welfare, criminal justice, healthcare, hiring, and consumer lending. Recent history has illuminated numerous examples where these algorithms proved unreliable or inequitable. This thesis proposes a principled approach to the use of machine learning in societally high-stakes settings, guided by three pillars: validity, equity, and oversight. We draw on methods from a variety of fields including statistics, machine learning, and the social sciences to develop novel methods that address data challenges and complex biases embedded in sociotechnical systems. We address data problems that challenge the validity of algorithmic decision support systems by developing methods for algorithmic risk assessments that account for selection bias, confounding, and bandit feedback. We conduct causal audits for bias throughout the system in which algorithms are used to inform decision-making. Throughout we propose novel methods that use doubly-robust techniques for bias correction. We present empirical results in the child welfare, consumer credit lending, and criminal justice settings using data from Allegheny County's Department of Human Services, the Commonwealth Bank of Australia, and the Stanford Open Policing Project.

Acknowledgments

I was incredibly fortunate to have Edward Kennedy and Alex Chouldechova as my advisors. Alex and Edward unfailingly inspired me with their brilliance and passion for research. Alex was an advisor to the fullest extent, advising on all things research, career, and many dimensions of life. She was really my highly-specialized and overqualified life coach. Alex is a master at what appears to be everything, and I try to emulate her in much of what I do, from communication to problem-solving and navigating complex domains. To Edward I owe my love of doubly-robust estimators, whiteboard derivations, and that *a-ha* moment. Edward is a superpower. He unerringly shines a path forward even during the most complicated, jumbled parts of research. From Edward, I learned the beauty and value of a well-written theoretical paper. His ability to make complex ideas accessible is something I am tremendously grateful for and something I strive toward. Edward and Alex both devoted so much time and energy to supporting me. They were always one step ahead, encouraging me to strive for more than I thought I could. What's more, they are simply delightful people to spent time with, and I had so much fun working together.

I would also like to thank the other phenomenal members of my thesis committee, Hoda Heidari and Sendhil Mullainathan, who provided invaluable guidance in this process. Hoda's wealth of knowledge across many disciplines was incredibly helpful. She essentially became a third advisor for me in the final couple years of my PhD, generously devoting her time to supporting my career and working together on exciting interdisciplinary research projects. I could not have asked for a better external committee member than Sendhil. I have long admired the originality and clarity that characterizes his approach to research, and I benefited immensely from his insight.

My collaborators are some of the absolute best and brightest, and I'm so happy we could work together: Ashesh Rambachan, Anjalie Field, Luke Guerdan, Ken Holstein, Alan Mishler, Anna Kawakami, Steven Wu, Haiyi Zhu, Dan Ho, Yulia Tsvetkov, David Steier, Emily Putnam-Hornstein, Nupoor Gandhi, Kush Varshney, Karthik Natesan Ramamurthy, Dennis Wei, Skyler Speakman, Zairah Mustahsan, Supriyo Chakraborty, Han Zhao, Neel Guha, Lisa Lu, and Derek Ouyang. I would especially like to highlight my co-authors on the work presented in this thesis: Alexandra Chouldechova, Edward H. Kennedy, Alan Mishler, Ashesh Rambachan, Hoda Heidari, Haiyi Zhu, Ken Holstein, and Anna Kawakami.

I would also like to thank many other faculty members who supported me in my PhD: Roni Rosenfeld, Rayid Ghani, George Chen, Amelia Haviland, Brian Kovak, Mike Smith, Dave Choi, Siva Balakrishnan, Zach Branson, Dave Childers, Yan Huang, Zach Lipton, Edson Severnini, Lowell Taylor, as well as many others.

I am lucky to have wonderful mentors. My official mentor for several years and unofficially since is the inimitable Maria De-Arteaga. Emma Pierson has long given me valuable advice, from where to apply to graduate programs to when and how to navigate the faculty job market. I know that Emma is a brilliant and kind mentor to many, but I like to think we forged a special bond on the playground in preschool.

It was Rob Schapire who first taught me what research is, and I am grateful for his guidance during my formative years. Brian Kernighan and Mark Braverman were also influential during my undergrad and beyond.

I'd like to acknolwedge all the I assistance I received from the incredible administrators Michelle Writz, Diane Stidle, and Emily Marshall.

My PhD brought my together with amazing people like Kevin Lin, Anjalie Field, Leqi Liu, Sebastián Caldas, Ifigeneia Apostolopoulou, Jin Li, Alan Mishler, Nicholay Topin, Ashesh Rambachan, Matteo Bonvini, Priya Donti, Riccardo Fogliato, Biswajit Paria, Lisa Lee, Channing Huang, Han Zhao, Devendra Chaplot, Natalia Lombardi de Oliveira, Lingwei Chen, Kartik Gupta, Max Rubinstein, Mikaela Meyer, Shaojie Bai, Cheng Cheng, Lingxiao Zhao, Helen Zeng, Shubhranshu Shekhar, Will Guss, Ritesh Noothigattu, Andre Ribeiro Cardoso, Brandon Houghton, Charvi Rastogi.

I'd also like to acknowledge the support of many long-time friends, particularly Nora Taranto, Jessica Martin, Esther Kim, Jared Griffin, Faaez Ul Haq, Myra Iqbal, Diego Vargas, and Boris Chen. Good friends can steer you in the right direction and push you past any doubts. Faaez and Boris convinced me to apply to PhD programs despite my initial misgivings, and I am so glad they did!

My family provided steady support throughout my PhD. My first and forever role model is my mom. She suggested the idea, when I was about age seven or eight, that I could grow up to be a researcher. And my father, who gave me every opportunity I could have hoped for, and then some. My grandparents provided love and inspiration. Always curious about what I was learning, their pride kept me going. I could count on my aunt Jay's boundless enthusiasm to energize me. I'd also like to acknowledge the wonderful support of my brother AJ and his wife Tori, and my new family, Prameela, Samarth and Ayyeshwar.

My other half, Tiggy, worked tirelessly to make sure I was positioned for success. He stationed at my door, growling at anyone who dared to disrupt me during work hours. He intervened when I'd been staring at the screen for too long and demanded cuddles. He is the best work buddy. I'd also like to thank the other important furry friends along the way: Belle, Riley, and Coco.

My fiancé, Sid, for everything. I met Sid during my PhD, and this was by far the greatest part of the experience. Sid believed in me, encouraged me, supported me, and helped me in more ways that I could ever describe. He kept my spirits up and always had my back.

Table of Contents

Abstract vii

Acknowledgments viii

Table of Contents x

List of Figures xii

List of Tables xv

List of Algorithms xvi

0.1 Notation 2

1 Counterfactual Risk Assessments, Evaluation, and Fairness **5**

1.1 Background and Related Work 6

1.2 Problem Formulation and Additional Notation 7

1.3 Standard Practice for Learning Models of Risk 8

1.4 Methodology for Learning Valid Models of Risk 8

1.5 Problem Formulation for Evaluating Models of Risk 9

1.6 Standard Practice for Evaluating Models of Risk 10

1.7 Methodology for Valid Evaluations of Models of Risk 10

1.8 Empirical Results for Evaluating Models of Risk on Synthetic Data 13

1.9 Empirical Results for Evaluating Models of Risk on Real-world Child Welfare Data 13

1.10 Problem Formulation for Algorithmic Fairness 18

1.11 Theoretical Results for Algorithmic Fairness 19

1.12 Empirical Results for Algorithmic Fairness on Synthetic Data 25

1.13 Conclusion 27

2 Counterfactual Predictions under Runtime Confounding **29**

2.1 Background and Related Work 30

2.2 Problem Formulation and Additional Notation 32

2.3 Standard Practice for Learning Counterfactual Prediction Models 33

2.4 Methodology for Learning Valid Counterfactual Prediction Models under Runtime Confounding 33

2.5 Theoretical Results for Counterfactual Prediction Methods 35

2.6 Methodology for Valid Evaluations of Counterfactual Prediction Models using Observational Data . . . 38
2.7 Empirical Results on Synthetic Data . . . 39
2.8 Empirical Results on Real-world Child Welfare Data . . . 42
2.9 Conclusion . . . 43

3 Characterizing Fairness Properties over the Set of Good Models 45
3.1 Background and Related Work . . . 46
3.2 Problem Formulation and Additional Notation . . . 48
3.3 Methodology for Optimizing over the Set of Good Models . . . 49
3.4 Theoretical Results for Optimizing over the Set of Good Models . . . 52
3.5 Methodology for Optimizing Over the Set of Good Models Under Selective Labels . . . 57
3.6 Theoretical Results for Optimizing over the Set of Good Models under Selective Labels . . . 58
3.7 Empirical Results on Benchmark Recidivism Data . . . 63
3.8 Empirical Results on Real-World Consumer Lending Data . . . 65
3.9 Conclusion . . . 67

4 Framework for Evaluating the Validity of Algorithms in Consequential Decisions 69
4.1 A Taxonomy of Criteria for Justified Use of Data-driven Algorithms . . . 70
4.2 Background and Related Work . . . 72
4.3 A Taxonomy for Common Threats to Validity of Predictive Models . . . 77
4.4 Methodology for Deliberating the Validity of Predictive Models . . . 80
4.5 Conclusion . . . 84

5 Assessing Racial Bias in Police Stops 85
5.1 Background and Related Work . . . 86
5.2 Problem Formulation and Additional Notation . . . 86
5.3 Identification of the Measure of Racial Bias . . . 88
5.4 Efficiency Theory for our Measure of Racial Bias . . . 91
5.5 Methodology for Estimating our Measure of Racial Bias . . . 93
5.6 Empirical Results on Real-World Traffic Stop Data . . . 97
5.7 Conclusion . . . 99

6 Conclusion 101

Bibliography 103

List of Figures

1.1 Synthetic data results for several approaches to evaluating risk assessments. Our evaluation (DR) most accurately represents the true counterfactual evaluation. The observational evaluation erroneously suggests the observational model performs better than the counterfactual model because it evaluates against observed outcomes which includes units whose risk was mitigated by treatment. The control evaluation produces inaccurate curves because it does not assess how well the models perform on the treated population. (See § 1.8 for details) 14

1.2 Child welfare results for several approaches to evaluating risk assessments. Our evaluation (DR) is the only method that exposes the poor performance of the observational model on treated cases. The control evaluation suggests the two models perform similarly on cases that did not receive treatment. The observational evaluation would suggest the observational model has better discrimination and calibration than the counterfactual model because it evaluates against the observed outcomes which include cases whose risk was likely mitigated by child welfare services.(See § 1.9 and 1.9 for details) 16

1.3 Calibration curves by race for child welfare data. The counterfactual model is well-calibrated by race according to the control and DR evaluations but shows inequities according to the observational evaluation because black cases were more likely to get treatment which mitigates risk (see § 1.9 for more details). The observational model is poorly calibrated for both black and white cases according to the DR evaluation. 17

1.4 Child welfare risk distributions by treatment type for counterfactual and observational risk models. We expect risk to increase with the severity of treatment assigned, with ‘Placed’ out-of-home having the highest risk distribution and ‘Screened out’ of investigation having the lowest (see § 1.9). The counterfactual model displays this expected trend whereas the observational model does not. The observational model underestimates risk on cases where child welfare effectively mitigated the risk 18

1.5 Recall of the counterfactual and observational risk models for downstream child welfare decisions. At current screen-in rates (0.5), the observational model would screen out nearly 50% of very high risk cases that were placed out-of-home. The counterfactual model has higher recall at 73%. The gap is even larger for cases that were accepted for services. (See § 1.9). 18

1.6 Counterfactual and observational base rates before and after applying a fairness-corrective method that reweighs training data (§ 1.12). X-axis controls the bias of treatment assignment toward group $A = 1$. Before reweighing (“Original"), counterfactual base rates are equal (cBP holds), but observational base rates are different (oBP doesn't hold) for $k > 0$ since group $A = 1$ is more likely to get treated. Reweighing achieves oBP but cBP no longer holds. 25

1.7 Counterfactual ROC curves before and after post-processing to equalize odds (§ 1.12). Before post-processing, ROC curves are identical for both groups, indicating that counterfactual equalized odds (cEO) holds. Post-processing induces imbalance, harming group $A = 0$ and compounding initial unfairness in treatment assignment. 26

2.1 Results for several approaches (each a different color) to learning predictive models under runtime confounding on synthetic data where the runtime predictors are uncorrelated with the runtime confounders. **(a)** MSE as we vary k_z using cross-validated LASSO to learn $\hat{\pi}$, $\hat{\mu}$, $\hat{\nu}_{\text{TCR}}$, $\hat{\nu}_{\text{PL}}$, $\hat{\nu}_{\text{DR}}$ for $\rho = 0$, $d_{\text{V}} = 400$ and $k_v = 25$. At low levels of confounding (k_z), the TCR method does well but performance degrades with k_z. For any non-zero confounding, our DR method performs best. **(b)** MSE against d_{V} using cross-validated LASSO and $\rho = 0$, $k_v = 25$ and $k_z = 20$. The DR method performs the best across the range of d_{V}. When d_{V} is small, the TCR method also does well since its estimation error is small. The PL method has higher error since it suffers from the full d-dimensional estimating error in the first stage. **(c)** MSE as we vary k_z using random forests and $\rho = 0$, $d_{\text{V}} = 400$ and $k_v = 25$. Compared to LASSO in (a), there is a relatively small increase in error as we increase k_z, suggesting that estimation error dominates the confounding error. The TCR method performs best at lower levels of confounding and on par with the DR method for larger values of k_z. Error bars denote 95% confidence intervals. 40

2.2 Results for several approaches (each a different color) to learning predictive models under runtime confounding on synthetic data where the runtime predictors are correlated with the runtime confounders. **(a)** MSE against correlation ρ_{V_i,Z_i} for $k_z = 20$, $k_v = 25$, and $d_{\text{V}} = 400$. Error decreases with ρ for all methods. Our DR method achieves the lowest error under confounding ($\rho < 1$). **(b)** MSE as we increase k_z for $\rho = 0.25$, $k_v = 25$, and $d_{\text{V}} = 400$. Compare to Figure 2.1a; the weak positive correlation reduces MSE, particularly for $k_v < i \leq k_z$ when V_i is only a correlate for the confounder Z_i but not a confounder itself. **(c)** MSE against d_{V} for $\rho = 0.25$, $k_z = 20$, and $k_v = 25$. The DR method is among the best-performing for all d_{V}. As with the uncorrelated setting (2.1b), the DR and TCR methods are better able to take advantage of low d_{V} than the PL method. Error bars denote 95% confidence intervals. 41

3.1 Semi-synthetic results for characterizing fairness properties over the set of good models trained on real-world credit lending data. This plot shows the area under the ROC curve (AUC) with respect to the synthetic outcome against disparity in the average risk prediction for the disadvantaged $(A_i = 1)$ vs advantaged $(A_i = 0)$ groups. FaiRS reduces disparities for the RIE and IE approaches while maintaining AUCs comparable to the benchmark models (first row). Evaluation on only funded applicants (second row) overestimates the performance of TFCS and KGB models and underestimates disparities for all models. Error bars show the 95% confidence intervals. See § 3.8 for details. 67

4.1 Visual representation of our proposed taxonomy for the justified use of algorithms in high-stakes decision making. Validity, reliability and value alignment are required for justified use. These concepts are overlapping and interconnected, encompassing many aspects of responsible machine learning. 72

5.1 Real-world results of our counterfactual method to estimate racial bias in police officers' decisions to stop drivers in Madison, WI. Red suggests racial bias under our experimental design; white no effect; blue that daylight would reduce the risk of a stop for black drivers more than for white drivers. 98

5.2 Real-world results of our aggregated measure of racial bias in the officer's decision to stop drivers for five cities. Values below the dotted line suggest racial bias. Dots indicate the point estimate of our bias-corrected method, and the bars indicate the 95% confidence intervals. 98

List of Tables

1.1 Empirical analysis of whether the models of re-referral risk transfer to related risk tasks in the child welfare domain. This table presents the area under ROC and PR curves using the models of re-referral risk to predict a related risk task, out-of-home placement (95% confidence intervals given in parentheses). The observational model performs worse than a random classifier. The counterfactual model performs better; it learns a model of risk that transfers to related risk tasks whereas the observational model does not. (See § 1.9) 19

1.2 Empirical results on synthetic data show that post-processing methods to achieve parity in standard group fairness metrics can induce unfairness in the counterfactual fairness metric. This table gives the counterfactual and observational generalized FNR/FPR before and after post-processing to equalize odds (§ 1.12) using threshold $= 0.5$. Before post-processing (“Original"), the counterfactual generalized rates (cGFNR and cGFPR) are the same for both groups. Post-processing equalizes the observational rates (oGFNR and oGFPR) but induces noticeable disparity in both cGFNR and cGFPR. 26

2.1 Synthetic results for prediction error of several learning methods under runtime confounding. This table provides the MSE $\mathbb{E}[\Big(\nu(V) - \hat{\nu}(V)\Big)^2]$ under correct specification and misspecification in the 2nd stage for $d = 500$, $d_{\mathrm{V}} = 400$, $k_v = 24$, $k_z = 20$ and $n = 3000$ (with 95% confidence intervals). Our DR method has the lowest error in both settings. Errors are larger for all methods under misspecification. 42

2.2 Child welfare screening results for prediction error of several learning methods under runtime confounding. This table shows the MSE estimated via our evaluation procedure (§ 2.6). The PL and DR approaches achieve lower MSE than the TCR approach. 95% confidence intervals given in parentheses. 42

3.1 Results showing fairness properties over the set of good models trained on Propublica's recidivism prediction dataset. Our optimization method shows that the redicivism prediction model used in practice, COMPAS, fails to satisfy the "business necessity" defense for disparate impact by race. The set of good models (performing within 1% of COMPAS's training loss) includes models that achieve significantly lower disparities than COMPAS. The first panel (SP) displays the disparity in average predictions for black versus white defendants (Def. 3.2.1). The second panel (BFPC) analyzes the disparity in average predictions for black versus white defendants in the positive class, and the third panel examines the disparity in average predictions for black versus white defendants in the negative class (Def. 3.2.2). Standard errors are reported in parentheses. See § 3.7 for details. . . . 65

List of Algorithms

1 The plug-in (PL) approach for counterfactual predictions under runtime confounding . . . 34

2 The plug-in (PL) approach for counterfactual predictions under runtime confounding with cross-fitting . . . 34

3 Our doubly-robust (DR) approach for counterfactual predictions under runtime confounding . . . 35

4 Our doubly-robust (DR) approach for counterfactual predictions under runtime confounding with cross fitting . . . 35

5 Cross-fitting approach to evaluation of counterfactual prediction methods . 39

6 Algorithm for finding the predictive disparity minimizing model 50

7 The Reject inference by extrapolation (RIE) approach for addressing missing outcomes due to the the selective labels problem . . . 57

8 The Interpolation and extrapolation (IE) method for addressing missing outcomes due to the the selective labels problem . . . 58

Introduction

Machine learning is increasingly used to make decisions in high-stakes settings, such as child welfare, criminal justice, consumer lending, education, and healthcare (Saxena et al., 2020; Vaithianathan et al., 2017; Raghavan et al., 2020a; Chouldechova, 2017; Cattell et al., 2021). These decisions affect future health and economic opportunities, and in aggregate they shape our societal structures. Often the data available for such tasks is abundant but nonetheless noisy, biased, or incomplete. Failure to properly address these data challenges can disproportionately harm vulnerable and historically marginalized groups (Barocas and Selbst, 2016b; Obermeyer et al., 2019a; Coston et al., 2020b, 2021a). In this thesis, we develop statistical methodologies and a deliberation framework to identify and address data issues that challenge the responsible use of machine learning in consequential settings.

When machine learning is used for high-stakes decision making, a common approach applies the standard supervised learning paradigm. Under this approach, one identifies an outcome of interest (typically a proxy for the actual outcome of interest) in the available data and builds a predictive model for this outcome using the other variables as predictors. This standard approach is often ill-suited when, as is common in real-world applications, the datasets are not representative of the target population on which the machine learning tool will be deployed and the predicted outcome can be markedly different from the outcome relevant to the decision-making task (Mullainathan and Obermeyer, 2021; Coston et al., 2020b; Fogliato et al., 2021; Wang et al., 2022). Moreover, the standard approach to performance evaluation that computes test metrics on a held-out set often fails to provide a valid assessment of performance on the target population (Kallus and Zhou, 2018b; Coston et al., 2021b).

A core principle of this thesis is that we must have alignment between what we intend to measure (e.g., what we intend for the ML tool to predict, or what we intend for the evaluation to assess) and what the method *actually* measures.This property is known as ***validity*** (Coston et al., 2023). Unaddressed data problems such as selection bias or missing data can induce misalignment and render machine learning tools invalid. We discuss examples of these problems in the child welfare, criminal justice, and consumer lending settings, and we propose methods of solution.

We demonstrate the connection between validity and our second principle, ***equity***, which requires that the ML tool does not unjustifiably advantage certain demographic groups over others. We show that often it is vulnerable or historically disadvantaged populations who are most likely to be under-represented or misrepresented in the available data. We present methods for reliably assessing demographic biases in algorithms and for scrutinizing validity and equity in the broader context in which the algorithmic tools are deployed. Analyzing validity and equity effectively in practice requires tools of ***governance*** that provide safeguarding and structure processes to carefully design and evaluate ML tools. We develop a framework to guide deliberation around common issues that threaten the validity and legitimacy of predictive algorithms.

The methods presented in this thesis constitute an alternative approach to the standard machine learning paradigm for consequential decision making. Our principled approach makes explicit the target population and target outcome, makes adjustments for any differences between the data sample and the target population, and makes reasonable assumptions to identify the target outcome and evaluation metrics. We develop efficient methods to estimate these quantities using influence-function based techniques from causal inference, a discipline that is suited to decision-making settings where the decision can change downstream outcomes. We present theoretical analysis for our methods that informs how to appropriately quantify uncertainty. The suite of methods proposed in this thesis comprise a toolkit for responsible use in model construction, evaluation, and fairness assessments.

We describe the problem setting and notation in § 0.1. The subsequent three sections consider how to obtain valid predictions, evaluations, and fairness assessments under different assumptions on the nature of missing data. § 1 describes methods when we have measured all confounding factors that jointly affect the decision and outcome of interest. In a number of decision support settings, confounding factors may be difficult to measure and input into a prediction model at runtime but may nonetheless be available in an offline dataset for training and evaluation. § 2 provides methods for this “runtime confounding" setting. § 3 delves into issues of fairness and equity through the lens of the Rashomon effect, an empirical phenomenon whereby a multiplicity of models achieve comparably good performance overall but differ notably in their individual predictions. § 4 describes a framework to scrutinize for validity in algorithmic design and ultimately to inform the decision to deploy a tool in a high-stakes setting. Expanding our scope to the broader contexts in which algorithms are trained and used, § 5 proposes a retrospective statistical audit for racial bias in human decisions in the criminal justice system. The methodology we propose in this section also shows how machine learning can be used in societally consequential domains to assess these systems and the actors who wield power within in them. We connect the work presented in this thesis to our three guiding principles of validity, equity, and governance.

0.1 Notation

We use Y to denote the observed outcome which we generally assume is binary except in § 2 which considers the more general case that $Y \in \mathcal{Y} \subseteq \mathbb{R}$. We let $T \in \{0, 1\}$ denote a binary decision of interest. We will use the terms ‘decision’ and ‘treatment’ and ’intervention’ interchangeably. In describing our proposed learning and evaluation methods, we rely on the potential outcomes framework common in causal inference (Rubin, 2005; Neyman, 1923; Kennedy et al., 2013). In this framework, Y^t denotes the outcome under treatment t. For any given case we only get to observe Y^0 or Y^1, depending on whether the case was treated. We let $X \in \mathcal{X} \subseteq \mathbb{R}^d$ denote the covariates (or features) which may include a protected or sensitive attribute $A \in \{0, 1\}$. We use subscripts i to index our data; e.g., X_i are the features for instance i. $\pi(X) = \mathbb{P}(T = 1 \mid X)$ denotes the propensity score, whose estimate we denote by $\hat{\pi}(X)$. For the binary outcome setting, we use $\hat{Y} : \mathcal{X} \mapsto \{0, 1\}$ to denote our predicted label and $\hat{f} : \mathcal{X} \mapsto [0, 1]$ to denote the predicted score which is the model's estimate of the target outcome.[1] We let $p(x)$ denote probability density functions; $\hat{f}$ denote an estimate of f; $L \lesssim R$ indicate that $L \leq C \cdot R$ for some universal constant C; $\mathbb{I}$ denote the indicator function; and define $\|f\|^2 := \int (f(x))^2 p(x) dx$.

[1] $\hat{Y}(X)$ is typically obtained by thresholding $\hat{f}(X)$.

1

Counterfactual Risk Assessments, Evaluation, and Fairness

Much of the activity in using machine learning to help address societal problems focuses on decision-making algorithms. In settings such as health, education, child welfare and criminal justice, decision-making algorithms commonly take the form of risk assessment instruments (RAIs), which distill rich case information into risk scores that reflect the likelihood of the case resulting in one or more adverse outcomes (Chouldechova et al., 2018; Kube et al., 2019; Ferguson, 2016; Kehl and Kessler, 2017; Stevenson, 2018; Caruana et al., 2015; Smith et al., 2012). RAIs are typically trained and evaluated as though the task were prediction when in reality the associated decision-making tasks are often interventions. In this chapter, we show how RAIs trained as such typically fall short of our requirement for validity. Models trained and evaluated in this way answer the question: What is the likelihood of an adverse outcome *under the observed historical decisions*? Yet the question relevant to the decision maker is: What is the likelihood of an adverse outcome *under the proposed decision*?

In order to meet requirements for validity, RAIs for these settings must be developed and evaluated taking into account the effect of historical decisions on the observed outcomes. Failure to do so will result in RAIs that, despite appearing to perform well according to standard evaluation practices, are not valid predictors for the target of interest. We demonstrate the resulting harms, showing how these RAIs underperform on cases that have been historically receptive to intervention.

In this chapter we present an approach to address this using **counterfactual risk modeling and evaluation** based on work in Coston et al. (2020b). Counterfactual modeling has been proposed for medical RAIs (Schulam and Saria, 2017b; Shalit et al., 2017; Alaa and van der Schaar, 2017), and prior work has used counterfactual evaluation for off-policy learning in bandit settings (Dudík et al., 2011). However, the question of adapting counterfactual evaluation for risk assessments and in particular for predictive bias assessments remains open. As a solution, we propose a new evaluation method for RAIs that uses doubly-robust estimation techniques from causal inference (Van der Laan et al., 2003; Robins and Rotnitzky, 2001). We also argue that fairness metrics that are functions of the outcome should be defined counterfactually, and we use our evaluation method to estimate these metrics. We theoretically and empirically characterize the relationship between the standard fairness metrics and their counterfactual analogues. Our results suggest that in many cases, achieving parity in the standard metric will not achieve parity in the counterfactual metric.

In this chapter we make the following contributions: 1) We define counterfactual versions of standard predictive performance metrics and propose doubly-robust estimators of these metrics (§ 1.2); 2) We provide empirical support that this evaluation outperforms existing methods using a synthetic dataset and a real-world child welfare hotline screening dataset (§ 1.2); 3) We propose counterfactual formulations of three standard fairness metrics that are more appropriate for decision-making settings (§ 1.10); 4) We provide theoretical results showing that only under strong conditions, which are unlikely to hold in general, does fairness according to standard metrics imply fairness according to counterfactual metrics (§ 1.10); 5) We demonstrate empirically that applying existing fairness-corrective methods can increase disparity in the counterfactual redefinition of the metric they target (§ 1.10).

1.1 Background and Related Work

1.1.1 Counterfactual learning and evaluation

Literature on contextual bandits has considered counterfactual learning and evaluation of decision *policies*. While this literature is methodologically relevant, as we discuss below, it addresses a different problem. In the *decision support* setting we are considering, human users will ultimately decide what action to take. The goal of the learning and evaluation task is not to learn a decision policy, but rather to learn a risk model that will inform human decisions. That is, the risk assessment task is to accurately and fairly estimate the probability of an outcome under a given intervention.

While the underlying task is different, the statistical methods used in evaluation are related. Swaminathan and Joachims (2015) use propensity score weighting, a form of importance sampling, to correct for the effect of the historical treatment on the observed outcome, and they propose learning the optimal policy based on the minimization of the propensity-score weighted empirical risk. Propensity-score methods are a good candidate when one has a good model of the historical decision-making policy, but may otherwise be biased. Doubly robust (DR) methods, by contrast, are robust to parametric misspecification of the propensity score model if instead one has the correct specification of the model of the regression outcome $\mathbb{E}[Y|X]$ where Y is the outcome and X are the features/covariates (Van der Laan et al., 2003; Robins et al., 1994; Robins and Rotnitzky, 1995). In a nonparametric setting, DR methods have faster rates of convergence than propensity-score methods (Kennedy, 2016). DR methods have been used for policy learning in the offline bandit setting (Dudík et al., 2011). The policy learned minimizes a DR estimate of the loss. Their framework can also be used to evaluate a policy by computing the DR estimate of its expected reward.

Prior work has considered counterfactual RAIs in a temporal setting (Schulam and Saria, 2017b). In this work, the trained model is evaluated on real data using the observed outcomes, and on simulated data. Evaluating against the observed outcomes can be misleading in settings in which treatment was not assigned randomly (see § 1.7). In our work we propose instead to adapt DR techniques, as have been used in the bandit literature for evaluating policies, to provide evaluations of counterfactual RAIs.

Counterfactual learning in the causal inference literature uses model selection based on DR estimation of counterfactual loss (Van der Laan et al., 2003). Whereas this approach evaluates counterfactual metrics implicitly, our approach does so explicitly, providing the estimators for standard classification metrics in § 1.7.

1.1.2 Fairness and causality

The literature on counterfactual fairness offers notions of fairness based on the counterfactual of the protected attribute (or its proxy) (Kusner et al., 2017; Wang et al., 2019; Kilbertus et al., 2017). In this work, a policy is considered fair if it would have made the same decision had the individual had a different value of the protected attribute (and hence, potentially different values of features affected by the attribute). In this setting, the treatment decision is the outcome, and the protected attribute is the 'treatment'. By contrast, we consider counterfactual treatment decisions and consider a future observation to be the outcome.[1]

Another line of work considers unfair causal pathways between the protected attribute (or its proxy) and the outcome variable or target of prediction (Nabi and Shpitser, 2018; Zhang and Bareinboim, 2018b). These papers characterize or explain discrimination via path-specific effects, which are defined by interventions on the protected attribute. We do not consider interventions on (i.e. counterfactuals of) the protected attribute; rather, we propose methods that account for interventions on treatment decisions in training and evaluation.

Fairness definitions based on the counterfactual of the protected attribute are not widely used in RAI settings for two reasons: one technical and one practical. The technical challenge is that the assumptions required to estimate these counterfactual metrics prohibit the use of important features, such as prior history, or require full specification of the structural causal model (SCM) (Zhang and Bareinboim, 2018a; Kusner et al., 2017, 2019) These requirements are too restrictive for our settings of interest where we have insufficient domain knowledge to construct the SCM and where we are unable to disregard important predictors like prior history. More significantly, the practical concern is that these definitions are ill-suited for risk assessment settings like child welfare screening. As we discuss in § 1.10, decisions made based on the counterfactual protected attribute may cause further harm to the protected groups.

Our work bears conceptual similarity to the analysis of residual unfairness when there is selection bias in the training data that induces covariate shift at test time as discussed in (Kallus and Zhou, 2018b). In settings where cases are systematically screened out from the training set, such as loan approvals in which we do not get to see whether someone who was denied a loan would have repaid, they find that applying fairness-corrective methods is insufficient to achieve parity. We consider a different but related setting in which we observe outcomes for all cases, but these outcomes are under different treatments. We propose fairness definitions that account for the effect of these treatments on the observed outcomes, and analyze the conditions under which existing methods can achieve this notion of counterfactual fairness.

Before proceeding to introduce the learning approaches and evaluation methods considered in this work, we pause to clarify the types of risk-based decision policies to which our evaluation strategy as presented is tailored, and provide some background on algorithm-assisted decision making in child welfare hotline screening.

1.2 Problem Formulation and Additional Notation

RAIs typically inform human decisions either by identifying cases that are the most (or least) *risky*, or by identifying cases that are the most (or least) *responsive*. The evaluation metrics we consider are most directly relevant in the paradigm where human decision-makers wish to

[1] This distinction is also made in a survey of fairness literature (Mitchell et al., 2018).

intervene on the *riskiest* cases. However, our method can readily be adapted (as discussed in § 1.5) for paradigms in which interventions are being targeted based on responsiveness.

The motivating application for our work is child welfare screening. Child welfare service agencies across the nation field over 4.1 million child abuse and neglect calls each year (U.S. Department of Health & Human Services, 2019). Call workers must decide whether to "screen in" a call, which refers to opening an investigation into the family. The child welfare system is responsible for responding to all cases where there is significant suspicion that the child is in present or impending danger. The standard of practice is therefore to identify the *riskiest* cases. Jurisdictions in California, Colorado, Oregon, Texas and Pennsylvania have either considered or are using RAIs for call screening processes. The RAIs are trained on historical data to predict adverse child welfare outcomes, such as re-referral to the hotline or out-of-home foster care placement (Chouldechova et al., 2018). The decision to investigate a call can affect the likelihood of the target outcomes.

Recall that we use Y^t to denote the outcome under treatment t. We will take $T = 0$ to be the baseline treatment, the decision under which it is relevant to assess risk. Most risk assessment settings have a natural baseline, which is often the decision to not intervene. For instance, in education one might wish to assess the likelihood of poor outcomes if a student is not offered support; in child welfare it is natural to assess the risk of re-referral if the call is not investigated. We refer to the baseline treatment as *control* and the not-baseline treatment as *treatment*.

1.3 Standard Practice for Learning Models of Risk

In this section we introduce the standard practice for model training, which we term the "observational" method of model training.

Observational

The *observational* RAI produces risk estimates by regressing Y on X for the entire observed dataset. i.e., this RAI estimates $\mathbb{E}[Y \mid X]$. This model answers the question: What is the likelihood of an adverse outcome under the *observed historical decisions*? The observational RAI is not a valid predictor for the likelihood of outcomes under the proposed decision. The observational RAI is therefore ill-suited for guiding future decisions; it will, for instance, underestimate (baseline) risk for cases that were historically responsive to treatment.

1.4 Methodology for Learning Valid Models of Risk

In this section we introduce the "counterfactual" form of model training.

Counterfactual

The counterfactual model of risk estimates the outcome under the baseline treatment. Our counterfactual model of risk targets $\mathbb{E}[Y^0 \mid X]$. Even though we only observe Y^0 or Y^1 for any given observation, we may nevertheless draw valid inference about both potential outcomes under a set of standard identifying assumptions[2]. These assumptions hold by design in our

[2] Identification is the process of using a set of assumptions to write a counterfactual quantity in terms of observable quantities.

synthetic dataset, and we discuss why they may be reasonable in the child welfare setting under each point. When these assumptions hold, the counterfactual RAI is a valid predictor of outcomes under the proposed decision.

1. Consistency: $Y = TY^1 + (1-T)Y^0$.
   This assumes there is no interference between treated and control units. This is a reasonable assumption in the child welfare setting since opening an investigation into one case will not likely affect another case's observed outcome.[3]

2. Exchangeability: $Y^0 \perp T \mid X$. This assumes that we measured all variables X that jointly influence the intervention decision T and the potential outcome Y^0. This is an untestable assumption but it may be reasonable in the child welfare setting where the measured variables capture most of the information the call screeners use to make their decision (see § 1.9 for more details).

3. Weak positivity requirement: $\mathbb{P}(\pi(X) < 1) = 1$ requires that each example have some non-zero chance of the baseline treatment. This can hold by construction in decision support settings. We can filter out cases that violate this assumption since the decision for these cases is nearly certain.[4]

Our assumptions identify the target $\mathbb{E}[Y^0|X] = \mathbb{E}[Y|X, T = 0]$.

The counterfactual model estimates $\mathbb{E}[Y^0 \mid X]$ by computing an estimate of $\mathbb{E}[Y \mid X, T = 0]$. We can train such a model by applying any probabilistic classifier to the control population. Since the control population may have a different covariate distribution than the full population, reweighing can be used to correct this covariate shift (Quionero-Candela et al., 2009). This may be useful in a setting with limited data or where model misspecification is a concern (Sugiyama et al., 2007).

1.5 Problem Formulation for Evaluating Models of Risk

To evaluate how well our models of risk might inform decision-making in the paradigm where interventions should be targeted at the riskiest cases, we assess performance metrics such as precision, true positive rate (TPR), false positive rate (FPR), and calibration.[5] Since the task is to evaluate how well the model predicts risk under a baseline intervention, we specify the performance metrics in terms of Y^0. The target counterfactual TPR is

$$\mathbb{E}[\hat{Y} \mid Y^0 = 1] \tag{1.1}$$

The target counterfactual precision is

$$\mathbb{E}[Y^0 \mid \hat{Y} = 1] \tag{1.2}$$

The target counterfactual FPR is

$$\mathbb{E}[\hat{Y} \mid Y^0 = 0] \tag{1.3}$$

[3] We set the treatment to be the same value for all children in a family.

[4] Risk assessments are unnecessary for these cases since the decision-maker already knows what to do.

[5] In the paradigm where interventions are to be targeted at the *most responsive* cases, performance metrics such as discounted cumulative gain (DCG) or Spearman's rank correlation coefficients are more natural choices for evaluation. DR estimates can be constructed for these metrics as well.

A model is well-calibrated in the counterfactual sense when

$$\mathbb{E}\Big[Y^0 \mid r_1 \leq \hat{f}(X) \leq r_2\Big] \approx \frac{r_1 + r_2}{2} \tag{1.4}$$

where r_1, r_2 define a bin of predictions. We next describe two standard practice approaches for evaluation, noting why these approaches do not adequately estimate the counterfactual targets. We subsequently introduce our proposed approach for obtaining valid estimates of the target.[6]

1.6 Standard Practice for Evaluating Models of Risk

Observational Evaluation

A standard practice approach evaluates the model against the observed outcomes. An observational Precision-Recall (PR) curve plots estimates of observational precision, $\mathbb{E}[Y \mid \hat{Y} = 1]$, against estimates of observational TPR[7], $\mathbb{E}[\hat{Y} \mid Y = 1]$. An observational ROC curve plots estimated observational TPR against estimates of observational FPR, $\mathbb{E}[\hat{Y} \mid Y = 0]$. An observational calibration curve plots our estimate of $\mathbb{E}[Y \mid r_1 < \hat{f}(X) < r_2]$, the observational outcome rate for scores in the interval $[r_1, r_2]$, across intervals. The observational evaluation answers the question: Does the RAI accurately predict the likelihood of an adverse outcome under the *observed historical decisions*? This evaluation approach can be misleading since $Y \neq Y^0$. For instance, it will conclude that a valid counterfactual model of risk under baseline performs poorly because its predictions will be systematically inaccurate for cases that are responsive to treatment.

Evaluation on the Control Population

The standard practice counterfactual approach to evaluation computes error metrics on the control population (Schulam and Saria, 2017b). The PR curve evaluated on the control population plots estimates of $\mathbb{E}[Y \mid \hat{Y} = 1, T = 0]$ against estimates of $\mathbb{E}[\hat{Y} \mid Y = 1, T = 0]$, and the ROC and calibration curve target estimands are similarly defined by conditioning on $T = 0$. When the control population is not representative of the full population (i.e. $T \not\perp X$), as is the case in non-experimental settings, this evaluation may be misleading since $\mathbb{E}[Y \mid T = 0] = \mathbb{E}[Y^0 \mid T = 0] \neq \mathbb{E}[Y^0]$. A method that performs well on the control population may perform poorly on the treated population (or vice-versa). In child welfare, cases where the perpetrator has a history of abuse are more likely to be screened in. Since there is more information associated with these cases, a model may be able to discriminate risk better for these cases than on cases in the control population with little history.

1.7 Methodology for Valid Evaluations of Models of Risk

Doubly-robust (DR) Counterfactual Evaluation

We propose to improve upon the control population evaluation procedure by using DR estimation to perform counterfactual evaluation using both treated and control cases. This ensures that performance is assessed on a representative sample of the population. Our method

[6] All evaluations are computed on a test partition that is separate from the train partition

[7] TPR and recall are equivalent.

estimates the counterfactual outcome for all cases and evaluates metrics on this estimate. Other approaches such as inverse-probability weighing (IPW) or plug-in estimates could be used for a counterfacutal evaluation, but DR techniques are preferable because they have faster rates of convergence for nonparametric methods, and for parametric methods they are robust to misspecification in one of the nuisance functions, which estimate treatment propensity $\pi(X)$ and the outcome regression $\mathbb{E}[Y^0 \mid X]$ (Robins et al., 1994; Robins and Rotnitzky, 1995; Kennedy, 2016). Under sample splitting and $n^{1/4}$ convergence in the nuisance function error terms, these estimates are $\sqrt{n}$-consistent and asymptotically normal. This enables us to compute confidence intervals (see *Calibration* below for an example).

We first consider estimates of the average outcome under control $\mathbb{E}[Y^0]$. Under our causal assumptions in § 1.4, $\mathbb{E}[Y^0] = \mathbb{E}[\mathbb{E}[Y \mid X, T = 0]]$. The plug-in estimate is:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{s}_0(X_i)$$

where $\hat{s}_0(X)$ denotes the score of our counterfactual model. The IPW estimate uses the observed outcome on the control population and reweighs the control population to resemble the full population:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1-T_i}{1-\hat{\pi}(X_i)} Y_i$$

DR estimators[8] combine the plug-in estimate with an IPW-residual bias-correction term for the control cases:

$$DR_{Y^0} = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{1-T_i}{1-\hat{\pi}(X_i)}(Y_i - \hat{s}_0(X_i)) + \hat{s}_0(X_i)\right] \tag{1.5}$$

Next we consider the counterfactual targets in Equations 1.1- 1.4. We identify the target under our causal assumptions and then state the DR estimator. We emphasize the distinction that $\hat{f}$ is the score of any model we wish to evaluate whereas $\hat{s}_0$ is the score of our counterfactual model in § 1.4.

TPR (Recall): Counterfactual TPR is identified as

$$\mathbb{E}[\hat{Y} \mid Y^0 = 1] = \frac{\mathbb{E}\Big[\hat{Y}\mathbb{E}[Y \mid X, T = 0]\Big]}{\mathbb{E}\Big[\mathbb{E}[Y \mid X, T = 0]\Big]} \tag{1.6}$$

.

The DR estimate for the numerator is

$$\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i \Big[\frac{1-T_i}{1-\hat{\pi}(X_i)}(Y_i - \hat{s}_0(X_i)) + \hat{s}_0(X_i)\Big] \tag{1.7}$$

The DR estimate for the denominator is DR_{Y^0} in Equation 1.5.

[8] In survey inference, this is known as the generalized regression estimator (Särndal et al., 1989).

Precision: The target counterfactual precision is identified as

$$\mathbb{E}[Y^0 \mid \hat{Y} = 1] = \mathbb{E}[\mathbb{E}[Y \mid X, T = 0] \mid \hat{Y} = 1] \tag{1.8}$$

The DR estimator for precision is

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1-T_i}{1-\hat{\pi}(X_i)}(Y_i - \hat{s}_0(X_i)) + \hat{s}_0(X_i)\right]\mathbb{I}\{\hat{Y}_i = 1\}}{P(\hat{Y}_i = 1)} \tag{1.9}$$

where $\mathbb{I}$ denotes the indicator function.

Calibration: The target in Equation 1.4 is identified as

$$\mathbb{E}\Big[\mathbb{E}[Y \mid X, T = 0] \mid r_1 \leq \hat{f}(X) \leq r_2\Big]$$

The DR estimate for calibration is

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1-T_i}{1-\hat{\pi}(X_i)}(Y_i - \hat{s}_0(X_i)) + \hat{s}_0(X_i)\right]\mathbb{I}\{r_1 \leq \hat{f}(X_i) \leq r_2\}}{P(r_1 \leq \hat{f}(X_i) \leq r_2)} \tag{1.10}$$

To compute the confidence interval for this estimate, we compute the number of data points in the bin $n_r = \sum_{i=1}^{n}\mathbb{I}\{r_1 \leq \hat{f}(X_i) \leq r_2\}$ and the variance in the bin

$$var(\phi_r) = var\Big(\frac{1-T_i}{1-\hat{\pi}(X_i)}(Y_i - \hat{s}_0(X_i)) + \hat{s}_0(X_i) \mid r_1 \leq \hat{f}(X_i) \leq r_2\Big)$$

.

Then we use the normal approximation to compute the interval: $\pm z\sqrt{\frac{var(\phi_r)}{n_r}}$ where $z = 1.96$ for a 95% confidence interval.

FPR: The target counterfactual FPR is identified as

$$\mathbb{E}[\hat{Y} \mid Y^0 = 0] = \frac{\mathbb{E}\Big[\hat{Y}\mathbb{E}[1 - Y \mid X, T = 0]\Big]}{\mathbb{E}\Big[\mathbb{E}[1 - Y \mid X, T = 0]\Big]} \tag{1.11}$$

The DR estimator for the numerator is

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i\Big[\frac{1-T_i}{1-\hat{\pi}(X_i)}(\hat{s}_0(X_i) - Y_i) + (1 - \hat{s}_0(X_i))\Big] \tag{1.12}$$

For the denominator we use $1 - DR_{Y^0}$ where DR_{Y^0} is in Eq 1.5.

We next present the results of these three evaluations on a synthetic dataset and our child welfare dataset. Comparing to the true counterfactual for the synthetic data, we find that our DR evaluation is more accurate than either the observational or control evaluations. On the real world child welfare data, where we do not have access to all counterfactuals, we perform a comparison to expert assessment of risk to give further credence to the conclusions from our DR evaluation.

1.8 Empirical Results for Evaluating Models of Risk on Synthetic Data

We begin our empirical analysis with a synthetic dataset so that we can compare methods in a setting where we observe both potential outcomes. We specify two groups with different treatment propensities, but the treatment is constructed to be equally effective at reducing the likelihood of adverse outcome ($Y = 1$) for both groups. We generate 100,000 data points (X_i, Y_i^0, Y_i^1, T_i) where $X_i = (Z_i, A_i)$ and $Z_i \sim \mathcal{N}(0, 1)$, a normal distribution with mean 0 and variance 1. $A_i \sim Bern(0.5)$, a Bernoulli with mean 0.5. $Y_i^0 \sim Bern(\sigma(Z_i - 0.5))$ where $\sigma(z) = \frac{1}{1+e^{-z}}$. $Y_i^1 \sim Bern(c\sigma(Z_i - 0.5))$ where $c = 0.1$ controls the treatment effect. $T_i \sim Bern(\sigma(Z_i - 0.5 + kA_i))$ where $k = 1.6$ describes the bias in treatment assignment toward group $A = 1$. We set $Y = TY^1 + (1 - T)Y^0$. The base rates are $\mathbb{E}[Y] = 0.17$; $\mathbb{E}[Y^0] = 0.4$; and $\mathbb{E}[Y^1] = 0.04$. The treatment rates are $\mathbb{E}[T] = 0.55$; $\mathbb{E}[T \mid a = 0] = 0.4$; and $\mathbb{E}[T \mid a = 1] = 0.71$.

We use logistic regression to train both the observational model of $\mathbb{E}[Y \mid X]$ and counterfactual model of $\mathbb{E}[Y^0 \mid X]$ as well as the propensity model of $\mathbb{E}[T \mid X]$. Under this choice of model, the propensity model and counterfactual model are both correctly specified, and accordingly, the plug-in and IPW estimates are both consistent in this setting. However, in practice, there is no way to know whether the models are correctly specified, so DR estimates are preferable for real-world settings. We use $X = (Z, A)$ as the features.

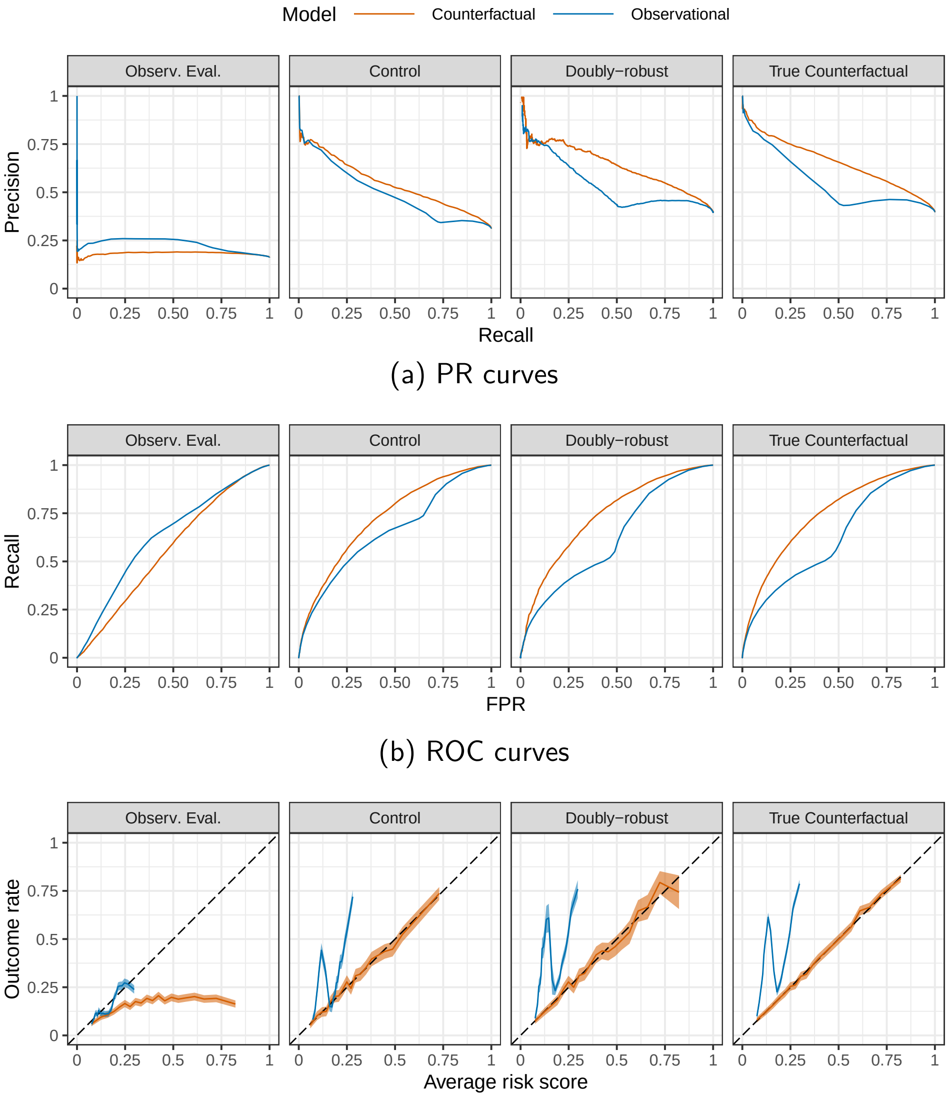
Figure 1.1 displays PR, ROC, and calibration curves.[9] DR evaluation most closely aligns with the true counterfactual evaluation. Notably, the observational evaluation suggests that the observational model outperforms the counterfactual model whereas the true counterfactual evaluation shows the counterfactual model performs better.

1.9 Empirical Results for Evaluating Models of Risk on Real-world Child Welfare Data

We also apply counterfactual learning and evaluation to the problem of child welfare screening. The baseline intervention is screen-out (which means no investigation occurs). The data consists of over 30,000 calls to the hotline in Allegheny County, Pennsylvania, each containing more than 1000 features describing the call information as well as county records for all individuals associated with the call. The call features are categorical variables describing the allegation types and worker-assessed risk and danger ratings. The county records include demographic information such as age, race and gender as well as criminal justice, child welfare, and behavioral health history. The outcome is re-referral within a six month period. Our approach contrasts to prior work which used placement out-of-home as the outcome (Chouldechova et al., 2018; De-Arteaga et al., 2018). This outcome is only observed for cases under investigation; therefore it cannot be used to identify Y^0, the risk under no investigation.

We use random forests to train the observational and counterfactual risk assessments as well as the propensity score model. We used reweighing to correct for covariate shift but did not observe a boost in performance, likely because we have sufficient data and we used a nonparametric model.

[9] The code for this analysis is given in https://github.com/mandycoston/counterfactual



(a) PR curves

(b) ROC curves

(c) Calibration curves. 95% pointwise confidence bounds shown.

Figure 1.1: Synthetic data results for several approaches to evaluating risk assessments. Our evaluation (DR) most accurately represents the true counterfactual evaluation. The observational evaluation erroneously suggests the observational model performs better than the counterfactual model because it evaluates against observed outcomes which includes units whose risk was mitigated by treatment. The control evaluation produces inaccurate curves because it does not assess how well the models perform on the treated population. (See § 1.8 for details)

We present the PR, ROC and calibration curves for the child welfare screening task in Figure 1.2. The observational evaluation suggests that the observational model performs better. The control evaluation suggests that the counterfactual and observational models of risk perform equally well. Our DR evaluation suggests the counterfactual model has both better discrimination and calibration in estimating the probability of re-referral under screen-out. In Figure 1.2c, the observational evaluation suggests that the observational model is well-calibrated whereas the counterfactual model is overestimating risk; this is expected because the counterfactual model assesses risk under no investigation whereas the observed outcomes include cases whose risk was mitigated by child welfare services. The control evaluation suggests that the two models are similarly calibrated. The DR evaluation shows that the counterfactual model is well-calibrated and the observational model underestimates risk. This makes intuitive sense because the observational model is not accounting for that fact that treatment likely reduced risk for the screened-in cases.

We see further evidence that the observational model performs poorly on the treated population in the drop in ROC curves between the control evaluation and DR evaluation in Figure 1.2b. Deploying such a model would mean failing to identify the people who need and would benefit from treatment. The observational and control evaluations do not show this significant limitation; DR evaluation is the only evaluation that illustrates the poor performance of the observational model on the treated population.

We also evaluate the different models according to whether they are equally predictive, in the sense of being equally well calibrated, across racial groups. Research suggests child welfare processes may disproportionately involve black families (Dettlaff et al., 2011). Here we ask whether the observational or counterfactual model is more equitable. We compare calibration rates by race in Figure 1.3. The observational evaluation suggests that the counterfactual model of risk is poorly calibrated by race. The DR evaluation shows that the counterfactual model is well-calibrated by race and indicates that the observational model underestimates risk on both black and white cases.

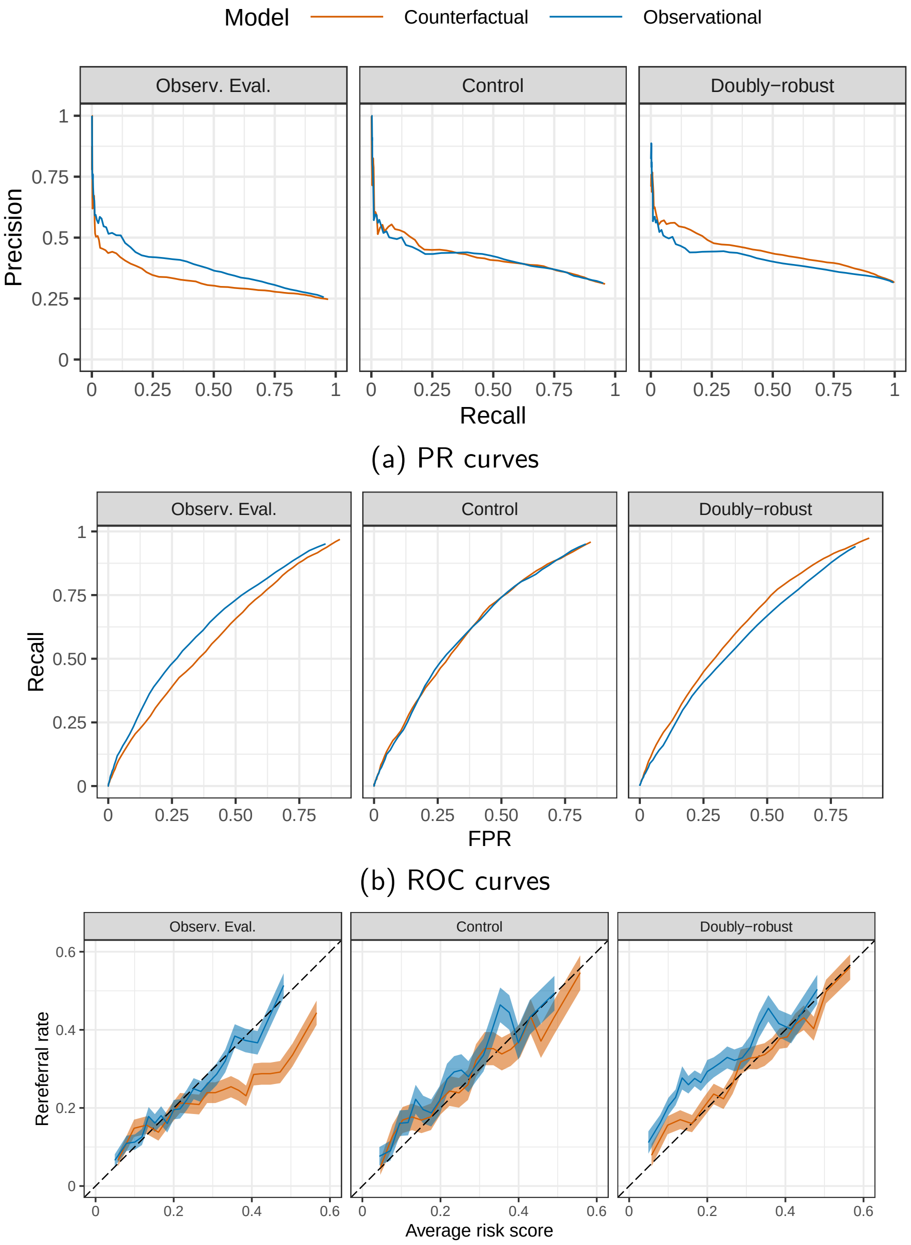
Overall the observational evaluation suggests that the observational model performs better whereas the DR evaluation suggests the counterfactual model performs better. Since we do not have access to the true counterfactual to validate these results, we further consider how well the models align with expert assessment of risk.

Expert Evaluation

At various stages in the child welfare process, social workers assign treatment based on their assessment of risk. Social workers sequentially make three treatment decisions:

1. Whether to screen in a case for investigation;
2. Whether to offer services for a case under investigation; and
3. Whether to place a child out-of-home after an investigation.

Assuming that social workers are competent at assessing risk, we expect the group placed out-of-home (3) to have the highest risk distribution, followed by the group offered services (2), followed by those screened in, and finally we expect the screened out group to have the lowest risk. Figure 1.4 shows that the counterfactual model exhibits this expected behavior whereas the observational model does not. The observational model assesses the screened



(a) PR curves

(b) ROC curves

(c) Calibration curves. 95% pointwise confidence bounds shown.

Figure 1.2: Child welfare results for several approaches to evaluating risk assessments. Our evaluation (DR) is the only method that exposes the poor performance of the observational model on treated cases. The control evaluation suggests the two models perform similarly on cases that did not receive treatment. The observational evaluation would suggest the observational model has better discrimination and calibration than the counterfactual model because it evaluates against the observed outcomes which include cases whose risk was likely mitigated by child welfare services.(See § 1.9 and 1.9 for details)

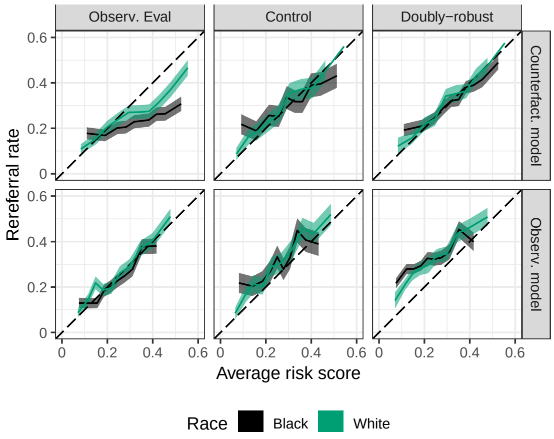


Figure 1.3: Calibration curves by race for child welfare data. The counterfactual model is well-calibrated by race according to the control and DR evaluations but shows inequities according to the observational evaluation because black cases were more likely to get treatment which mitigates risk (see § 1.9 for more details). The observational model is poorly calibrated for both black and white cases according to the DR evaluation.

out population to have more high risk cases than any other treatment group. This suggests that the observational model may be underestimating risk on the treated groups (investigated, services, and placed) since it fails to account for any risk-mitigating effects of these treatments. The observational model underestimates risk on those who were assigned effective treatments. These cases *should* be assigned treatment, but the observational model would suggest that they are low risk and should be screened out.

Such a mistake can have cascading effects downstream. We are particularly concerned about screening out cases that, had they been screened in, would have been accepted for services or placed out-of-home. Figure 1.5 shows the recall for placed cases and serviced cases as we vary the proportion of cases classified as high-risk. This plot shows that at any proportion the counterfactual model has significantly higher recall for both services and placement cases. In particular, at the 0.5 proportion (which is the rate of screen in), the counterfactual model screens in 74% of cases that were placed whereas the observational model only screens in 53%. At the 0.5 proportion the counterfactual model screens in 69% of cases that were accepted for services versus 31% for the observational model.

Task adaptation: Predicting Placement

Another way to evaluate the models is to assess their performance on related risk tasks. While the counterfactual risk models $\mathbb{E}[Y^0|X]$, we can assess how well it estimates $\mathbb{E}[Y^1|X]$, which is the risk under investigation. If we have reason to believe there will be common risk factors for risk under no investigation and risk under investigation, then we expect our model to perform well on this task. We use placement out-of-home, an adverse child welfare outcome that is observed for cases under investigation.

Table 1.1 shows the area under the ROC and PR curves for the placement task. The observational model performs worse than a random classifier, whereas the counterfactual model shows some degree of discrimination. This suggests that the counterfactual model is learning a risk model that is useful in related risk tasks whereas the observational model is not.

The comparison to expert assessment of risk and the performance on a downstream risk task support the conclusions of our DR evaluation: the counterfactual model outperforms the

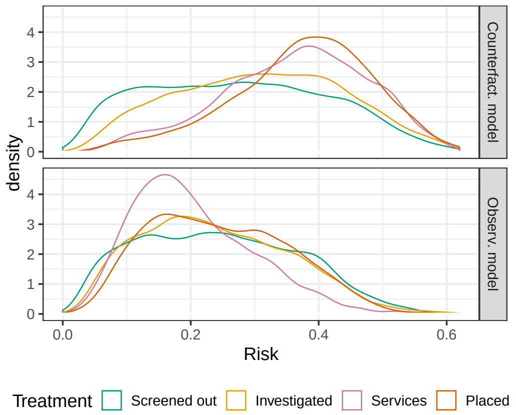


Figure 1.4: Child welfare risk distributions by treatment type for counterfactual and observational risk models. We expect risk to increase with the severity of treatment assigned, with 'Placed' out-of-home having the highest risk distribution and 'Screened out' of investigation having the lowest (see § 1.9). The counterfactual model displays this expected trend whereas the observational model does not. The observational model underestimates risk on cases where child welfare effectively mitigated the risk

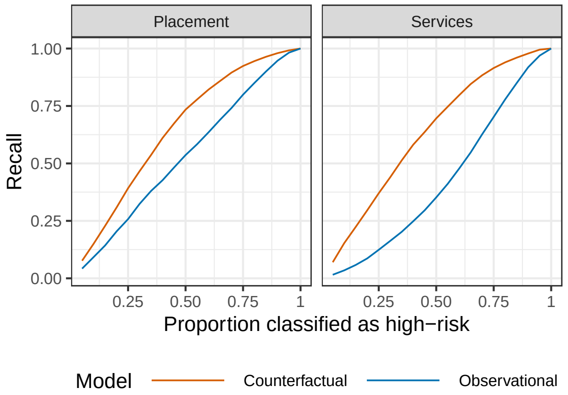


Figure 1.5: Recall of the counterfactual and observational risk models for downstream child welfare decisions. At current screen-in rates (0.5), the observational model would screen out nearly 50% of very high risk cases that were placed out-of-home. The counterfactual model has higher recall at 73%. The gap is even larger for cases that were accepted for services. (See § 1.9).

observational model. In decision-making contexts, failure to account for treatment effects can lead one to the wrong conclusions about model performance, even potentially leading to the deployment of a model that underestimates risk for those who stand to gain most from treatment. In the next section, we consider how failure to account for treatment effects can impact fairness.

1.10 Problem Formulation for Algorithmic Fairness

Standard observational notions of algorithmic fairness are subject to the same pitfalls as observational model evaluation. In this section we propose counterfactual formulations of several fairness metrics and analyze the conditions under which the standard (observational) metric implies the counterfactual one.

	Observ. model	Counterfact. model	Random
AUROC	0.48 (0.46,0.49)	0.62 (0.61,0.63)	0.50
AUPR	0.13 (0.11,0.14)	0.18 (0.16,0.19)	0.14

Table 1.1: Empirical analysis of whether the models of re-referral risk transfer to related risk tasks in the child welfare domain. This table presents the area under ROC and PR curves using the models of re-referral risk to predict a related risk task, out-of-home placement (95% confidence intervals given in parentheses). The observational model performs worse than a random classifier. The counterfactual model performs better; it learns a model of risk that transfers to related risk tasks whereas the observational model does not. (See § 1.9)

We motivate the importance of defining these metrics counterfactually with an example. Suppose teachers are assessing the effectiveness and fairness of a model that predicts who is likely to fail an exam which they intend to use to assign tutoring resources. Suppose anyone tutored will pass. The tutoring session conflicts with girls' sports practice so only male students are tutored. A model that perfectly predicts who will fail without the help of a tutor will have a higher observational FPR for men than women because some male students were tutored, which enabled them to pass. It would be wrong to conclude that this model is unfair with regards to FPR. Someone who would have been high-risk had they not been treated but whose risk was mitigated under treatment should not be considered a false positive. Failure to make this distinction could lead to unfairness, not only in settings where the treatment assignment varies according to the protected attribute but also in settings where the risk under treatment varies according to the protected attribute, as we can see in the next example.

Suppose that the classroom next door is also evaluating the model. This classroom offers tutoring during lunch so girls and boys both can attend; however they hired a tutor who happens to only be effective in preparing male students to pass. The teachers don't know this and randomly assign this tutor to students regardless of gender. The model that perfectly predicts who will fail without a tutor has a higher observational FPR for men, but as before, it is wrong to conclude that the model is unfair with regards to FPR.

We distinguish our notion of counterfactual fairness from prior work which considered counterfactuals of the protected attribute (Kusner et al., 2017; Kilbertus et al., 2017; Wang et al., 2019), an approach which is counterproductive in our settings of interest. Consider a female student who is at high risk of failing because of gender discrimination at home or in the classroom e.g. parents or previous teachers have not given her the support they would have had she been male. Treating this student "counterfactually as if she had been male all along" may suggest that we should not assign this student a tutor. In fact we *must* assign her a tutor in order to correct historical discrimination. Similar arguments can be made in settings like child welfare screening and loan approvals.

1.11 Theoretical Results for Algorithmic Fairness

For three definitions of fairness (parity), we show that observational parity implies counterfactual parity if and only if a balance condition holds. We further show that an independence condition is sufficient for observational parity to imply counterfactual parity. We discuss why it is generally unlikely that the independence condition holds and even more unlikely that the finer balance condition holds when the independence condition fails.

Base Rate Parity

Base rate plays a core role in statistical definitions of fairness (also known as group fairness). Base rate parity is similar to the fairness notion of demographic parity, which requires $\hat{Y} \perp A$ (Dwork et al., 2012a; Calders et al., 2009; Zafar et al., 2015). In § 1.12, we perform empirical analysis on a fairness corrective method that targets base rate parity in order to encourage demographic parity (Kamiran and Calders, 2012). A related fairness notion, prediction-prevalence parity, requires $\mathbb{E}[Y \mid a] = \mathbb{E}[\hat{Y} \mid a]$. Satisfying both prediction-prevalence parity and demographic parity requires parity in the base rates. We distinguish observational base rate parity (oBP) $Y \perp A$ from counterfactual base rate parity (cBP), which requires $Y^0 \perp A$, where Y^0 is the potential outcome under the baseline treatment.

Theorem 1.11.1 (Base Rate Parity)**.** Assume $P(T = 0 \mid y^0, a) \neq 0$. If oBP holds, then cBP holds if and only if the following balance condition holds.

Condition 1.11.0.1 (balBP)**.**

$$\begin{aligned} &P(Y^1 = y)P(T = 1 \mid Y^1 = y) - P(Y^1 = y \mid a)P(T = 1 \mid Y^1 = y, a) \\ &= P(Y^0 = y)\Big(P(T = 1 \mid Y^0 = y) - P(T = 1 \mid Y^0 = y, a)\Big) \end{aligned} \tag{1.13}$$

BalBP holds under the following independence conditions, which provide sufficient conditions for oBP to imply cBP.

Condition 1.11.0.2 (indBP)**.**

$$\begin{gathered} T \perp A \mid Y^0 \\ (Y^1, T) \perp A \end{gathered} \tag{1.14}$$

It is unlikely that indBP (1.14) holds in many contexts. In settings such as child welfare screening and criminal justice, research suggests that even when controlling for the true risk, certain races are more likely to receive treatment (Dettlaff et al., 2011; Alexander, 2011; Mauer, 2010). indBP cannot hold in these settings since $T \not\perp A \mid Y^0$. Even in settings where there is no such bias, indBP will not hold if the risk distributions under treatment vary by protected attribute since indBP requires that $Y^1 \perp A$. indBP also requires $T \perp A \mid Y^1$, which forbids discrimination in treatment assignment when controlling for risk under treatment. If indBP does not hold, it is possible that balBP (1.13) still holds if the conditional and marginal probabilities are such that all terms in Condition 1.13 exactly cancel; however there is no semantic reason why this should hold. Theorem 1 assumes $P(T = 0 \mid y^0, a) \neq 0$, a mild positivity-like assumption that holds in all settings that are suitable for algorithmic risk assessment. Violations of this assumption indicate either completely perfect or imperfect treatment assignment historically for a demographic group.

Proof of Base Rate Necessary Condition. By consistency $Y = TY^1 + (1 - T)Y^0$. Then we have $P(Y = y) =$

$$P(Y^1 = y)\mathbb{P}(T = 1 \mid Y^1 = y) + \P(Y^0 = y)P(T = 0 \mid Y^0 = y)$$

Likewise for $P(Y = y \mid a) =$

$$P(Y^1 = y \mid a)P(T = 1 \mid Y^1 = y, a) + P(Y^0 = y \mid a)P(T = 0 \mid Y^0 = y, a)$$

By oBP, $P(Y=y)=P(Y=y\mid a)$. We assume cBP holds so $P(Y^0=y)=P(Y^0=y\mid a)$. Then, we have

$$\begin{aligned}&P(Y^1=y)P(T=1\mid Y^1=y)-P(Y^1=y\mid a)P(T=1\mid Y^1=y,a)\\&=P(Y^0=y)\Big(P(T=1\mid Y^0=y)-P(T=1\mid Y^0=0,a)\Big)\end{aligned}$$

□

Proof of Base Rate Parity Sufficiency.

$$\begin{aligned}P(Y=1\mid a)&=P(TY^1+(1-T)Y^0=1\mid a)\\&=P(TY^1=1)+P\big((1-T)Y^0=1\mid a\big)\end{aligned}$$

where the first line used consistency and the second line applied linearity of expectation and $(Y^1,T)\perp A$. By oBP, $P(Y=1)=P(Y=1\mid a)$, so it must be true that

$$\begin{aligned}&P\big((1-T)Y^0=1\big)=P\big((1-T)Y^0=1\mid a\big)\implies(T,Y^0)\perp A\\&\implies Y^0\perp A\end{aligned}$$

□

Predictive parity

Base parity and demographic parity may be ill-suited for settings where base rates differ by protected attribute due to disparate needs. Here we may instead desire parity in an error metric, such as precision. Positive predictive parity requires the precision (also known as positive predictive value) to be independent of the protected attribute, and negative predictive parity requires the negative predictive value to be independent of the protected attribute (Chouldechova, 2017; Kleinberg et al., 2016). We define observational Predictive Parity (oPP) as $Y\perp A\mid\hat{Y}=\hat{y}$ and counterfactual Predictive Parity (cPP) as $Y^0\perp A\mid\hat{Y}=\hat{y}$ where $\hat{y}=0$ corresponds to negative predictive parity and $\hat{y}=1$ corresponds to positive predictive parity.

Theorem 1.11.2 (Predictive Parity)**.** Assume $P(T=0\mid y^0,a,\hat{y})\neq 0$. If oPP holds, then cPP holds if and only if the following balance condition holds.

Condition 1.11.0.3 (balPP)**.**

$$\begin{aligned}&P(Y^1=y\mid\hat{y})P(T=1\mid Y^1=y,\hat{y})\\&-P(Y^1=y\mid a,\hat{y})P(T=1\mid Y^1=y,a,\hat{y})\\&=P(Y^0=y\mid\hat{y})\Big(P(T=1\mid Y^0=y,\hat{y})-P(T=1\mid Y^0=y,a,\hat{y})\Big)\end{aligned}\tag{1.15}$$

BalPP is satisfied under the following independence conditions, which provide sufficient conditions for oPP to imply cPP.

Condition 1.11.0.4 (indPP)**.**

$$\begin{aligned}T\perp A\mid Y^0,\hat{Y}\\(Y^1,T)\perp A\mid\hat{Y}\end{aligned}\tag{1.16}$$

IndPP will not hold in many settings. Note that $(Y^1, T) \perp A \mid \hat{Y} \iff T \perp A \mid Y^1, \hat{Y}$ and $Y^1 \perp A \mid \hat{Y}$. Conditions $T \perp A \mid Y^t, \hat{Y}$ require $\hat{Y}$ to contain all the information that A tells us about treatment assignment that is not contained in Y^t. Since $\hat{Y}$ is typically trained to predict Y and not T, it is quite unlikely that these conditions will hold in settings where there is bias in treatment assignment even when controlling for true risk. Condition $Y^1 \perp A \mid \hat{Y}$ allows differences in the risk distribution under treatment if we can fully explain these differences with $\hat{Y}$. In the best case $\hat{Y} \approx Y$, but it is unlikely that the observed outcome, which is not causally well-defined, would explain differences in the risk distribution under treatment. As above, even if indPP does not hold, balPP may hold but it is difficult to reason why this should hold in any setting. Like Theorem 1, Theorem 2 also assumes a mild positivity-like assumption that is reasonable in risk assessment settings.

The proofs use the same techniques as for base rate parity.

Equalized odds

In settings where TPR and FPR are more important than predictive value, we may desire parity in TPR and FPR, a fairness notion known as Equalized Odds (Hardt et al., 2016b). Let observational Equalized Odds (oEO) require that $\hat{Y} \perp A \mid Y$ and counterfactual Equalized Odds (cEO) require that $\hat{Y} \perp A \mid Y^0$.

Theorem 1.11.3 (Equalized Odds)**.** Assume $P(Y = y \mid a) \neq 0$ and $P(T = 0 \mid y^0, a, \hat{y}) \neq 0$. If oEO holds, then cEO holds if and only if the following balance condition holds. .

Condition 1.11.0.5 (balEO)**.**

$$\begin{aligned}
& P(\hat{Y} = 1 \mid Y^1 = y)\frac{P(T = 1 \mid \hat{Y} = 1, Y^1 = y)P(Y^1 = y)}{P(Y = y)} \\
& - P(\hat{Y} = 1 \mid Y^1 = y, a)\frac{P(T = 1 \mid \hat{Y} = 1, Y^1 = y, a)P(Y^1 = y \mid a)}{P(Y = y \mid a)} \\
& = P(\hat{Y} = 1 \mid Y^0 = y)\left(\frac{P(T = 0 \mid \hat{Y} = 1, Y^0 = y, a)P(Y^0 = y \mid a)}{P(Y = y \mid a)}\right. \\
& \left. - \frac{P(T = 0 \mid \hat{Y} = 1, Y^0 = y)P(Y^0 = y)}{P(Y = y)}\right)
\end{aligned} \tag{1.17}$$

The balance condition is satisfied under the following independence conditions, which comprise sufficient conditions for oEO to imply cEO.

Condition 1.11.0.6 (indEO)**.**

$$\begin{aligned}
Y &\perp A \\
Y^0 &\perp A \\
T &\perp A \mid \hat{Y}, Y^0 \\
(Y^1, \hat{Y}, T) &\perp A
\end{aligned} \tag{1.18}$$

The first two conditions of indEO require oBP and cBP, so indEO requires balBP to hold. In settings where there is discrimination in treatment assignment even when controlling for true risk, indEO is unlikely to hold. Even if there is no such discrimination, indEO will not hold if there are differences in the risk distributions under treatment since the last condition of 1.18 requires $Y^1 \perp A$. indEO requires further conditions such as parity in the TPR/FPR against the outcome under treatment. If these conditions are not met, oEO could imply cEO if balEO holds, but it is difficult to reason about why this would hold for a setting when the independencies do not. Theorem 3 assumes two mild assumptions: the positivity-like assumption of Theorem 2 and $P(Y^0 = y \mid a) \neq 0$.

Proof that BalEO is Necessary and Sufficient. We first expand $P(\hat{Y} = 1 \mid Y = y)$

$$\begin{aligned}
&= \frac{P(\hat{Y} = 1, Y = y)}{P(Y = y)} \\
&= \frac{P(\hat{Y} = 1, Y^1 = y, T = 1) + P(\hat{Y} = 1, Y^0 = y, T = 0)}{P(Y = y)}
\end{aligned} \tag{1.19}$$

which we can further expand to get $P(\hat{Y} = 1 \mid Y = y)$

$$\begin{aligned}
&= \frac{P(T = 1 \mid \hat{Y} = 1, Y^1 = y)P(\hat{Y} = 1 \mid Y^1 = y)P(Y^1 = y)}{P(Y = y)} \\
&+ \frac{P(T = 0 \mid \hat{Y} = 1, Y^0 = y)P(\hat{Y} = 1 \mid Y^0 = y)P(Y^0 = y)}{P(Y = y)}
\end{aligned} \tag{1.20}$$

Since oEO holds by assumption, then $P(\hat{Y} = 1 \mid Y = y) = P(\hat{Y} = 1 \mid Y = y, A = a)$. Using the expansion in Equation 1.20, we have

$$\begin{aligned}
&\frac{P(T = 1 \mid \hat{Y} = 1, Y^1 = y)P(\hat{Y} = 1 \mid Y^1 = y)P(Y^1 = y)}{P(Y = y)} \\
&+ \frac{P(T = 0 \mid \hat{Y} = 1, Y^0 = y)P(\hat{Y} = 1 \mid Y^0 = y)P(Y^0 = y)}{P(Y = y)} \\
&= \frac{P(T = 1 \mid \hat{Y} = 1, Y^1 = y, a)P(\hat{Y} = 1 \mid Y^1 = y, a)P(Y^1 = y \mid a)}{P(Y = y \mid a)} \\
&+ \frac{P(T = 0 \mid \hat{Y} = 1, Y^0 = y, a)P(\hat{Y} = 1 \mid Y^0 = y, a)P(Y^0 = y \mid a)}{P(Y = y \mid a)}
\end{aligned} \tag{1.21}$$

Rearranging gives

$$\begin{aligned}
&P(\hat{Y} = 1 \mid Y^1 = y)\frac{P(T = 1 \mid \hat{Y} = 1, Y^1 = y)P(Y^1 = y)}{P(Y = y)} \\
&- P(\hat{Y} = 1 \mid Y^1 = y, a)\frac{P(T = 1 \mid \hat{Y} = 1, Y^1 = y, a)P(Y^1 = y \mid a)}{P(Y = y \mid a)} \\
&= -P(\hat{Y} = 1 \mid Y^0 = y)\frac{P(T = 0 \mid \hat{Y} = 1, Y^0 = y)P(Y^0 = y)}{P(Y = y)} \\
&+ P(\hat{Y} = 1 \mid Y^0 = y, a)\frac{P(T = 0 \mid \hat{Y} = 1, Y^0 = y, a)P(Y^0 = y \mid a)}{P(Y = y \mid a)}.
\end{aligned} \tag{1.22}$$

Necessary For oEO to imply cEO, both conditions must hold. By cEO, $P(\hat{Y}=1 \mid Y^0 = y) = P(\hat{Y}=1 \mid Y^0 = y, A=a)$ which would imply that

$$\begin{aligned}
&P(\hat{Y}=1 \mid Y^1 = y)\frac{P(T=1 \mid \hat{Y}=1, Y^1=y)P(Y^1=y)}{P(Y=y)} \\
&- P(\hat{Y}=1 \mid Y^1 = y, a)\frac{P(T=1 \mid \hat{Y}=1, Y^1=y, a)P(Y^1=y \mid a)}{P(Y=y \mid a)} \\
&= P(\hat{Y}=1 \mid Y^0 = y)\left(\frac{P(T=0 \mid \hat{Y}=1, Y^0=y, a)P(Y^0=y \mid a)}{P(Y=y \mid a)}\right. \\
&\left. - \frac{P(T=0 \mid \hat{Y}=1, Y^0=y)P(Y^0=y)}{P(Y=y)}\right)
\end{aligned} \tag{1.23}$$

Sufficient In addition to oEO, we assume balEO holds. From oEO we have equation 1.22 and balEO is equation 1.23. The left-hand sides of equations 1.22 and 1.23 are the same so by the transitive property,

$$\begin{aligned}
&- P(\hat{Y}=1 \mid Y^0 = y)\frac{P(T=0 \mid \hat{Y}=1, Y^0=y)P(Y^0=y)}{P(Y=y)} \\
&+ P(\hat{Y}=1 \mid Y^0 = y, a)\frac{P(T=0 \mid \hat{Y}=1, Y^0=y, a)P(Y^0=y \mid a)}{P(Y=y \mid a)} \\
&= P(\hat{Y}=1 \mid Y^0 = y)\left(\frac{P(T=0 \mid \hat{Y}=1, Y^0=y, A=a)P(Y^0=y \mid a)}{P(Y=y \mid a)}\right. \\
&\left. - \frac{P(T=0 \mid \hat{Y}=1, Y^0=y)P(Y^0=y)}{P(Y=y)}\right)
\end{aligned} \tag{1.24}$$

Simplifying gives

$$\begin{aligned}
&P(\hat{Y}=1 \mid Y^0 = y, a)\frac{P(T=0 \mid \hat{Y}=1, Y^0=y, a)P(Y^0=y \mid a)}{P(Y=y \mid a)} \\
&= P(\hat{Y}=1 \mid Y^0 = y)\left(\frac{P(T=0 \mid \hat{Y}=1, Y^0=y, a)P(Y^0=y \mid a)}{P(Y=y \mid a)}\right)
\end{aligned} \tag{1.25}$$

Assuming $P(T=0 \mid \hat{Y}=1, y^0, a) \neq 0$ and $P(Y=y \mid a) \neq 0$, then we conclude that $P(\hat{Y}=1 \mid Y^0 = y, a) = P(\hat{Y}=1 \mid Y^0 = y) \implies cEO$

□

indEO Sufficiency The following conditions are sufficient for oEO to imply cEO:

$$\begin{aligned}
Y &\perp A \\
Y^0 &\perp A \\
T \perp A &\mid \hat{Y}, Y^0 \\
(Y^1, \hat{Y}, T) &\perp A
\end{aligned} \tag{1.26}$$

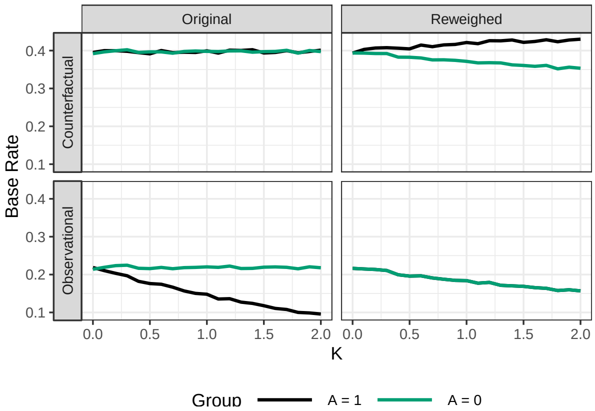


Figure 1.6: Counterfactual and observational base rates before and after applying a fairness-corrective method that reweighs training data (§ 1.12). X-axis controls the bias of treatment assignment toward group $A = 1$. Before reweighing (“Original"), counterfactual base rates are equal (cBP holds), but observational base rates are different (oBP doesn't hold) for $k > 0$ since group $A = 1$ is more likely to get treated. Reweighing achieves oBP but cBP no longer holds.

Proof of indEO Sufficiency. By contraction, the indEO conditions are equivalently written as $\hat{Y} \perp A \mid Y^1$; $Y \perp A$; $Y^1 \perp A$; $Y^0 \perp A$; $T \perp A \mid \hat{Y}, Y^0$; $T \perp A \mid \hat{Y}, Y^1$. Under these assumptions, both sides of Equation 1.23 are 0, so the balEO condition holds under these independencies. Since balEO is sufficient, then indEO is sufficient for oEO to imply cEO. □

Our theoretical analysis suggests that in many settings equalizing the observational fairness metric will not equalize the counterfactual fairness metric. We conclude by noting that the theorems hold when conditioning on any feature(s) $\subseteq X$, and in this context, these theorems are relevant to individual notions of fairness.

1.12 Empirical Results for Algorithmic Fairness on Synthetic Data

We empirically demonstrate that equalizing the observational metric via fairness-corrective methods can increase disparity in the counterfactual metric on the synthetic data described in § 1.8.[10]

Reweighing

One approach to encourage demographic parity reweighs the training data to achieve base rate parity (Kamiran and Calders, 2012). Figure 1.6 shows that without any processing (“Original"), the counterfactual base rates are equal while the observational base rates show increasing disparity with K. Reweighing applied to the observational outcome achieves oBP but induces disparity in the counterfactual base rate. Theorem 1.11.1 suggested this result: For $k > 0$, $A \not\perp T \mid Y^0$; then it is unlikely that oBP implies cBP.

[10] We do not perform this empirical analysis on the child welfare data since it is balanced in terms of base rates and FPR/TPR with respect to race.

Group	Method	cGFNR	cGFPR	oGFNR	oGFPR
A=1	Original	0.50	0.33	0.58	0.39
A=0	Original	0.50	0.33	0.56	0.39
A=1	Post-Proc.	0.58	0.30	0.63	0.35
A=0	Post-Proc.	0.64	0.34	0.63	0.35

Table 1.2: Empirical results on synthetic data show that post-processing methods to achieve parity in standard group fairness metrics can induce unfairness in the counterfactual fairness metric. This table gives the counterfactual and observational generalized FNR/FPR before and after post-processing to equalize odds (§ 1.12) using threshold $= 0.5$. Before post-processing (“Original"), the counterfactual generalized rates (cGFNR and cGFPR) are the same for both groups. Post-processing equalizes the observational rates (oGFNR and oGFPR) but induces noticeable disparity in both cGFNR and cGFPR.

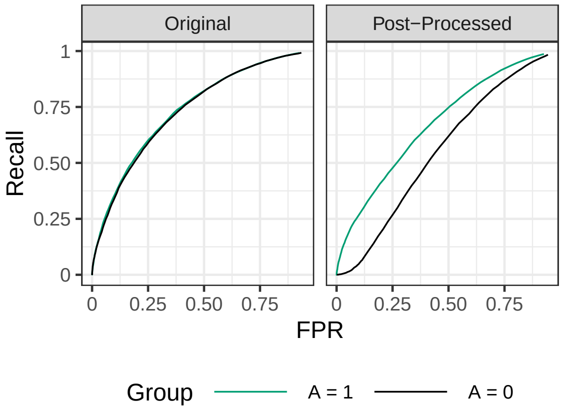


Figure 1.7: Counterfactual ROC curves before and after post-processing to equalize odds (§ 1.12). Before post-processing, ROC curves are identical for both groups, indicating that counterfactual equalized odds (cEO) holds. Post-processing induces imbalance, harming group $A = 0$ and compounding initial unfairness in treatment assignment.

Post-processing for equalized odds

We evaluate a method that modifies scores to achieve a generalized version of equalized odds (Pleiss et al., 2017; Hardt et al., 2016b).[11] This method targets parity in the generalized FNR/FPR, where GFPR is $\mathbb{E}[\hat{f}(X) \mid Y = 0]$ and GFNR is $\mathbb{E}[1 - \hat{f}(X) \mid Y = 1]$. We refer to these observational rates as oGFPR/oGFNR and define their counterfactual counterpart: cGFPR $= \mathbb{E}[\hat{f}(X) \mid Y^0 = 0]$ and cGFNR $= \mathbb{E}[1 - \hat{f}(X) \mid Y^0 = 1]$. We use the scores of the counterfactual model as inputs. We compute the cGFNR and cGFPR using our DR method from § 1.7.[12]

Table 1.2 shows that post-processing to equalize oGFPR and oGFNR induces imbalance in cGFPR and cGFNR. In Figure 1.7 we see that the original model achieved cEO, but post-processing induced disparity to the detriment of the group that was less likely to be treated. Since treatment is beneficial, this “fairness" adjustment actually compounded the discrimination in the treatment assignment.

[11] We use the Pleiss implementation on https://github.com/gpleiss/equalized_odds_and_calibration that extends the method in Hardt et al. (2016b) to probabilistic classifiers.

[12] The estimator is nearly identical to the estimators for FPR/FNR if we use $\hat{f}(X)$ in place of the predicted label $\hat{Y}(X)$.

1.13 Conclusion

This chapter demonstrates that training and evaluating models using observed outcomes produces invalid models that notably misestimate risk for those likely to be receptive to treatment. Furthermore, fairness-correcting methods that seek to achieve observational parity can lead to disparities on the relevant counterfactual metrics, and may further compound inequities in intial treatment assignment. To obtain valid risk assessments, evaluation metrics, and predictive fairness assessments, we developed counterfactual approaches to learning, evaluation and predictive fairness assessment. A key condition for the validity of these approaches is measuring all confounding factors. However, in many consequential decision-making settings, this condition may be hard to attain. We next consider how to proceed when some confounding factors are inaccessible.

2

Counterfactual Predictions under Runtime Confounding

Generally, to learn counterfactual prediction models from observational data on historical decisions and corresponding outcomes, one must measure all factors that jointly affect the outcome and the decision taken. Motivated by decision support applications, we study the counterfactual prediction task in the setting where all relevant factors are captured in the historical data, but it is infeasible, undesirable, or impermissible to use some such factors in the prediction model. We refer to this setting as **runtime confounding**. We propose a doubly-robust procedure for learning counterfactual prediction models in this setting. Our theoretical analysis and empirical results suggest that our method often outperforms competing approaches. We also present a validation procedure for evaluating the performance of counterfactual prediction methods. The methods and results presented in this chapter comprise work first published in Coston et al. (2020a).

Runtime confounding naturally arises in a number of different settings. First, relevant factors may not yet be available at the desired runtime. For instance, in child welfare screening, call workers decide which allegations coming in to the child abuse hotline should be investigated based on the information in the call and historical administrative data (Chouldechova et al., 2018). The call worker's decision-making process can be informed by a risk assessment if the call worker can access the risk score in real-time. Since existing case management software cannot run speech/NLP models in realtime, the call information (although recorded) is not available at runtime, thereby leading to runtime confounding. Second, runtime confounding arises when historical decisions and outcomes have been affected by sensitive or protected attributes which for legal or ethical reasons are deemed ineligible as inputs to algorithmic predictions. We may for instance be concerned that call workers implicitly relied on race in their decisions, but it would not be permissible to include race as a model input. Third, runtime confounding may result from interpretability or simplicity requirements. For example, a university may require algorithmic tools used for case management to be interpretable. While information conveyed during student-advisor meetings is likely informative both of case management decisions and student outcomes, natural language processing models are not classically interpretable, and thus the university may wish instead to only use structured information like GPA in their tools.

Drawing upon techniques used in low-dimensional treatment effect estimation (Van der Laan et al., 2003; Zimmert and Lechner, 2019; Chernozhukov et al., 2018b), we develop a procedure for the full pipeline of learning and evaluating prediction models under runtime confounding. We (1) formalize the problem of counterfactual prediction with runtime confounding [§ 2.2];

(2) propose a solution based on doubly-robust techniques that has desirable theoretical properties [§ 2.4.2]; (3) theoretically and empirically compare this solution to an alternative counterfactually valid approach as well as the standard practice, describing the conditions under which we expect each to perform well [§ 2.4 & 2.7]; and (4) provide an evaluation procedure to assess performance of the methods in the real-world [§ 2.6].

2.1 Background and Related Work

This chapter builds on the work in the previous chapter. As before, our goal here is to predict outcomes under a proposed decision (interchageably referred to as 'treatment' or 'intervention') in order to inform human decision-makers about what is likely to happen under that treatment.

Our proposed prediction (Contribution 2) and evaluation methods (Contribution 4) draw upon the literature on double machine learning and doubly-robust estimation, which uses the efficient influence function to produce estimators with reduced bias (Van der Laan et al., 2003; Robins et al., 1994; Robins and Rotnitzky, 1995; Kennedy, 2016; Chernozhukov et al., 2018a; Kennedy, 2020). These techniques are commonly used for treatment effect estimation, and of particular note for our setting are methods for estimating treatment effects conditional on only a subset of confounders (Semenova and Chernozhukov, 2017; Chernozhukov et al., 2018b; Zimmert and Lechner, 2019; Fan et al., 2020). Semenova and Chernozhukov (2017) propose a two-stage doubly-robust procedure that uses series estimators in the second stage to achieve asymptotic normality guarantees. Zimmert and Lechner (2019) propose a similar approach that uses local constant regression in the second stage, and Fan et al. (2020) propose using a local linear regression in the second stage. These approaches can obtain rate double-robustness under the notably strict condition that the product of nuisance errors to converge faster than $\sqrt{n}$ rates. In a related work, Foster and Syrgkanis (2019) proposes an orthogonal estimator of treatment effects which, under certain conditions, guarantees the excess risk is second-order but not doubly robust.[1] Our work is most similar to the approach taken in Kennedy (2020), which proposes a model-agnostic two-stage doubly robust estimation procedure for conditional average treatment effects that attains a model-free doubly robust guarantee on the prediction error. Treatment effects can be identified under weaker assumptions than required to individual the potential outcomes, and prior work has proposed a procedure to find the minimal set of confounders for estimating conditional treatment effects (Makar et al., 2019).

Our prediction task is different from the common causal inference problem of treatment effect estimation, which targets a contrast of outcomes under two different treatments (Wager and Athey, 2018; Shalit et al., 2017). Treatment effects are useful for describing responsiveness to treatment. While responsiveness is relevant to some types of decisions, it is insufficient, or even irrelevant, to consider for others. For instance, a doctor considering an invasive procedure may make a different recommendation for two patients with the same responsiveness if one has a good probability of successful recovery without the procedure and the other does not. In lending settings, the responsiveness to different loan terms is irrelevant; all that matters is that the likelihood of default be sufficiently small under feasible terms. In such settings, we are interested in *predictions* conditional on only those features that are permissible or desirable to consider at runtime. Our methods are specifically designed for minimizing prediction error,

[1] A second order but not doubly robust guarantee requires sufficiently fast rates on *both* nuisance functions. By contrast, rate double robustness imposes a weaker assumption on the *product* of nuisance function errors, allowing e.g., fast rates on the propensity function and slow rates on the outcome regression function.

rather than providing inferential guarantees such as confidence intervals, as is common in the treatment effect estimation setting.

The practical challenge that we often need to make decisions based on only a subset of the confounders has been discussed in the policy learning literature (Zhang et al., 2012; Athey and Wager, 2017; Kitagawa and Tetenov, 2018). For instance, it may be necessary to use only a subset of confounders to meet ethical requirements, model simplicity desiderata, or budget limitations (Athey and Wager, 2017). Doubly robust methods for learning treatment assignment policies have been proposed for such settings (Zhang et al., 2012; Athey and Wager, 2017).

Our work is also related to the literature on marginal structure models (MSMs) (Robins et al., 2000; Robins, 2000). An MSM is a model for a marginal mean of a counterfactual, possibly conditional on a subset of baseline covariates. The standard MSM approach is semiparametric, employing parametric assumptions for the marginal mean but leaving other components of the data-generating process unspecified (Van der Laan et al., 2003). Nonparametric variants were studied in the unconditional case for continuous treatments by Rubin and van der Laan (2006). In contrast our setting can be viewed as a nonparametric MSM for a binary treatment, conditional on a large subset of covariates. This is similar in spirit to partly-conditional treatment effect estimation (van der Laan and Luedtke, 2014); however we do not target a contrast since our interest is in predictions rather than treatment effects. Our results are also less focused on model selection (Van Der Laan and Dudoit, 2003), and more on error rates for particular estimators. We draw on techniques for sample-splitting and cross-fitting, which have been used in the regression setting for model selection and tuning (Györfi et al., 2006; Van der Laan et al., 2003) and in treatment effect estimation (Robins et al., 2008; Zheng and van der Laan, 2010; Chernozhukov et al., 2018b).

Our method is relevant to settings where the outcome is selectively observed. This *selective labels* problem (Lakkaraju et al., 2017; Kleinberg et al., 2018) is common in settings like lending where the repayment/default outcome is only observed for applicants whose loan is approved. Runtime confounding can arise in such settings if some factors that are used for decision-making are unavailable for prediction.

Recent work has considered methods to accommodate confounding due to sources other than missing confounders at runtime. A line of work has considered how to use causal techniques to correct runtime dataset shift (Subbaswamy et al., 2018; Magliacane et al., 2018; Subbaswamy and Saria, 2018). In our case the runtime setting is different from the training setting not because of distributional shift but because we can no longer access all confounders. These methods also differ from ours in that they are not seeking to predict outcomes under specific decisions.

There is also a line of work that considers confounding in the *training* data (Kallus and Zhou, 2018a; Madras et al., 2019). While confounded training data is common in various applications, our work targets decision support settings where the factors used by decision-makers are recorded in the training data but are not available for prediction.

Lastly, there are connections between runtime confounding and the literature on privileged learning and algorithmic fairness that use features during training time that are not available for prediction. Learning using Privileged Information (LUPI) has been proposed for settings in which the training data contains additional features that are not available at runtime (Vapnik and Vashist, 2009). In algorithmic fairness, disparate learning processes (DLPs) use the sensitive attribute during training to produce models that achieve a target notion of parity without requiring access to the protected attribute at test time (Lipton et al., 2018). LUPI and DLPs both make use of variables that are only available at train time, but if these variables affect the decisions under which outcomes are observed, predictions from LUPI and DLPs will be confounded because neither accounts for how these variables affect decisions. By contrast, our method uses confounding variables during training to produce valid counterfactual predictions.

2.2 Problem Formulation and Additional Notation

We denote the confounding factors as $X = (V, Z)$ where $V \in \mathcal{V} \subseteq \mathbb{R}^{d_\mathrm{V}}$ are available at both traintime and runtime while $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_\mathrm{Z}}$ are only available for training (not at runtime).[2] Our goal is to predict outcomes under a proposed decision t based on runtime-available predictors V. Using the potential outcomes framework (Rubin, 2005; Neyman, 1923), our prediction target is $\nu_t(v) := \mathbb{E}[Y^t \mid V = v]$ where $Y^t \in \mathcal{Y} \subseteq \mathbb{R}$ is the potential outcome we would observe under treatment $T = t$. We denote the propensity to receive treatment t by $\pi_t(v, z) := P(T = t \mid V = v, Z = z)$. We also define the outcome regression by $\mu_a(v, z) := \mathbb{E}[Y^t \mid V = v, Z = z]$. For brevity, we will generally omit the subscript, using notation ν, π and μ to denote the functions for a generic treatment t.

Definition 2.2.1. Formally, the task of counterfactual prediction under **runtime-only confounding** is to estimate $\nu(v)$ from iid training data (V, Z, T, Y) under the following two conditions:

Condition 2.2.1.1 (Training Ignorability)**.** Decisions are unconfounded given V and Z: $Y^t \perp T \mid V, Z$.

Condition 2.2.1.2 (Runtime Confounding)**.** Decisions are confounded given only V: $Y^t \not\perp T \mid V$; equivalently, $T \not\perp Z \mid V$ and $Y^t \not\perp Z \mid V$

To ensure that the target quantity is identifiable, we require two further assumptions, which are standard in causal inference and not specific to the runtime confounding setting.

Condition 2.2.1.3 (Consistency)**.** A case that receives treatment t has outcome $Y = Y^t$.

Condition 2.2.1.4 (Positivity)**.** $P(\pi_t(V, Z) \geq \epsilon > 0) = 1 \quad \forall t$

Identifications. Under these conditions, we can write the counterfactual regression functions μ and ν in terms of observable quantities. We can identify $\mu(v, z) = \mathbb{E}[Y \mid V = v, Z = z, T = t]$ and our target $\nu(v) = \mathbb{E}[\mathbb{E}[Y \mid V = v, Z = z, T = t] \mid V = v] = \mathbb{E}[\mu(V, Z) \mid V = v]$. The iterated expectation in the identification of ν suggests a two-stage approach that we propose in § 2.4.1 after reviewing current approaches.

[2] For settings where it is impossible to obtain an offline dataset containing Z, an outcome sensitivity model can be used to partially identify the target learning and evaluation estimands (Rambachan et al., 2022).

2.3 Standard Practice for Learning Counterfactual Prediction Models

2.3.1 Standard practice: Treatment-conditional regression (TCR)

As we saw in the previous chapter, standard counterfactual prediction methods train models on the cases that received treatment t (Schulam and Saria, 2017b; Coston et al., 2020b), a procedure we will refer to as **treatment-conditional regression** (TCR). This procedure estimates $\omega(v) = \mathbb{E}[Y \mid T = t, V = v]$. This method works well given access to all the confounders at runtime; if $T \perp Y^t \mid V$, then $\omega(v) = \mathbb{E}[Y^t \mid V = v] = \nu(v)$. However, under runtime confounding, this method does not produce valid counterfactual predictions because $\omega(v) \neq \mathbb{E}[Y^t \mid V = v]$. This method does not target the right counterfactual quantity, and may produce misleading predictions.[3] For instance, consider a risk assessment setting that historically assigned risk-mitigating treatment to cases that have higher risk under the null treatment ($T = 0$). Using TCR to predict outcomes under the null treatment will underestimate risk since $\mathbb{E}[Y \mid V, T = 0] = \mathbb{E}[Y^0 \mid V, T = 0] < \mathbb{E}[Y^0 \mid V]$. We can characterize the bias of this approach by analyzing $b(v) := \omega(v) - \nu(v)$, a quantity we term the pointwise *confounding bias.*

Proposition 2.3.1. Under runtime confounding, $\omega(v)$ has pointwise confounding bias

$$b(v) = \int_{\mathcal{Z}} \mu(v, z)\Big(p(z \mid V = v, T = t) - p(z \mid V = v)\Big)dz \quad \neq \quad 0 \tag{2.1}$$

By Condition 2.2.1.2, this confounding bias will be non-zero. Nonetheless we might expect the TCR method to perform well if $b(v)$ is small enough. We can formalize this intuition by decomposing the error of a TCR predictive model $\hat{\nu}_{\text{TCR}}$ into estimation error and confounding bias:

Proposition 2.3.2. The pointwise regression error of the TCR method can be bounded as follows:

$$\mathbb{E}[(\nu(v) - \hat{\nu}_{\text{TCR}}(v))^2] \lesssim \mathbb{E}[(\omega(v) - \hat{\nu}_{\text{TCR}}(v))^2] + b(v)^2$$

The first term gives the estimation error and the second term bounds the bias in targeting the wrong counterfactual quantity.

2.4 Methodology for Learning Valid Counterfactual Prediction Models under Runtime Confounding

2.4.1 A simple proposal: Plug-in (PL) approach

We can avoid the confounding bias of TCR through a simple two-stage procedure we call the **plug-in** approach that targets the proper counterfactual quantity. This approach, described in Algorithm 1, first estimates μ as $\hat{\mu}$ and then uses $\hat{\mu}$ to construct a pseudo-outcome which is regressed on V to yield prediction $\hat{\nu}_{\text{PL}}$. Cross-fitting techniques (Alg. 2) can be applied to prevent issues that may arise due to potential overfitting when learning both $\hat{\mu}$ and $\hat{\nu}_{\text{PL}}$ on the same training data. Sample-splitting (or cross-fitting) also enables us to get the following upper bound on the error of the PL approach.

[3] Runtime imputation of Z will not eliminate this bias since $E[Y \mid T = t, V = v, f(v)] = \omega(v)$.

Algorithm 1 The plug-in (PL) approach for counterfactual predictions under runtime confounding

```
*Stage 1:* Learn $\hat{\mu}(v,z)$ by regressing $Y \sim V,Z \mid T=t$
*Stage 2:* Learn $\hat{\nu}_{\text{PL}}(v)$ by regressing $\hat{\mu}(V,Z) \sim V$
```

Algorithm 2 The plug-in (PL) approach for counterfactual predictions under runtime confounding with cross-fitting

```
Randomly divide training data into two partitions $\mathcal{W}^1$ and $\mathcal{W}^2$.
for $(p,q) \in \{(1,2),(2,1)\}$ do
    *Stage 1:* On partition $\mathcal{W}^p$, learn $\hat{\mu}^p(v,z)$ by regressing $Y \sim V,Z \mid T=t$
    *Stage 2:* On partition $\mathcal{W}^q$, learn $\hat{\nu}^q_{\text{PL}}(v)$ by regressing $\hat{\mu}^p(V,Z) \sim V$
PL prediction: $\hat{\nu}_{\text{PL}}(v) = \frac{1}{2}\sum_{i=1}^{2} \hat{\nu}^i_{\text{PL}}(v)$
```

We now introduce stability conditions that will be used in the subsequent analysis.

Definition 2.4.1. (Stability conditions) The results assume the following two stability conditions on the second-stage regression estimators:

Condition 2.4.1.1. $\hat{\mathbb{E}}_n[Y \mid V=v] + c = \hat{\mathbb{E}}_n[Y + c \mid V=v]$ for any constant c

Condition 2.4.1.2. For two random variables R and Q, if $\mathbb{E}[R \mid V=v] = \mathbb{E}[Q \mid V=v]$, then

$$\mathbb{E}\left[\left(\hat{\mathbb{E}}_n[R \mid V=v] - \mathbb{E}[R \mid V=v]\right)^2\right] \asymp \mathbb{E}\left[\left(\hat{\mathbb{E}}_n[Q \mid V=v] - \mathbb{E}[Q \mid V=v]\right)^2\right]$$

where $L \asymp R$ denotes $L \lesssim R$ and $R \lesssim L$ and $\hat{\mathbb{E}}_n[Y \mid V=v]$ denotes an estimator of the regression function $\mathbb{E}[Y \mid V=v]$.

The second condition is satisfied for instance by local estimation techniques. While global methods (such as linear regression) may not satisfy this property, a weaker stability condition (see Kennedy (2020)) can be used to achieve a bound on the integrated mean squared error.

Proposition 2.4.1. Under these stability conditions on the 2nd stage estimators and sample-splitting for stages 1 and 2, the PL method has pointwise regression error bounded by

$$\mathbb{E}\left[\left(\hat{\nu}_{\text{PL}}(v) - \nu(v)\right)^2\right] \lesssim \mathbb{E}\left[\left(\tilde{\nu}(v) - \nu(v)\right)^2\right] + \mathbb{E}\left[\left(\hat{\mu}(V,Z) - \mu(V,Z)\right)^2 \mid V=v\right]$$

where the oracle-quantity $\tilde{\nu}(v)$ describes the function we would get in the second-stage if we had oracle access to Y^t.

This simple approach can consistently estimate our target $\nu(v)$. However, it solves a harder problem (estimation of $\mu(v,z)$) than what our lower-dimensional target ν requires. Notably the bound depends *linearly* on the MSE of $\hat{\mu}$. We next propose an approach that avoids such strong dependence.

2.4.2 Our main proposal: Doubly-robust (DR) approach

Our main proposed method is what we call the **doubly-robust** (DR) approach, which improves upon the PL procedure by using a bias-corrected pseudo-outcome in the second stage (Alg. 4). The DR approach estimates both μ and π, which enables the method to perform well in situations in which π is easier to estimate than μ. We propose a cross-fitting (Alg. 3) variant that satisfies the sample-splitting requirements of Theorem 2.5.1.

Algorithm 3 Our doubly-robust (DR) approach for counterfactual predictions under runtime confounding

```
*Stage 1:* Learn $\hat{\mu}(v, z)$ by regressing $Y \sim V, Z \mid T = t$.
           Learn $\hat{\pi}(v, z)$ by regressing $\mathbb{I}\{T = t\} \sim V, Z$
*Stage 2:* Learn $\hat{\nu}_{\text{DR}}(v)$ by regressing $\left(\frac{\mathbb{I}\{T=t\}}{\hat{\pi}(V,Z)}(Y - \hat{\mu}(V, Z)) + \hat{\mu}(V, Z)\right) \sim V$
```

Algorithm 4 Our doubly-robust (DR) approach for counterfactual predictions under runtime confounding with cross fitting

```
Randomly divide training data into three partitions $\mathcal{W}^1$, $\mathcal{W}^2$, $\mathcal{W}^3$.
for $(p, q, r) \in \{(1, 2, 3), (3, 1, 2), (2, 3, 1)\}$ do
    *Stage 1:* On $\mathcal{W}^p$, learn $\hat{\mu}^p(v, z)$ by regressing $Y \sim V, Z \mid T = t$.
               On $\mathcal{W}^q$, learn $\hat{\pi}^q(v, z)$ by regressing $\mathbb{I}\{T = t\} \sim V, Z$
    *Stage 2:* On $\mathcal{W}^r$, learn $\hat{\nu}^r_{\text{DR}}$ by regressing $\left(\frac{\mathbb{I}\{T=t\}}{\hat{\pi}^q(V,Z)}(Y - \hat{\mu}^p(V, Z)) + \hat{\mu}^p(V, Z)\right) \sim V$
DR prediction: $\hat{\nu}_{\text{DR}}(v) = \frac{1}{3}\sum_{i=1}^{3} \hat{\nu}^i_{\text{DR}}(v)$
```

2.5 Theoretical Results for Counterfactual Prediction Methods

Theorem 2.5.1. Under sample-splitting to learn $\hat{\mu}$, $\hat{\pi}$, and $\hat{\nu}_{\text{DR}}$ and stability conditions on the 2nd stage estimators, the DR method has pointwise error bounded by:

$$
\begin{aligned}
\mathbb{E}\Big[\Big(\hat{\nu}_{\text{DR}}(v) - \nu(v)\Big)^2\Big] \lesssim \mathbb{E}\Big[\Big(\tilde{\nu}(v) - \nu(v)\Big)^2\Big] \\
+ \mathbb{E}\Big[(\hat{\pi}(V, Z) - \pi(V, Z))^2 \mid V = v\Big]\mathbb{E}\Big[(\hat{\mu}(V, Z) - \mu(V, Z))^2 \mid V = v\Big]
\end{aligned}
$$

The DR error is bounded by the error of an oracle with access to Y^t and a *product* of nuisance function errors. This product can be substantially smaller than the error of $\hat{\mu}$ in the PL bound. When this product is less than the oracle error, the DR approach is oracle-efficient, in the sense that it achieves (up to a constant factor) the same error rate as an oracle.

Proof. We begin with additional notation needed for the proof. For brevity let $W = (V, Z, A, Y)$ indicate a training observation. The theoretical guarantees for our methods rely on a two-stage training procedure that assumes independent training samples. We denote the first-stage training dataset as $\mathcal{W}^1 := \{W^1_1, W^1_2, W^1_3, ...W^1_n\}$ and the second-stage training dataset as $\mathcal{W}^2 := \{W^2_1, W^2_2, W^2_3, ...W^2_n\}$.

The first step is to derive the form of the error function for our DR approach. For clarity and brevity, we denote the measure of the expectation in the subscript.

$$\begin{aligned}
\hat{r}_{\mathrm{DR}}(v) &= \mathbb{E}_{W|V=v,\mathcal{W}^1}\left[\frac{\mathbb{I}\{T=t\}}{\hat{\pi}(v,Z)}(Y-\hat{\mu}(v,Z))+\hat{\mu}(v,Z)\right]-\nu(v)\\
&= \mathbb{E}_{Z,T|V=v,\mathcal{W}^1}\left[\mathbb{E}_{W|T=t,V=v,Z=z,\mathcal{W}^1}\left[\frac{\mathbb{I}\{T=t\}}{\hat{\pi}(v,Z)}(Y-\hat{\mu}(v,z))+\hat{\mu}(v,z)\right]\right]-\nu(v)\\
&= \mathbb{E}_{Z,T|V=v,\mathcal{W}^1}\left[\mathbb{E}_{Y|T=t,V=v,Z=z,\mathcal{W}^1}\left[\frac{\mathbb{I}\{T=t\}}{\hat{\pi}(v,Z)}(Y-\hat{\mu}(v,z))\right]+\hat{\mu}(v,Z)\right]-\nu(v)\\
&= \mathbb{E}_{Z,T|V=v,\mathcal{W}^1}\left[\frac{\mathbb{I}\{T=t\}}{\hat{\pi}(v,Z)}(\mathbb{E}_{Y|T=t,V=v,Z=z,\mathcal{W}^1}[Y]-\hat{\mu}(v,Z))+\hat{\mu}(v,Z)\right]-\nu(v)\\
&= \mathbb{E}_{W|V=v,\mathcal{W}^1}\left[\frac{\mathbb{I}\{T=t\}}{\hat{\pi}(v,Z)}(\mu(v,Z)-\hat{\mu}(v,Z))+\hat{\mu}(v,Z)\right]-\nu(v)\\
&= \mathbb{E}_{Z|V=v,,\mathcal{W}^1}\left[\mathbb{E}_{W|V=v,Z=z,\mathcal{W}^1}\left[\frac{\mathbb{I}\{T=t\}}{\hat{\pi}(v,Z)}(\mu(v,z)-\hat{\mu}(v,z))+\hat{\mu}(v,z)\right]\right]-\nu(v)\\
&= \mathbb{E}_{Z|V=v,\mathcal{W}^1}\left[\frac{P(A=a\mid V=v,Z=z)}{\hat{\pi}(v,Z)}(\mu(v,Z)-\hat{\mu}(v,Z))+\hat{\mu}(v,Z)\right]-\nu(v)\\
&= \mathbb{E}_{Z|V=v,\mathcal{W}^1}\left[\frac{\pi(v,Z)}{\hat{\pi}(v,Z)}(\mu(v,Z)-\hat{\mu}(v,Z))+\hat{\mu}(v,Z)\right]-\nu(v)\\
&= \mathbb{E}_{Z|V=v,\mathcal{W}^1}\left[\frac{\pi(v,Z)}{\hat{\pi}(v,Z)}(\mu(v,Z)-\hat{\mu}(v,Z))+\hat{\mu}(v,Z)-\mu(v,Z)\right]\\
&= \mathbb{E}\left[\frac{(\mu(v,Z)-\hat{\mu}(v,Z))(\pi(v,Z)-\hat{\pi}(v,Z))}{\hat{\pi}(v,Z)}\mid V=v,\mathcal{W}^1\right].
\end{aligned}$$

Where the first line holds by definition of the error function $\hat{r}$ and the second line by iterated expectation. The third line uses the fact that conditional on $Z=z, V=v, T=t$, then the only randomness in W is Y (and therefore $\hat{\mu}$ is constant). The fourth line makes use of the ($\mathbb{I}\{T=t\}$) term to allow us to condition on only $T=t$ (since the term conditioning on any other $t' \neq t$ will evaluate to zero). The fifth line applies the definition of μ. The sixth line again uses iterated expectation and the seventh makes use of the fact that conditional on Z, the only randomness now is in T and that $\mathcal{W}^1$ is an independent randomly sampled set. The seventh line applies the definition of $\pi(v,z)=\mathbb{P}(T=1\mid V=v,Z=z)$ which since $T\in\{0,1\}$ is equal to $\mathbb{E}[T\mid V=v,Z=z]$. The eight line uses iterated expectation and the fact that $\mathcal{W}^1$ is an independent randomly sampled set to rewrite $\nu(v)=E_{Z|V=v,\mathcal{W}^1}[\mu(v,Z)]$. The ninth line rearranges the terms.

By Cauchy-Schwarz and the positivity assumption,

$$\hat{r}_{\mathrm{DR}}(v)\leq C\sqrt{\mathbb{E}[(\mu(v,Z)-\hat{\mu}(v,Z))^2\mid V=v,\mathcal{W}^1]}\sqrt{\mathbb{E}[(\pi(v,Z)-\hat{\pi}(v,Z))^2\mid V=v,\mathcal{W}^1]}$$

for a constant C.

Squaring both sides yields

$$\hat{r}^2_{\mathrm{DR}}(v)\leq C^2\;\mathbb{E}[(\mu(v,Z)-\hat{\mu}(v,Z))^2\mid V=v,\mathcal{W}^1]\;\mathbb{E}[(\pi(v,Z)-\hat{\pi}(v,Z))^2\mid V=v,\mathcal{W}^1]$$

If $\hat{\pi}$ and $\hat{\mu}$ are estimated using separate training samples, then taking the expectation over the first-stage training sample $\mathcal{W}^1$ yields:

$$\mathbb{E}[\hat{r}^2_{\mathrm{DR}}(v)] \leq C^2\ \mathbb{E}[(\mu(v,Z)-\hat{\mu}(v,Z))^2] \mid V=v]\ \mathbb{E}[(\pi(v,Z)-\hat{\pi}(v,Z))^2] \mid V=v]$$

Applying Theorem 1 of Kennedy (2020) gets the pointwise bound:

$$\mathbb{E}\left[\left(\hat{\nu}_{\mathrm{DR}}(v)-\nu(v)\right)^2\right] \lesssim \mathbb{E}\left[\left(\tilde{\nu}(v)-\nu(v)\right)^2\right] \\ +\mathbb{E}\Big[(\hat{\pi}(V,Z)-\pi(V,Z))^2 \mid V=v\Big]\mathbb{E}\Big[(\hat{\mu}(V,Z)-\mu(V,Z))^2 \mid V=v\Big]$$

□

This model-free result provides bounds that hold for *any* regression method. It is nonetheless instructive to consider the form of these bounds in a couple common contexts. The next result is specialized to the sparse high-dimensional setting, and subsequently we consider the smooth non-parametric setting.

Corollary 2.5.1. Assume stability conditions on the 2nd stage regression estimator and that a k-sparse model can be estimated with squared error $k^2\sqrt{\frac{\log d}{n}}$ (e.g. Chatterjee (2013)).[4] With k_ω-sparse ω, the pointwise error for the TCR method is

$$\mathbb{E}\left[\left(\hat{\nu}_{\mathrm{TCR}}(v)-\nu(v)\right)^2\right] \lesssim\ k^2_\omega\sqrt{\frac{\log d_{\mathrm{V}}}{n}}+b(v)^2$$

With k_μ-sparse μ and k_ν-sparse ν, the pointwise error for the PL method is

$$\mathbb{E}\left[\left(\hat{\nu}_{\mathrm{PL}}(v)-\nu(v)\right)^2\right] \lesssim\ k^2_\nu\sqrt{\frac{\log d_{\mathrm{V}}}{n}}+k^2_\mu\sqrt{\frac{\log d}{n}}$$

Additionally with k_π-sparse π, the pointwise error for the DR method is

$$\mathbb{E}\left[\left(\hat{\nu}_{\mathrm{DR}}(v)-\nu(v)\right)^2\right] \lesssim\ k^2_\nu\sqrt{\frac{\log d_{\mathrm{V}}}{n}}+k^2_\mu k^2_\pi\frac{\log d}{n}$$

The DR approach is therefore oracle efficient when $\left(\frac{k_\mu k_\pi}{k_\nu}\right)^2 \lesssim \left(\frac{\sqrt{n\log d_{\mathrm{V}}}}{\log d}\right)$.

Based on the upper bound, we cannot claim efficiency for the PL approach because $k_\mu > k_\nu$ and $d > d_{\mathrm{V}}$. For exposition, consider the simple case where $k_\nu \approx k_\mu \approx k_\pi$. Corollary 2.5.1 indicates that when $d_{\mathrm{V}} \approx d$, the DR and PL methods will perform similarly. When $d_{\mathrm{V}} \ll d$, we expect the DR to outperform the PL method because the second term of the PL bound dominates the error whereas the first term of the DR bound dominates in high-dimensional settings. When $d_{\mathrm{V}} \ll d$ and the amount of confounding is small, we expect the TCR to perform well.

[4] We use the sparsity parameter k to indicate k covariates have non-zero coefficients in the model.

Corollary 2.5.2. Assume stability conditions on the 2nd stage regression estimator and that a β-smooth function of a p-dimensional vector can be estimated with squared error $n^{\frac{-2\beta}{2\beta+p}}$. With β_ω-smooth ω, the pointwise error for the TCR method is

$$\mathbb{E}\Big[\Big(\hat{\nu}_{\mathrm{TCR}}(v) - \nu(v)\Big)^2\Big] \lesssim \; n^{-2\beta_\omega/(2\beta_\omega + d_{\mathrm{V}})} + b(v)^2$$

With β_μ-smooth μ and β_ν-smooth ν, the pointwise error for the PL method is

$$\mathbb{E}\Big[\Big(\hat{\nu}_{\mathrm{PL}}(v) - \nu(v)\Big)^2\Big] \lesssim \; n^{-2\beta_\nu/(2\beta_\nu + d_{\mathrm{V}})} + n^{-2\beta_\mu/(2\beta_\mu + d)}$$

Additionally with β_π-smooth π, the pointwise error for the DR method is

$$\mathbb{E}\Big[\Big(\hat{\nu}_{\mathrm{DR}}(v) - \nu(v)\Big)^2\Big] \lesssim \; n^{-2\beta_\nu/(2\beta_\nu + d_{\mathrm{V}})} + n^{\frac{-2\beta_\mu}{2\beta_\mu+d} + \frac{-2\beta_\pi}{2\beta_\pi+d}}$$

The DR approach is therefore oracle efficient when $\frac{\beta_\nu}{\beta_\nu + d_{\mathrm{V}}/2} \leq \frac{\beta_\mu}{\beta_\mu + d/2} + \frac{\beta_\pi}{\beta_\pi + d/2}$ which simplifies to $s \geq \frac{d/2}{1 + \frac{d_{\mathrm{V}}}{\beta_\nu}}$ when $\beta_\pi = \beta_\mu = s$.

As in the sparse setting above, we cannot claim oracle efficiency for PL approach based on this upper bound because $\beta_\mu \leq \beta_\nu$ and $d > d_{\mathrm{V}}$. For exposition, consider an example where $\beta_\nu \approx \beta_\mu \approx \beta_\pi$. Corollary 2.5.2 indicates that when $d_{\mathrm{V}} \approx d$, the DR and PL methods will perform similarly. When $d_{\mathrm{V}} \ll d$, we expect the DR to outperform the PL method because the second term of the PL bound dominates the error whereas the first term of the DR bound dominates. When $d_{\mathrm{V}} \ll d$ and the amount of confounding is small, we expect the TCR to perform well.

This theoretical analysis helps us understand when we expect the prediction methods to perform well. However, in practice, these upper bounds may not be tight and the degree of confounding is typically unknown. To compare the prediction methods in practice, we require a method for counterfactual model evaluation.

2.6 Methodology for Valid Evaluations of Counterfactual Prediction Models using Observational Data

We describe an approach for evaluating the prediction methods using observed data. In our problem setting (§ 2.2.1), the mean-squared prediction error of a model $\hat{\nu}$ is identified as $\mathbb{E}[(Y^t - \hat{\nu}(V))^2] = \mathbb{E}[\mathbb{E}[(Y - \hat{\nu}(V))^2 \mid V, Z, T = t]]$. We propose a doubly-robust procedure to estimate the prediction error that follows the approach in Chapter 1, which focused on classification metrics. Defining the error regression $\eta(v, z) := \mathbb{E}[(Y^t - \hat{\nu}(V))^2 | V = v, Z = z]$, which is identified as $\mathbb{E}[(Y - \hat{\nu}(V))^2 \mid V = v, Z = z, T = t]$, the **doubly-robust estimate of the MSE of** ν is

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{I}\{T_i = t\}}{\hat{\pi}(V_i, Z_i)}\Big(\Big(Y_i - \hat{\nu}(V_i)\Big)^2 - \hat{\eta}(V_i, Z_i)\Big) + \hat{\eta}(V_i, Z_i)\right]$$

The doubly-robust estimation of MSE is $\sqrt{n}$-consistent under sample-splitting and $n^{1/4}$ convergence in the nuisance function error terms, enabling us to get estimates with confidence intervals. Algorithm 5 describes this procedure. This evaluation method can also be used to select the regression estimators for the first and second stages.

Algorithm 5 Cross-fitting approach to evaluation of counterfactual prediction methods

```
**Input:** Test samples $\{(V_j, Z_j, T_j, Y_j)\}_{j=1}^{2n}$ and prediction models $\{\hat{\nu}_1, ...\hat{\nu}_h\}$
Randomly divide test data into two partitions $\mathcal{W}^0 = \{(V_j^0, Z_j^0, T_j^0, Y_j^0)\}_{j=1}^n$ and $\mathcal{W}^1 = \{(V_j^1, Z_j^1, T_j^1, Y_j^1)\}_{j=1}^n$.
for $(p, q) \in \{(0,1), (1,0)\}$ do
    On $\mathcal{W}^p$, learn $\hat{\pi}^p(v, z)$ by regressing $\mathbb{I}\{T = t\} \sim V, Z$.
    for $m \in \{1, ...., h\}$ do
        On $\mathcal{W}^p$, learn $\hat{\eta}_m^p(v, z)$ by regressing $(Y - \hat{\nu}_m(V))^2 \sim V, Z \mid T = t$
        On $\mathcal{W}^q$, for $j \in \{1, ..., n\}$ compute $\phi_{m,j}^q = \frac{\mathbb{I}\{T_j^q = t\}}{\hat{\pi}^p(V_j^q, Z_j^q)}((Y_j^q - \hat{\nu}_m(V_j^q))^2 - \hat{\eta}_m^p(V_j^q, Z_j^q)) + \hat{\eta}_m^p(V_j^q, Z_j^q)$
**Output error estimate confidence intervals:** for $m \in \{1, ..., h\}$:
$\mathrm{MSE}_m = \left(\frac{1}{2n}\sum_{i=0}^{1}\sum_{j=1}^{n}\phi_{m,j}^i\right) \pm 1.96\sqrt{\frac{1}{2n}\mathrm{var}(\phi_m)}$
```

2.7 Empirical Results on Synthetic Data

We evaluate our methods against ground truth by performing empirical analysis on simulated data, where we can vary the amount of confounding in order to assess the effect on predictive performance. While our theoretical results for PL and DR are obtained under sample splitting, in practice there may be a reluctance to perform sample splitting in training predictive models due to the potential loss in efficiency. We present results where we use the full training data to learn the 1st-stage nuisance functions and 2nd-stage regressions for DR and PL and we use the full training data for the one-stage TCR.[5] This allows us to examine performance in a setting outside what our theory covers.

We first analyze how the methods perform in a sparse linear model. This simple setup enables us to explore how properties like correlation between V and Z impact performance. We simulate data as

$$V_i \sim \mathcal{N}(0,1) \qquad ; \ 1 \le i \le d_{\mathrm{V}}$$

$$Z_i \sim \mathcal{N}(\rho V_i, 1 - \rho^2) \qquad ; \ 1 \le i \le d_{\mathrm{Z}}$$

$$\mu(V, Z) = \frac{k_v}{k_v + \rho k_z}\left(\sum_{i=1}^{k_v} V_i + \sum_{i=1}^{k_z} Z_i\right) \qquad Y^t = \mu(V, Z) + \epsilon \ ; \ \epsilon \sim \mathcal{N}\left(0, \frac{1}{2n}\|\mu(V, Z)\|_2^2\right)$$

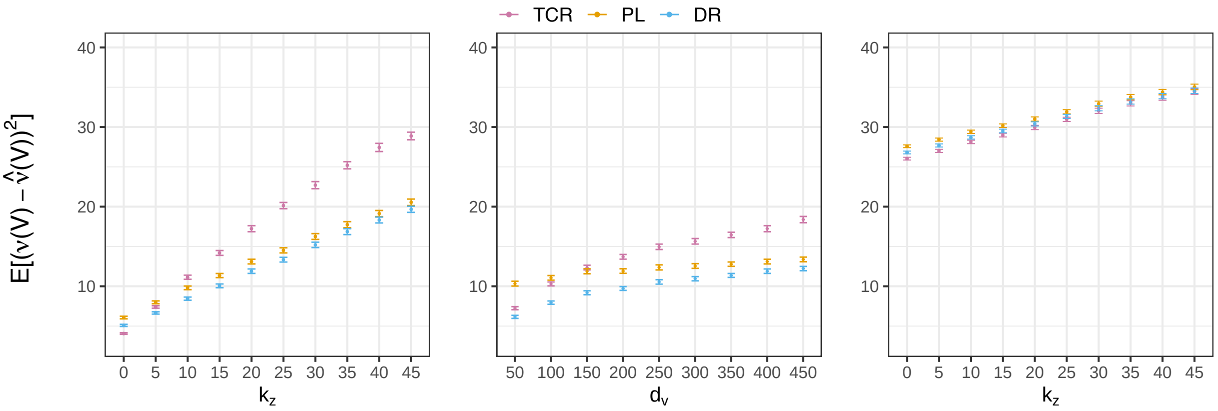
$$\nu(V) = \frac{k_v}{k_v + \rho k_z}\left(\sum_{i=1}^{k_v} V_i + \rho \sum_{i=1}^{k_z} V_i\right)$$

$$\pi(V, Z) = 1 - \sigma\left(\frac{1}{\sqrt{k_v + k_z}}\left(\sum_{i=1}^{k_v} V_i + \sum_{i=1}^{k_z} Z_i\right)\right) \qquad T \sim \mathrm{Bernoulli}(\pi(V, Z))$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$. We normalize $\pi(v, z)$ by $\frac{1}{\sqrt{k_v + k_z}}$ to satisfy Condition 2.2.1.4 and use the coefficient $k_v/(k_v + \rho k_z)$ to facilitate a fair comparison as we vary ρ. For all empirical results, we report test MSE for 300 simulations where each simulation generates $n = 2000$ data points split randomly and evenly into train and test sets.[6] In the first set of experiments, for fixed $d = d_{\mathrm{V}} + d_{\mathrm{Z}} = 500$, we vary d_{V} (and correspondingly d_{Z}). We also vary k_z, which governs the runtime confounding. Larger values of k_z correspond to more confounding variables. The theoretical analysis (§ 2.4) suggests that when confounding (k_z) is small, then the TCR and DR methods will perform well. More confounding (larger k_z) should increase error for all methods, and we expect this increase to be significantly larger for the TCR method that has confounding bias. We expect the TCR and DR methods to perform better at smaller values

[5] We report error metrics on a random heldout test set.

[6] Source code is available at https://github.com/mandycoston/confound_sim



(a) LASSO

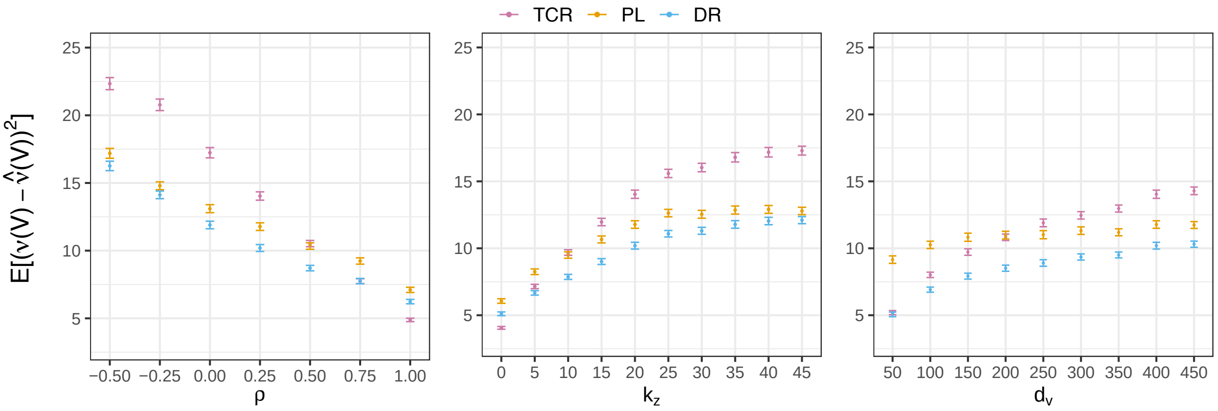
(b) LASSO

(c) Random forests

Figure 2.1: Results for several approaches (each a different color) to learning predictive models under runtime confounding on synthetic data where the runtime predictors are uncorrelated with the runtime confounders. **(a)** MSE as we vary k_z using cross-validated LASSO to learn $\hat{\pi}$, $\hat{\mu}$, $\hat{\nu}_{\text{TCR}}$, $\hat{\nu}_{\text{PL}}$, $\hat{\nu}_{\text{DR}}$ for $\rho = 0$, $d_{\text{V}} = 400$ and $k_v = 25$. At low levels of confounding (k_z), the TCR method does well but performance degrades with k_z. For any non-zero confounding, our DR method performs best.
(b) MSE against d_{V} using cross-validated LASSO and $\rho = 0$, $k_v = 25$ and $k_z = 20$. The DR method performs the best across the range of d_{V}. When d_{V} is small, the TCR method also does well since its estimation error is small. The PL method has higher error since it suffers from the full d-dimensional estimating error in the first stage.
(c) MSE as we vary k_z using random forests and $\rho = 0$, $d_{\text{V}} = 400$ and $k_v = 25$. Compared to LASSO in (a), there is a relatively small increase in error as we increase k_z, suggesting that estimation error dominates the confounding error. The TCR method performs best at lower levels of confounding and on par with the DR method for larger values of k_z.
Error bars denote 95% confidence intervals.

of d_{V}; by contrast, we expect the PL performance to vary less with d_{V} since the PL method suffers from the full d-dimensionality in the first stage regardless of d_{V}. For large values of d_{V}, we expect the PL method to perform similarly to the DR method. Fig. 2.1 plots the MSE in estimating ν for $\rho = 0$ and $k_v = 25$ using LASSO and random forests. The LASSO plots in Fig. 2.1a and 2.1b show the expected trends. Random forests have much higher error than the LASSO (compare Fig. 2.1a to 2.1c) and we only see a small increase in error as we increase confounding (Fig. 2.1c) because the random forest estimation error dominates the confounding error. In this setting, the TCR method may outperform the other methods, and in fact the TCR performs best at low levels of confounding.

We next consider the case were V and Z are correlated. If V and Z are perfectly correlated, there is no confounding. For our data where higher values of V and Z both decrease π and increase μ, a positive correlation should reduce confounding, and a negative correlation may exacerbate confounding by increasing the probability that Z is small given $T = t$ and V is large and therefore increasing the gap $\mathbb{E}[Y^t \mid V = v] - \mathbb{E}[Y^t \mid V = v, T = t]$. Fig. 2.2 gives MSE for correlated V and Z. As expected, error overall decreases with ρ (Fig. 2.2a). Relative to the uncorrelated setting (Fig. 2.1), the weak positive correlation reduces MSE for all methods, particularly for large k_z and d_{V}. The DR method achieves the lowest error for settings with confounding, performing on par with the TCR when $d_{\text{V}} = 50$.



(a) LASSO

(b) LASSO

(c) LASSO

Figure 2.2: Results for several approaches (each a different color) to learning predictive models under runtime confounding on synthetic data where the runtime predictors are correlated with the runtime confounders. **(a)** MSE against correlation ρ_{V_i,Z_i} for $k_z = 20$, $k_v = 25$, and $d_{\mathrm{V}} = 400$. Error decreases with ρ for all methods. Our DR method achieves the lowest error under confounding ($\rho < 1$). **(b)** MSE as we increase k_z for $\rho = 0.25$, $k_v = 25$, and $d_{\mathrm{V}} = 400$. Compare to Figure 2.1a; the weak positive correlation reduces MSE, particularly for $k_v < i \leq k_z$ when V_i is only a correlate for the confounder Z_i but not a confounder itself. **(c)** MSE against d_{V} for $\rho = 0.25$, $k_z = 20$, and $k_v = 25$. The DR method is among the best-performing for all d_{V}. As with the uncorrelated setting (2.1b), the DR and TCR methods are better able to take advantage of low d_{V} than the PL method.
Error bars denote 95% confidence intervals.

Empirical Results under Second-Stage Misspecification Next, we explore a more complex data generating process through the lens of model interpretability. Interpretability requirements allow for a complex training process as long as the final model outputs interpretable predictions (Tan et al., 2018; Zeng et al., 2017; Rudin, 2019b). Since the PL and DR first stage regressions are only a part of the training process, we can use any flexible model to learn the first stage functions as accurately as possible without impacting interpretability. Constraining the second-stage learning class to interpretable models (e.g. linear classifiers) may cause misspecification since the interpretable class may not contain the true model. We simulate such a setting by modifying the setup (for $\rho = 0$):

$$V_i \sim \mathcal{N}(0,1) \;\; \text{for} \;\; 1 \leq i \leq \frac{d_{\mathrm{V}}}{2} \quad ; \quad V_i := V_j^2 \;\; \text{for} \;\; \frac{d_{\mathrm{V}}}{2} < i \leq d_{\mathrm{V}}, \quad j = i - \frac{d_{\mathrm{V}}}{2}$$

$$\mu(V,Z) = \sum_{i=1}^{k_v/2} \left(V_i + (2(i \bmod 2) - 1)V_i^2 \right) + \sum_{i=1}^{k_z} Z_i \;\; ; \; \nu(V) = \sum_{i=1}^{k_v/2} \left(V_i + (2(i \bmod 2) - 1)V_i^2 \right)$$

We restrict our second stage models and the TCR model to predictors V_i for $1 \leq i \leq \frac{d_{\mathrm{V}}}{2}$ to simulate a real-world setting where we are constrained to linear classifiers using only V at runtime. We allow the first stage models access to the full V and Z since the first stage is not constrained by variables or model class. We use cross-validated LASSO models for both stages and compare this setup to the setting where the model is correctly specified. The DR method achieves the lowest error for both settings (Table 2.1), although the error is significantly higher for all methods under misspecification.

Method	Correct specification	2nd-stage misspecification
TCR	16.64 (16.28, 17.00)	35.52 (35.18, 35.85)
PL	12.32 (12.03, 12.61)	32.09 (31.82, 32.36)
DR (ours)	**11.10 (10.84, 11.37)**	**31.33 (31.06, 31.59)**

Table 2.1: Synthetic results for prediction error of several learning methods under runtime confounding. This table provides the MSE $\mathbb{E}[\big(\nu(V) - \hat{\nu}(V)\big)^2]$ under correct specification and misspecification in the 2nd stage for $d = 500$, $d_{\mathrm{V}} = 400$, $k_v = 24$, $k_z = 20$ and $n = 3000$ (with 95% confidence intervals). Our DR method has the lowest error in both settings. Errors are larger for all methods under misspecification.

2.8 Empirical Results on Real-world Child Welfare Data

In this section we present empirical results on the child welfare screening task introduced in the previous chapter. In agencies that have adopted risk assessment tools, the worker relies on (immediate risk) information communicated during the call and an algorithmic risk score that summarizes (longer term) risk based on historical administrative data (Chouldechova et al., 2018). The call is recorded but is not used as a predictor for three reasons: (1) the inadequacy of existing case management software to run speech/NLP models on calls in realtime; (2) model interpretability requirements; and (3) the need to maintain distinction between immediate risk (as may be conveyed during the call) and longer-term risk the model seeks to estimate. Since it is not possible to use call information as a predictor, we encounter runtime confounding. Additionally, we would like to account for the disproportionate involvement of families of color in the child welfare system (Dettlaff et al., 2011), but due to its sensitivity, we do not want to use race in the prediction model.

The task is to predict which cases are likely to be offered services under the decision $t =$ "screened in for investigation" using historical administrative data as predictors (V) and accounting for confounders race and allegations in the call (Z). Our dataset consists of over 30,000 calls to the hotline in Allegheny County, PA. We use random forests in the first stage for flexibility and LASSO in the second stage for interpretability. Table 2.2 presents the MSE using our evaluation method (§ 2.6).[7] The PL and DR methods achieve a statistically significant lower MSE than the TCR approach, suggesting these approaches could help workers better identify at-risk children than standard practice.

	MSE
TCR	0.290 (0.287, 0.293)
PL	**0.249 (0.246, 0.251)**
DR (ours)	**0.248 (0.245, 0.250)**

Table 2.2: Child welfare screening results for prediction error of several learning methods under runtime confounding. This table shows the MSE estimated via our evaluation procedure (§ 2.6). The PL and DR approaches achieve lower MSE than the TCR approach. 95% confidence intervals given in parentheses.

[7] We report error metrics on a random held-out test set.

2.9 Conclusion

This chapters presents a generic procedure for learning counterfactual predictions under runtime confounding that can be used with any parametric or nonparametric learning algorithm. Our theoretical and empirical analysis suggests this procedure will often outperform other methods, particularly when the level of runtime confounding is significant. Our method is backed by a doubly-robust guarantee on the mean-squared error (MSE) that implies oracle efficiency when the product of nuisance function errors is less than the MSE of an oracle with access to the potential outcomes.

In this chapter we focused largely on issues of validity – how to obtain valid counterfactual predictions and evaluations under runtime confounding. Next, we delve into our second principle for responsible use, equity, as we analyze the impact of algorithms on historically marginalized groups. During this discussion, we will see how threats to validity, such as selection bias, impact equity.

3

Characterizing Fairness Properties over the Set of Good Models

Algorithms used in high-stakes settings can disproportionately harm marginalized groups (Barocas and Selbst, 2016a; Dastin, 2018; Vigdor, 2019). The vast literature on algorithmic fairness offers numerous methods for learning anew the best performing model among those that satisfy a chosen notion of predictive fairness (e.g. Zemel et al. (2013), Agarwal et al. (2018), Agarwal et al. (2019)). However, for real-world settings where a risk assessment is already in use, practitioners and auditors may instead want to assess disparities with respect to the current model, which we term the *benchmark model*. In this chapter, we propose a method to answer the question: Can we improve upon the benchmark model in terms of predictive fairness with minimal change in overall accuracy? To answer this question, this chapter provides methods that were first published in Coston et al. (2021b).

We explore this question through the lens of the "Rashomon Effect," a common empirical phenomenon whereby multiple models perform similarly overall but differ markedly in their predictions for individual cases (Breiman, 2001). These models may perform quite differently over various groups, and therefore have different predictive fairness properties. We propose an algorithm, Fairness in the Rashomon Set (FaiRS), to characterize predictive fairness properties over the set of models that perform similarly to a chosen benchmark model. We refer to this set as *the set of good models* (Dong and Rudin, 2020). FaiRS is designed to efficiently answer the following questions: What are the range of predictive disparities that could be generated over the set of good models? What is the disparity minimizing model within the set of good models?

A key empirical challenge to validity in domains such as credit lending is that outcomes are not observed for all cases (Lakkaraju et al., 2017; Kleinberg et al., 2018). This *selective labels problem* is particularly vexing in the context of assessing predictive fairness. Our framework addresses selectively labelled data in contexts where the selection decision and outcome are unconfounded given the observed data features.

Our methods are useful for legal audits of disparate impact. In various domains, decisions that generate disparate impact must be justified by "business necessity" (CRA, 1964; ECOA, 1974; Barocas and Selbst, 2016a). For instance, financial regulators investigate whether credit lenders could have offered more loans to minority applicants without affecting default rates (Gillis, 2020). Employment regulators may investigate whether resume screening software screens out underrepresented applicants for reasons that cannot be attributed to the job criteria (Raghavan et al., 2020b). Our methods provide one possible formalization of the business necessity criteria. An auditor can use FaiRS to assess whether there exists an alternative model that reduces predictive disparities without compromising performance relative to the benchmark model. If possible, then it is difficult to justify the benchmark model on the grounds of business necessity.

Our methods can also be a useful tool for decision makers who want to improve upon an existing model. A decision maker may use FaiRS to search for a prediction function that reduces predictive disparities without compromising performance relative to the benchmark model. We emphasize that the effective usage of our methods requires careful thought about the broader social context surrounding the setting of interest (Selbst et al., 2019; Holstein et al., 2019a).

In this chapter, we develop an algorithmic framework, Fairness in the Rashomon Set (FaiRS), to investigate predictive disparities over the set of good models. We provide theoretical guarantees on the generalization error and predictive disparities of FaiRS [§ 3.3]. Next we propose a variant of FaiRS that addresses the selective labels problem and achieves the same guarantees under oracle access to the outcome regression function [§ 3.5]. We then use FaiRS to audit the COMPAS risk assessment, finding that it generates larger predictive disparities between black and white defendants than any model in the set of good models [§ 3.7]. Finally we use FaiRS on a selectively labelled credit-scoring dataset to build a model with lower predictive disparities than a benchmark model [§ 3.8].

3.1 Background and Related Work

3.1.1 Rashomon Effect

In a seminal paper on statistical modeling, Breiman (2001) observed that often a multiplicity of good models achieve similar accuracy by relying on different features, which he termed the "Rashomon effect." Even though they have similar accuracy, these models may differ along other key dimensions, and recent work considers the implications of the Rashomon effect for model simplicity, interpretability, and explainability (Fisher et al., 2019; Marx et al., 2019; Rudin, 2019a; Dong and Rudin, 2020; Semenova et al., 2020).

We introduce these ideas into research on algorithmic fairness by studying the range of predictive disparities that can be achieved over the set of good models. We provide computational techniques to directly and efficiently investigate the range of predictive disparities that may be generated over the set of good models. Our recidivism risk prediction and credit scoring applications demonstrate that the set of good models is a rich empirical object, and we illustrate how characterizing the range of achievable predictive fairness properties over this set can be used for model learning and evaluation.

3.1.2 Fair Classification and Fair Regression

An influential literature on fair classification and fair regression constructs prediction functions that minimize loss subject to a predictive fairness constraint chosen by the decision maker (Dwork et al., 2012b; Zemel et al., 2013; Hardt et al., 2016b; Menon and Williamson, 2018; Donini et al., 2018; Agarwal et al., 2018, 2019; Zafar et al., 2019). In contrast, we construct prediction functions that minimize a chosen measure of predictive disparities subject to a constraint on overall performance. This is useful when decision makers find it difficult to specify acceptable levels of predictive disparities, but instead know what performance loss is tolerable. It may be unclear, for instance, how a lending institution should specify acceptable differences in credit risk scores across groups, but the lending institution can easily specify an acceptable average default rate among approved loans. Our methods allow users to directly search for prediction functions that reduce disparities given such a specified loss tolerance. Similar in spirit to our work, Zafar et al. (2019) provide a method for selecting a classifier that minimizes a particular notion of predictive fairness, "decision boundary covariance," subject to a performance constraint. Our method applies more generally to a large class of predictive disparities and covers both classification and regression tasks.

While originally developed to solve fair classification and fair regression problems, we show that the "reductions approach" used in Agarwal et al. (2018, 2019) can be suitably adapted to solve general optimization problems over the set of good models. This provides a general computational approach that may be useful for investigating the implications of the Rashomon Effect for other model properties.

In constructing the set of good models with comparable performance to a benchmark model, our work bears resemblance to techniques that "post-process" existing models. Post-processing techniques typically modify the predictions from an existing model to achieve a target notion of fairness (Hardt et al., 2016b; Pleiss et al., 2017; Kim et al., 2019). By contrast, our methods only use the existing model to calibrate the performance constraint, but need not share any other properties with the benchmark model. While post-processing techniques often require access to individual predictions from the benchmark model, our approach only requires that we know its average loss.

3.1.3 Selective Labels and Missing Data

In settings such as criminal justice and credit lending, the training data only contain labeled outcomes for a selectively observed sample from the full population of interest. For example, banks use risk scores to assess all loan applicants, yet the historical data only contains default/repayment outcomes for those applicants whose loans were approved. This is a missing data problem (Little and Rubin, 2019). Because the outcome label is missing based on a selection mechanism, this type of missing data is known as the *selective labels problem* (Lakkaraju et al., 2017; Kleinberg et al., 2018). One solution treats the selectively labelled population as if it were the population of interest, and proceeds with training and evaluation on the selectively labelled population only. This is also called the "*known good-bad*" (KGB) approach (Zeng and Zhao, 2014; Nguyen et al., 2016). The problem with such an approach, as we saw in Chapter 1, is that evaluating a model on a population different than the one on which it will be used can produce invalid assessments, particularly with regards to predictive fairness measures. Unfortunately, most fair classification and fair regression methods do not offer modifications to address the selective labels problem. Our framework does [§ 3.5].

Popular in credit lending applications, "reject inference" procedures incorporate information from the selectively unobserved cases (i.e., rejected applicants) in model construction and evaluation by imputing missing outcomes using augmentation, reweighing or extrapolation-based approaches (Li et al., 2020; Mancisidor et al., 2020). These approaches are similar to domain adaptation techniques, and indeed the selective labels problem can be cast as domain adaptation since the labelled training data is not a random sample of the target distribution. Most relevant to our setting are covariate shift methods for domain adaptation. Reweighing procedures have been proposed for jointly addressing covariate shift and fairness (Coston et al., 2019; Singh et al., 2021). While FaiRS similarly uses iterative reweighing to solve our joint optimization problem, we explicitly use extrapolation to address covariate shift. Empirically we find extrapolation can achieve lower disparities than reweighing.

3.2 Problem Formulation and Additional Notation

The training data consist of n i.i.d. draws from the joint distribution $(X_i, A_i, T_i, Y_i^*) \sim P$ and may suffer from a *selective labels problem*: There exists $\mathcal{T}^* \subseteq \mathcal{T}$ such that the outcome is observed if and only if the decision satisfies $T_i \in \mathcal{T}^*$. Hence, the training data are $\{(X_i, A_i, T_i, Y_i)\}_{i=1}^n$, where $Y_i = Y_i^* \mathbb{I}\{T_i \in \mathcal{T}^*\})$ is the *observed outcome*.

Given a specified set of prediction functions $\mathcal{F}$ with elements $f \colon \mathcal{X} \to [0,1]$, we search for the prediction function $f \in \mathcal{F}$ that minimizes or maximizes a measure of predictive disparities with respect to the sensitive attribute subject to a constraint on predictive performance. We measure performance using average loss, where $l \colon \mathcal{Y} \times [0,1] \to [0,1]$ is the loss function and $\text{loss}(f) := \mathbb{E}\left[l(Y_i^*, f(X_i))\right]$. The loss function is assumed to be 1-Lipshitz under the l_1-norm following Agarwal et al. (2019). The constraint on performance takes the form $\text{loss}(f) \leq \epsilon$ for some specified *loss tolerance* $\epsilon \geq 0$. The set of prediction functions satisfying this constraint is the *set of good models*.

The loss tolerance may be chosen based on an existing benchmark model $\tilde{f}$ such as an existing risk score, e.g., by setting $\epsilon = (1+\delta)\,\text{loss}(\tilde{f})$ for some $\delta \in [0,1]$. The set of good models now describes the set of models whose performance lies within a δ-neighborhood of the benchmark model. When defined in this manner, the set of good models is also called the "Rashomon set" (Rudin, 2019a; Fisher et al., 2019; Dong and Rudin, 2020; Semenova et al., 2020).

3.2.1 Measures of Predictive Disparities

We consider measures of predictive disparity of the form

$$\text{disp}(f) := \beta_0 \mathbb{E}\left[f(X_i) | \mathcal{E}_{i,0}\right] + \beta_1 \mathbb{E}\left[f(X_i) | \mathcal{E}_{i,1}\right], \tag{3.1}$$

where $\mathcal{E}_{i,a}$ is a group-specific conditioning event that depends on (A_i, Y_i^*) and $\beta_a \in \mathbb{R}$ for $a \in \{0,1\}$ are chosen parameters. Note that we measure predictive disparities over the *full* population (i.e., not conditional on T_i).

For different choices of the conditioning events $\mathcal{E}_{i,0}, \mathcal{E}_{i,1}$ and parameters β_0, β_1, our predictive disparity measure summarizes violations of common definitions of predictive fairness.

Definition 3.2.1. Statistical parity (SP) requires the prediction $f(X_i)$ to be independent of the attribute A_i (Dwork et al., 2012b; Zemel et al., 2013; Feldman et al., 2015). By setting $\mathcal{E}_{i,a} = \{A_i = a\}$ for $a \in \{0,1\}$ and $\beta_0 = -1$, $\beta_1 = 1$, $\text{disp}(f)$ measures the difference in average predictions across values of the sensitive attribute.

Definition 3.2.2. Suppose $\mathcal{Y} = \{0,1\}$. **Balance for the positive class** (BFPC) and **balance for the negative class** (BFNC) requires the prediction $f(X_i)$ to be independent of the attribute A_i conditional on $Y_i^* = 1$ and $Y_i^* = 0$ respectively (e.g., Chapter 2 of (Barocas et al., 2019)). Defining $\mathcal{E}_{i,a} = \{Y_i^* = 1, A_i = a\}$ for $a \in \{0,1\}$ and $\beta_0 = -1, \beta_1 = 1$, $\text{disp}(f)$ describes the difference in average predictions across values of the sensitive attribute given $Y_i^* = 1$. If instead $\mathcal{E}_{i,a} = \{Y_i^* = 0, A_i = a\}$ for $a \in \{0,1\}$, then $\text{disp}(f)$ equals the difference in average predictions across values of the sensitive attribute given $Y_i^* = 0$.

Our focus on differences in average predictions across groups is a common relaxation of parity-based predictive fairness definitions (Corbett-Davies et al., 2017; Mitchell et al., 2019b).

Our predictive disparity measure can also be used for *fairness promoting interventions*, which aim to increase opportunities for a particular group. For instance, the decision maker may wish to search for the prediction function among the set of good models that minimizes the average predicted risk score $f(X_i)$ for a historically disadvantaged group.

Definition 3.2.3. Defining $\mathcal{E}_{i,1} = \{A_i = 1\}$ and $\beta_0 = 0, \beta_1 = 1$, $\mathrm{disp}(f)$ measures the average risk score for the group with $A_i = 1$. This is an **affirmative action**-based fairness promoting intervention. Further assuming $\mathcal{Y} = \{0, 1\}$ and defining $\mathcal{E}_{i,1} = \{Y_i^* = 1, A_i = 1\}$, $\mathrm{disp}(f)$ measures the average risk score for the group with both $Y_i^* = 1$, $A_i = 1$. This is a **qualified affirmative action**-based fairness promoting intervention.

3.2.2 Characterizing Predictive Disparities over the Set of Good Models

We develop the algorithmic framework, Fairness in the Rashomon Set (FaiRS), to solve two related problems over the set of good models. First, we characterize the range of predictive disparities by minimizing or maximizing the predictive disparity measure over the set of good models. We focus on the minimization problem

$$\min_{f \in \mathcal{F}} \mathrm{disp}(f) \text{ s.t. } \mathrm{loss}(f) \leq \epsilon. \tag{3.2}$$

Second, we search for the prediction function that minimizes the absolute predictive disparity over the set of good models

$$\min_{f \in \mathcal{F}} |\mathrm{disp}(f)| \text{ s.t. } \mathrm{loss}(f) \leq \epsilon. \tag{3.3}$$

The solutions to these problems tell us whether there exist alternative prediction functions that achieve similar performance yet generate different predictive disparities. The existence of such a model is relevant to both decision makers, who may want to replace an existing model with a more equitable but equally performant one, and to auditors, who may want to know whether "business necessity"-type defenses to disparate impact hold in a given setting (CRA, 1964; ECOA, 1974; Barocas and Selbst, 2016a).

3.3 Methodology for Optimizing over the Set of Good Models

We characterize the range of predictive disparities (3.2) and find the absolute predictive disparity minimizing model (3.3) over the set of good models using techniques inspired by the reductions approach in Agarwal et al. (2018, 2019). Although originally developed to solve fair classification and fair regression problems in the case without selective labels, we extend the reductions approach to solve general optimization problems over the set of good models in the presence of selective labels. For exposition, we present our method for (3.2) in the case without selective labels, where $\mathcal{T}^* = \mathcal{T}$ and the outcome Y_i^* is observed for all observations.

3.3.1 Computing the Range of Predictive Disparities

We consider randomized prediction functions that select $f \in \mathcal{F}$ according to some distribution $Q \in \Delta(\mathcal{F})$ where Δ denotes the probability simplex. Let $\mathrm{loss}(Q) := \sum_{f \in \mathcal{F}} Q(f)\,\mathrm{loss}(f)$ and $\mathrm{disp}(Q) := \sum_{f \in \mathcal{F}} Q(f)\,\mathrm{disp}(f)$. We solve

$$\min_{Q \in \Delta(\mathcal{F})} \mathrm{disp}(Q) \text{ s.t. } \mathrm{loss}(Q) \leq \epsilon. \tag{3.4}$$

While it may be possible to solve this problem directly for certain parametric function classes, we develop an approach that can be applied to any generic function class.[1] A key object for doing so will be classifiers obtained by thresholding prediction functions. For cutoff $z \in [0, 1]$, define $h_f(x, z) = \mathbb{I}\{f(x) \geq z\}$ and let $\mathcal{H} := \{h_f : f \in \mathcal{F}\}$ be the set of all classifiers obtained by thresholding prediction functions $f \in \mathcal{F}$. We first reduce the optimization problem (3.4) to a constrained classification problem through a discretization argument, and then solve the resulting constrained classification problem through a further reduction to finding the saddle point of a min-max problem.

Algorithm 6 Algorithm for finding the predictive disparity minimizing model

```
Input: Training data $\{(X_i, Y_i, A_i)\}_{i=1}^n$, parameters $\beta_0, \beta_1$, events $\mathcal{E}_{i,0}, \mathcal{E}_{i,1}$, empirical loss
       tolerance $\hat{\epsilon}$, bound $B_\lambda$, accuracy $\nu$ and learning rate $\eta$.
Result: $\nu$-approximate saddle point $(\hat{Q}_h, \hat{\lambda})$
Set $\theta_1 = 0 \in \mathbb{R}$   for $t = 1, 2, \ldots$ do
    Set $\lambda_t = B_\lambda \frac{\exp(\theta_t)}{1+\exp(\theta_t)}$;
     $h_t \leftarrow \mathrm{Best}_h(\lambda_t)$;
     $\hat{Q}_{h,t} \leftarrow \frac{1}{t}\sum_{s=1}^t h_s$,      $\bar{L} \leftarrow L(\hat{Q}_{h,t}, \mathrm{Best}_\lambda(\hat{Q}_{h,t}))$;
     $\hat{\lambda}_t \leftarrow \frac{1}{t}\sum_{s=1}^t \lambda_s$,      $\underline{L} \leftarrow L(\mathrm{Best}_h(\hat{\lambda}_t), \hat{\lambda}_t)$;
     $\nu_t \leftarrow \max\left\{L(\hat{Q}_{h,t}, \hat{\lambda}_t) - \underline{L}, \bar{L} - L(\hat{Q}_{h,t}, \hat{\lambda}_t)\right\}$;
     if $\nu_t \leq \nu$ then
         if $\widehat{\mathrm{cost}}(\hat{Q}_{h,t}) \leq \hat{\epsilon} + \frac{|\beta_0|+|\beta_1|+2\nu}{B_\lambda}$ then
             return $(\hat{Q}_{h,t}, \hat{\lambda}_t)$;
         else
             return null
         end
    end
    Set $\theta_{t+1} = \theta_t + \eta\left(\widehat{\mathrm{cost}}(h_t) - \hat{\epsilon}\right)$;
end
```

Following the notation in Agarwal et al. (2019), we define a discretization grid for $[0, 1]$ of size N with $\alpha := 1/N$ and $\mathcal{Z}_\alpha := \{j\alpha \colon j = 1, \ldots, N\}$. Let $\tilde{\mathcal{Y}}_\alpha$ be an $\frac{\alpha}{2}$-cover of $\mathcal{Y}$. The piecewise approximation to the loss function is $l_\alpha(y, u) := l(\underline{y}, [u]_\alpha + \frac{\alpha}{2})$, where $\underline{y}$ is the smallest $\tilde{y} \in \tilde{\mathcal{Y}}_\alpha$ such that $|y - \tilde{y}| \leq \frac{\alpha}{2}$ and $[u]_\alpha$ rounds u down to the nearest integer multiple of α. For a fine enough discretization grid, $\mathrm{loss}_\alpha(f) := \mathbb{E}\left[l_\alpha(Y_i^*, f(X_i))\right]$ approximates $\mathrm{loss}(f)$.

Define $c(y, z) := N \times \left(l(y, z + \frac{\alpha}{2}) - l(y, z - \frac{\alpha}{2})\right)$ and Z_α to be the random variable that uniformly samples $z_\alpha \in \mathcal{Z}_\alpha$ and is independent of the data (X_i, A_i, Y_i^*). For $h_f \in \mathcal{H}$, define

[1] Our error analysis only covers function classes whose Rademacher complexity can be bounded as in Assumption 3.4.1.

the cost-sensitive average loss function as $\text{cost}(h_f) := \mathbb{E}\left[c(\underline{Y}_i^*, Z_\alpha)h_f(X_i, Z_\alpha)\right]$. Lemma 1 in Agarwal et al. (2019) shows $\text{cost}(h_f) + c_0 = \text{loss}_\alpha(f)$ for any $f \in \mathcal{F}$, where $c_0 \geq 0$ is a constant that does not depend on f. Since $\text{loss}_\alpha(f)$ approximates $\text{loss}(f)$, $\text{cost}(h_f)$ also approximates $\text{loss}(f)$. For $Q \in \Delta(\mathcal{F})$, define $Q_h \in \Delta(\mathcal{H})$ to be the induced distribution over threshold classifiers h_f. By the same argument, $\text{cost}(Q_h) + c_0 = \text{loss}_\alpha(Q)$, where $\text{cost}(Q_h) := \sum_{h_f \in \mathcal{H}} Q_h(h)\,\text{cost}(h_f)$ and $\text{loss}_\alpha(Q)$ is defined analogously.

We next relate the predictive disparity measure defined on prediction functions to a predictive disparity measure defined on threshold classifiers. Define $\text{disp}(h_f) := \beta_0 \mathbb{E}\left[h_f(X_i, Z_\alpha) \mid \mathcal{E}_{i,0}\right] + \beta_1 \mathbb{E}\left[h_f(X_i, Z_\alpha) \mid \mathcal{E}_{i,1}\right]$.

Lemma 3.3.1. Given any distribution over (X_i, A_i, Y_i^*) and $f \in \mathcal{F}$, $|\text{disp}(h_f) - \text{disp}(f)| \leq (|\beta_0| + |\beta_1|)\,\alpha$.

Proof. Fix $f \in \mathcal{F}$. For $x \in \mathcal{X}$ and $z_\alpha \in \mathcal{Z}_\alpha$

$$h_f(x, z_\alpha) = 1\{f(x) \geq z_\alpha\} = 1\{\underline{f}(x) \geq z_\alpha\},$$

Therefore,

$$\mathbb{E}_{Z_\alpha}\left[h_f(x, Z_\alpha)\right] = \mathbb{E}_{Z_\alpha}\left[1\{\underline{f}(x) \geq Z_\alpha\}\right] = \underline{f}(x),$$

and for any $a \in \{0, 1\}$,

$$\begin{aligned} &|\mathbb{E}\left[h_f(X, Z_\alpha)|\mathcal{E}_{i,a}\right] - \mathbb{E}\left[f(X)|\mathcal{E}_{i,a}\right]| \\ &= |\mathbb{E}\left[\mathbb{E}_{Z_\alpha}\left[h_f(X, Z_\alpha)\right] - f(X)|\mathcal{E}_{i,a}\right]| \\ &= |\mathbb{E}\left[\underline{f}(X) - f(X)|\mathcal{E}_{i,a}\right]| \leq \alpha \end{aligned}$$

where the first equality uses iterated expectations plus the fact that Z_α is independent of (X, A, Y^*) and the final equality follows by the definition of $\underline{f}(X)$. The claim is immediate after noticing $\text{disp}(h_f) - \text{disp}(f)$ equals $\beta_0\left(\mathbb{E}\left[h_f(X, Z_\alpha) - f(X)|\mathcal{E}_{i,0}\right]\right) + \beta_1\left(\mathbb{E}\left[h_f(X, Z_\alpha) - f(X)|\mathcal{E}_{i,1}\right]\right)$ and applying the triangle inequality. □ □

Lemma 3.3.1 combined with Jensen's Inequality imply $|\,\text{disp}(Q_h) - \text{disp}(Q)| \leq (|\beta_0| + |\beta_1|)\,\alpha$.

Based on these results, we approximate (3.4) with its analogue over threshold classifiers

$$\min_{Q_h \in \Delta(\mathcal{H})} \text{disp}(Q_h) \text{ s.t. } \text{cost}(Q_h) \leq \epsilon - c_0. \tag{3.5}$$

We solve the sample analogue in which we minimize $\widehat{\text{disp}}(Q_h)$ subject to $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$, where $\hat{\epsilon} := \epsilon - \hat{c}_0$ plus additional slack, and $\hat{c}_0, \widehat{\text{disp}}(Q_h)$, $\widehat{\text{cost}}(Q_h)$ are the associated sample analogues. We form the Lagrangian $L(Q_h, \lambda) := \widehat{\text{disp}}(Q_h) + \lambda(\widehat{\text{cost}}(Q_h) - \hat{\epsilon})$ with primal variable $Q_h \in \Delta(\mathcal{H})$ and dual variable $\lambda \in \mathbb{R}^+$. Solving the sample analogue is equivalent to finding the saddle point of the min-max problem $\min_{Q_h \in \Delta(\mathcal{H})} \max_{0 \leq \lambda \leq B_\lambda} L(Q_h, \lambda)$, where $B_\lambda \geq 0$ bounds the Lagrange multiplier. We search for the saddle point by adapting the exponentiated gradient algorithm used in Agarwal et al. (2018, 2019). The algorithm delivers a ν-approximate saddle point of the Lagrangian, denoted $(\hat{Q}_h, \hat{\lambda})$.

3.4 Theoretical Results for Optimizing over the Set of Good Models

The suboptimality of the returned solution $\hat{Q}_h$ can be controlled under conditions on the complexity of the model class $\mathcal{F}$ and how various parameters are set.

Assumption 3.4.1. Let $R_n(\mathcal{H})$ be the Radermacher complexity of $\mathcal{H}$. There exists constants $C, C', C'' > 0$ and $\phi \leq 1/2$ such that $R_n(\mathcal{H}) \leq Cn^{-\phi}$ and $\hat{\epsilon} = \epsilon - \hat{c}_0 + C'n^{-\phi} - C''n^{-1/2}$.

Theorem 3.4.1. Suppose Assumption 3.4.1 holds for $C' \geq 2C + 2 + \sqrt{2\ln(8N/\delta)}$ and $C'' \geq \sqrt{\frac{-\log(\delta/8)}{2}}$. Let n_0, n_1 denote the number of samples satisfying the events $\mathcal{E}_{i,0}, \mathcal{E}_{i,1}$ respectively.

Then, the exponentiated gradient algorithm with $\nu \propto n^{-\phi}$, $B_\lambda \propto n^{\phi}$ and $N \propto n^{\phi}$ terminates in $O(n^{4\phi})$ iterations and returns $\hat{Q}_h$, which when viewed as a distribution over $\mathcal{F}$, satisfies with probability at least $1-\delta$ one of the following: 1) $\hat{Q}_h \neq null$, $\text{loss}(\hat{Q}_h) \leq \epsilon + \tilde{O}(n^{-\phi})$ and $\text{disp}(\hat{Q}_h) \leq \text{disp}(\tilde{Q}) + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi})$ for any $\tilde{Q}$ that is feasible in (3.4); or 2) $\hat{Q}_h = null$ and (3.4) is infeasible.[2]

Proof. The proof strategy follows that of Theorems 2-3 in Agarwal et al. (2019). We consider two cases.

Case 1: There is a feasible solution Q^* to the population problem (3.4) Using Lemmas 3.4.2-3.4.3, the ν-approximate saddle point $\hat{Q}_h$ satisfies

$$\widehat{\text{disp}}(\hat{Q}_h) \leq \widehat{\text{disp}}(Q_h) + 2\nu \tag{3.6}$$

$$\widehat{\text{cost}}(\hat{Q}_h) \leq \hat{\epsilon} + \frac{|\beta_0| + |\beta_1| + 2\nu}{B} \tag{3.7}$$

for any distribution Q_h that is feasible in the empirical problem. This implies that Algorithm 6 returns $\hat{Q} \neq null$. We now show that the returned $\hat{Q}_h$ provides an approximate solution to the discretized population problem.

First, define $\widehat{\text{cost}}_z(h) := \hat{\mathbb{E}}\left[c(\underline{Y}_i^*, z)h(X_i, z)\right]$ and $\text{cost}_z(h) := \mathbb{E}\left[c(\underline{Y}_i^*, z)h(X_i, z)\right]$. Since $c(\underline{Y}_i^*, z) \in [-1, 1]$, we invoke Lemma 2 in Agarwal et al. (2019) with $S_i = c(\underline{Y}_i^*, z_i)$, $U_i = (X_i, z)$, $\mathcal{G} = \mathcal{H}$ and $\psi(s,t) = st$ to obtain that with probability at least $1 - \frac{\delta}{4}$ for all $z \in \mathcal{Z}_\alpha$ and $h \in \mathcal{H}$

$$\left|\widehat{\text{cost}}_z(h) - \text{cost}_z(h)\right| \leq$$

$$2R_n(\mathcal{H}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2\ln(8N/\delta)}{n}} = \tilde{O}(n^{-\phi}),$$

where the last equality follows by the bound on $R_n(\mathcal{H})$ in Assumption 3.4.1 and setting $N \propto n^{\phi}$. Averaging over $z \in \mathcal{Z}_\alpha$ and taking a convex combination of according to $Q_h \in \Delta(\mathcal{H})$ then delivers via Jensen's Inequality that with probability at least $1 - \delta/4$ for all $Q \in \Delta(\mathcal{H})$

$$\left|\widehat{\text{cost}}(Q_h) - \text{cost}(Q_h)\right| \leq \tilde{O}(n^{-\phi}). \tag{3.8}$$

[2] The notation $\tilde{O}(\cdot)$ suppresses polynomial dependence on $\ln(n)$ and $\ln(1/\delta)$

Next, define $\widehat{\mathrm{disp}}_z(h) := \beta_0\hat{\mathbb{E}}\left[h(X_i,z)|\mathcal{E}_{i,0}\right]+\beta_1\hat{\mathbb{E}}\left[h(X_i,z)|\mathcal{E}_{i,1}\right]$ and $\mathrm{disp}_z(h) := \beta_0\mathbb{E}\left[h(X_i,z)|\mathcal{E}_{i,0}\right]+\beta_1\mathbb{E}\left[h(X_i,z)|\mathcal{E}_{i,1}\right]$, where the difference can be expressed as

$$\begin{aligned}
&\widehat{\mathrm{disp}}_z(h) - \mathrm{disp}_z(h) = \\
&\beta_0\left(\hat{\mathbb{E}}\left[h(X_i,z)|\mathcal{E}_{i,0}\right] - \mathbb{E}\left[h(X_i,z)|\mathcal{E}_{i,0}\right]\right) + \\
&\beta_1\left(\hat{\mathbb{E}}\left[h(X_i,z)|\mathcal{E}_{i,1}\right] - \mathbb{E}\left[h(X_i,z)|\mathcal{E}_{i,1}\right]\right).
\end{aligned}$$

Therefore, by the triangle inequality,

$$\begin{aligned}
&\left|\widehat{\mathrm{disp}}_z(h) - \mathrm{disp}_z(h)\right| \le \\
&|\beta_0|\left|\hat{\mathbb{E}}\left[h(X_i,z)|\mathcal{E}_{i,0}\right] - \mathbb{E}\left[h(X_i,z)|\mathcal{E}_{i,0}\right]\right| + \\
&|\beta_1|\left|\hat{\mathbb{E}}\left[h(X_i,z)|\mathcal{E}_{i,1}\right] - \mathbb{E}\left[h(X_i,z)|\mathcal{E}_{i,1}\right]\right|.
\end{aligned}$$

For each term on the right-hand side of the previous display, we invoke Lemma 2 in Agarwal et al. (2019) applied to the data distribution conditional on $\mathcal{E}_0$ and $\mathcal{E}_1$. We set $S = 1$, $U = (X_i, z)$, $\mathcal{G} = \mathcal{H}$ and $\psi(s,t) = st$. With probability at least $1 - \frac{\delta}{4}$ for all $z \in \mathcal{Z}_\alpha$,

$$\begin{aligned}
&\left|\hat{\mathbb{E}}\left[h(X_i,z)|\mathcal{E}_{i,0}\right] - \mathbb{E}\left[h(X_i,z)|\mathcal{E}_{i,0}\right]\right| \le \\
&R_{n_0}(\mathcal{H}) + \frac{2}{\sqrt{n_0}} + \sqrt{\frac{2\ln(8N/\delta)}{n_0}}, \\
&\left|\hat{\mathbb{E}}\left[h(X_i,z)|\mathcal{E}_{i,1}\right] - \mathbb{E}\left[h(X_i,z)|\mathcal{E}_{i,1}\right]\right| \le \\
&R_{n_1}(\mathcal{H}) + \frac{2}{\sqrt{n_1}} + \sqrt{\frac{2\ln(8N/\delta)}{n_1}}.
\end{aligned}$$

Then, averaging over $z \in \mathcal{Z}_\alpha$ and taking a convex combination according to $Q_h \in \Delta(\mathcal{H})$ delivers via Jensen's Inequality that with probability at least $1-\delta/4$ for all $Q \in \Delta(\mathcal{H})$

$$\begin{aligned}\left|\hat{\mathbb{E}}\left[Q_h|\mathcal{E}_{i,0}\right] - \mathbb{E}\left[Q_h|\mathcal{E}_{i,0}\right]\right| \le R_{n_0}(\mathcal{H}) + \frac{2}{\sqrt{n_0}} \\ + \sqrt{\frac{2\ln(8N/\delta)}{n_0}}\end{aligned} \tag{3.9}$$

$$\begin{aligned}\left|\hat{\mathbb{E}}\left[Q_h|\mathcal{E}_{i,1}\right] - \mathbb{E}\left[Q_h|\mathcal{E}_{i,1}\right]\right| \le R_{n_1}(\mathcal{H}) + \frac{2}{\sqrt{n_1}} \\ + \sqrt{\frac{2\ln(8N/\delta)}{n_1}}\end{aligned} \tag{3.10}$$

By the union bound, both inequalities hold with probability at least $1-\delta/2$.

Finally, Hoeffding's Inequality implies that with probability at least $1-\delta/4$,

$$|\hat{c}_0 - c_0| \le \sqrt{\frac{-\log(\delta/8)}{2n}}. \tag{3.11}$$

From Lemma 3.4.5, we have that Algorithm 6 terminates and delivers a distribution $\hat{Q}_h$ that compares favorably against any feasible Q in the discretized sample problem. That is, for any such Q_h,

$$\widehat{\text{disp}}(\hat{Q}_h) \leq \widehat{\text{disp}}(Q_h) + O(n^{-\phi}) \tag{3.12}$$

$$\widehat{\text{cost}}(\hat{Q}_h) \leq \hat{\epsilon} + O(n^{-\phi}) \tag{3.13}$$

where we used the fact that $\nu \propto n^{-\phi}$ and $B \propto n^{\phi}$ by assumption. First, (3.8), (3.11), (3.13) imply

$$\text{cost}(\hat{Q}_h) \leq \hat{\epsilon} + \tilde{O}(n^{-\phi}) \leq \epsilon - c_0 + \tilde{O}(n^{-\phi}), \tag{3.14}$$

where we used that $\hat{\epsilon} = \epsilon - \hat{\mathbb{E}}[l(\underline{Y}_i^*, \frac{\alpha}{2})] + C'n^{-\phi} - C''n^{-1/2}$. by assumption. Second, the bounds in (3.9), (3.10) imply

$$\text{disp}(\hat{Q}_h) \leq \text{disp}(Q_h) + \tilde{O}(n_0^{-\beta}) + \tilde{O}(n_1^{-\phi}). \tag{3.15}$$

We assumed that Q_h was a feasible point in the discretized sample problem. Assuming that (3.8) holds implies that any feasible solution of the population problem is also feasible in the empirical problem due to how we have set C' and C''. Therefore, we have just shown in (3.14), (3.15) that $\hat{Q}_h$ is approximately feasible and approximately optimal in the discretized population problem (3.5). Our last step is to relate $\hat{Q}_h$ to the original problem over $f \in \mathcal{F}$ (3.2).

From Lemma 1 in Agarwal et al. (2019) and (3.14), we observe that

$$\text{loss}_\alpha(\hat{Q}_h) \overset{(1)}{\leq} \epsilon + \tilde{O}(n^{-\phi}),$$
$$\text{loss}(\hat{Q}_h) \overset{(2)}{\leq} \epsilon + \tilde{O}(n^{-\phi}),$$

where (1) used Lemma 1 in Agarwal et al. (2019) and we now view $\hat{Q}_h$ as a distribution of risk scores $f \in \mathcal{F}$, (2) used that $\text{loss}(Q) \leq \text{loss}_\alpha(Q) + \alpha$. Next, from Lemma 3.3.1 and (3.15), we observe that

$$\text{disp}(\hat{Q}_h) \leq \text{disp}(\tilde{Q}) + (|\beta_0| + |\beta_1|)\, \alpha + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi}).$$

where $\hat{Q}_h$ is viewed as a distribution over risk scores $f \in \mathcal{F}$ and $\tilde{Q}$ is now any distribution over risk scores $f \in \mathcal{F}$ that is feasible in the fairness frontier problem. This proves the result for Case I.

Case II: There is no feasible solution to the population problem (3.4) This follows the proof of Case II in Theorem 3 of Agarwal et al. (2019). If the algorithm returns a ν-approximate saddle point $\hat{Q}_h$, then the theorem holds vacuously since there is no feasible $\tilde{Q}$. Similarly, if the algorithm returns $null$, then the theorem also holds. □

□

Theorem 3.4.1 shows that the returned solution $\hat{Q}_h$ is approximately feasible and achieves the lowest possible predictive disparity up to some error. Infeasibility is a concern if no prediction function $f \in \mathcal{F}$ satisfies the average loss constraint. Assumption 3.4.1 is satisfied for instance under LASSO and ridge regression. If Assumption 1 does not hold, FaiRS still delivers good solutions to the sample analogue of Eq. 3.5.

A practical challenge is that the solution returned by the exponentiated gradient algorithm $\hat{Q}_h$ is a stochastic prediction function with possibly large support. Therefore it may be difficult to describe, time-intensive to evaluate, and memory-intensive to store. Results from Cotter et al. (2019) show that the support of the returned stochastic prediction function may be shrunk while maintaining the same guarantees on its performance by solving a simple linear program. The linear programming reduction reduces the stochastic prediction function to have at most two support points and we use this linear programming reduction in our empirical work.

3.4.1 Auxiliary Lemmas for Theoretical Results

Here we provide auxiliary results and their proofs that are used to prove the main theorems and lemmas presented above.

Let $\Lambda := \{\lambda \in \mathbb{R}_+ \colon \lambda \leq B\}$ denote the domain of λ. Throughout this section, we assume we are given a pair $(\hat{Q}_h, \hat{\lambda})$ that is a ν-approximate saddle point of the Lagrangian

$$\begin{aligned} L(\hat{Q}_h, \hat{\lambda}) &\leq L(Q_h, \hat{\lambda}) + \nu \text{ for all } Q_h \in \Delta(\mathcal{H}), \\ L(\hat{Q}_h, \hat{\lambda}) &\geq L(\hat{Q}_h, \lambda) - \nu \text{ for all } 0 \leq \lambda \leq B. \end{aligned}$$

We extend Lemma 1, Lemma 2 and Lemma 3 of Agarwal et al. (2018) to our setting.

Lemma 3.4.1. The pair $(\hat{Q}_h, \hat{\lambda})$ satisfies

$$\hat{\lambda}\left(\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\right) \geq B\left(\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\right)_+ - \nu,$$

where $(x)_+ = \max\{x, 0\}$.

Proof. We consider a dual variable λ that is defined as

$$\lambda = \begin{cases} 0 \text{ if } \widehat{\text{cost}}(\hat{Q}_h) \leq \hat{\epsilon} \\ B \text{ otherwise.} \end{cases}$$

From the ν-approximate optimality conditions,

$$\begin{aligned} \widehat{\text{disp}}(\hat{Q}) + \hat{\lambda}\left(\widehat{\text{cost}}(\hat{Q}) - \hat{\epsilon}\right) &= L(\hat{Q}, \hat{\lambda}) \\ &\geq L(\hat{Q}, \lambda) - \nu \\ &= \widehat{\text{disp}}(\hat{Q}) + \lambda\left(\widehat{\text{cost}}(Q) - \hat{\epsilon}\right), \end{aligned}$$

and the claim follows by our choice of λ. □

Lemma 3.4.2. The distribution $\hat{Q}_h$ satisfies

$$\widehat{\text{disp}}(\hat{Q}_h) \leq \widehat{\text{disp}}(Q_h) + 2\nu$$

for any Q_h satisfying the empirical constraint (i.e., any Q_h such that $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$).

Proof. Assume Q_h satisfies $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$. Since $\hat{\lambda} \geq 0$, we have that

$$L(Q_h, \hat{\lambda}) = \widehat{\text{disp}}(Q_h) + \hat{\lambda}\left(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}\right) \leq \widehat{\text{disp}}(Q_h).$$

Moreover, the ν-approximate optimality conditions imply that $L(\hat{Q}_h, \hat{\lambda}) \leq L(Q_h, \hat{\lambda}) + \nu$. Together, these inequalities imply that

$$L(\hat{Q}_h, \hat{\lambda}) \leq \widehat{\text{disp}}(Q_h) + \nu.$$

Next, we use Lemma 3.4.1 to construct a lower bound for $L(\hat{Q}_h, \hat{\lambda})$. We have that

$$\begin{aligned} L(\hat{Q}_h, \hat{\lambda}) &= \widehat{\text{disp}}(\hat{Q}_h) + \hat{\lambda}\left(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}'\right) \\ &\geq \widehat{\text{disp}}(\hat{Q}_h) + B\left(\widehat{\text{cost}}(\hat{Q}) - \hat{\epsilon}'\right)_+ - \nu \\ &\geq \widehat{\text{disp}}(\hat{Q}_h) - \nu. \end{aligned}$$

By combining the inequalities $L(\hat{Q}_h, \hat{\lambda}) \geq \widehat{\text{disp}}(\hat{Q}_h) - \nu$ and $L(\hat{Q}_h, \hat{\lambda}) \leq \widehat{\text{disp}}(Q_h) + \nu$, we arrive at the claim. □

Lemma 3.4.3. Assume the empirical constraint $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$ is feasible. Then, the distribution $\hat{Q}_h$ approximately satisfies the empirical cost constraint with

$$\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon} \leq \frac{|\beta_0| + |\beta_1| + 2\nu}{B}.$$

Proof. Let Q_h satisfy $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$. Recall from the proof of Lemma 3.4.2, we showed that

$$\widehat{\text{disp}}(\hat{Q}_h) + B\left(\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\right)_+ - \nu \leq L(\hat{Q}_h, \hat{\lambda}) \leq$$

$$\widehat{\text{disp}}(Q_h) + \nu.$$

Therefore, we observe that

$$B\left(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}\right) \leq \left(\widehat{\text{disp}}(Q_h) - \widehat{\text{disp}}(\hat{Q}_h)\right) + 2\nu.$$

Since we can bound $\widehat{\text{disp}}(Q_h) - \widehat{\text{disp}}(\hat{Q}_h)$ by $|\beta_0| + |\beta_1|$, the result follows. □

Lemma 3.4.4. Letting $\rho := \max_{h \in \mathcal{H}} |\widehat{\text{cost}}(h) - \hat{\epsilon}|$, Algorithm 6 satisfies the inequality

$$\nu_t \leq \frac{B \log(2)}{\eta t} + \eta \rho^2 B.$$

For $\eta = \frac{\nu}{2\rho^2 B}$, Algorithm 6 will return a ν-approximate saddle point of L in at most $\frac{4\rho^2 B^2 \log(2)}{\nu^2}$. Since in our setting, $\rho \leq 1$, the iteration complexity of Algorithm 6 is $4B^2 \log(2)/\nu^2$.

Proof. Follows immediately from the proof of iteration complexity in Theorem 3 of Agarwal et al. (2019). Since the cost is bounded on $[-1, 1]$ and $\widehat{\text{cost}}(h) - \hat{\epsilon} \leq \widehat{\text{cost}}(h) \leq 1$ for any $h \in \mathcal{H}$, we see that $\rho \leq 1$. □

Lemma 3.4.5. Suppose that Q_h is any feasible solution to discretized sample problem. Then, the solution $\hat{Q}_h$ returned by Algorithm 6 satisfies

$$\begin{aligned} \widehat{\text{disp}}(\hat{Q}_h) &\leq \widehat{\text{disp}}(Q_h) + 2\nu \\ \widehat{\text{cost}}(\hat{Q}_h) &\leq \hat{\epsilon} + \frac{|\beta_0| + |\beta_1| + 2\nu}{B}. \end{aligned}$$

Proof. This is an immediate consequence of Lemma 3.4.4, Lemma 3.4.2 and Lemma 3.4.3. If the algorithm returns $null$, then these inequalities are vacuously satisfied. □

3.5 Methodology for Optimizing Over the Set of Good Models Under Selective Labels

We now modify the reductions approach to the empirically relevant case in which the training data suffer from the selective labels problem, whereby the outcome Y_i^* is observed only if $T_i \in \mathcal{T}^*$ with $\mathcal{T}^* \subset \mathcal{T}$. The main challenge concerns evaluating model properties over the target population when we only observe labels for a selective (i.e., biased) sample. We propose a solution that uses outcome modeling, also known as extrapolation, to estimate these properties.

To motivate this approach, we observe that average loss and measures of predictive disparity (3.1) that condition on Y_i^* are not identified under selective labels without further assumptions. We introduce the following assumption on the nature of the selective labels problem for the binary decision setting with $\mathcal{T} = \{0, 1\}$ and $\mathcal{T}^* = \{1\}$.

Assumption 3.5.1. The joint distribution $(X_i, A_i, T_i, Y_i^*) \sim P$ satisfies 1) **selection on observables**: $T_i \perp\!\!\!\perp Y_i^* \mid X_i$, and 2) **positivity**: $P(T_i = 1 \mid X_i = x) > 0$ with probability one.

This assumption is common in causal inference and selection bias settings (e.g., Chapter 12 of Imbens and Rubin (2015) and Heckman (1990))[3] and in covariate shift learning (Moreno-Torres et al., 2012). Under Assumption 3.5.1, the regression function $\mu(x) := \mathbb{E}[Y_i^* \mid X_i = x]$ is identified as $\mathbb{E}[Y_i \mid X_i, T_i = 1]$, and may be estimated by regressing the observed outcome Y_i on the features X_i among observations with $T_i = 1$, yielding the outcome model $\hat{\mu}(x)$, as we saw in Chapter 1.

We can use the outcome model to estimate loss on the full population. One approach, *Reject inference by extrapolation* (RIE), uses $\hat{\mu}(x)$ as pseudo-outcomes for the unknown observations (Crook and Banasik, 2004). We consider a second approach, *Interpolation & extrapolation* (IE), which uses $\hat{\mu}(x)$ as pseudo-outcomes for *all* applicants, replacing the $\{0, 1\}$ labels for known cases with smoothed estimates of their underlying risks. Algorithms 7-8 summarize the RIE and IE methods. If the outcome model could perfectly recover $\mu(x)$, then the IE approach recovers an oracle setting for which the FaiRS error analysis continues to hold (Theorem 3.6.1 below).

Algorithm 7 The Reject inference by extrapolation (RIE) approach for addressing missing outcomes due to the the selective labels problem

```
Input: $\{(X_i, Y_i, T_i = 1, A_i)\}_{i=1}^n$;
  Estimate $\hat{\mu}(x)$ by regressing $Y_i \sim X_i \mid T_i = 1$;
  $\hat{Y}(X_i) \leftarrow (1 - T_i)\hat{\mu}(X_i) + T_i Y_i$;
Output: $\{(X_i, \hat{Y}_i(X_i), T_i, A_i)\}_{i=1}^n$;
```

Estimating predictive disparity measures on the full population requires a more general definition of predictive disparity than previously given in Eq. 3.1. Define the modified predictive disparity

[3] Casting this into potential outcomes notation where Y_i^d is the counterfactual outcome if decision d were assigned, we define $Y_i^0 = 0$ and $Y_i^1 = Y_i^*$ (e.g., a rejected loan application cannot default). The observed outcome Y_i then equals $Y_i^1 T_i$.

Algorithm 8 The Interpolation and extrapolation (IE) method for addressing missing outcomes due to the the selective labels problem

```
Input: ${(X_i, Y_i, T_i = 1, A_i)}_{i=1}^n$;
  Estimate $\hat{\mu}(x)$ by regressing $Y_i \sim X_i \mid T_i = 1$;
  $\hat{Y}(X_i) \leftarrow \hat{\mu}(X_i)$;
Output: ${(X_i, \hat{Y}_i(X_i), T_i, A_i)}_{i=1}^n$;
```

measure over threshold classifiers as

$$\begin{aligned}\operatorname{disp}(h_f) = &\beta_0 \frac{\mathbb{E}\left[g(X_i, Y_i) h_f(X_i, Z_\alpha) \mid \mathcal{E}_{i,0}\right]}{\mathbb{E}[g(X_i, Y_i) \mid \mathcal{E}_{i,0}]} + \\ &\beta_1 \frac{\mathbb{E}\left[g(X_i, Y_i) h_f(X_i, Z_\alpha) | \mathcal{E}_{i,1}\right]}{\mathbb{E}[g(X_i, Y_i) \mid \mathcal{E}_{i,1}]},\end{aligned} \tag{3.16}$$

where the nuisance function $g(X_i, Y_i)$ is constructed to identify the measure of interest.[4] To illustrate, the qualified affirmative action fairness-promoting intervention (Def. 3.2.3) is identified as $\mathbb{E}[f(X_i)|Y_i^* = 1, A_i = 1] = \frac{\mathbb{E}[f(X_i)\mu(X_i)|A_i=1]}{\mathbb{E}[\mu(X_i)|A_i=1]}$ under Assumption 3.5.1. This may be estimated by plugging in the outcome model estimate $\hat{\mu}(x)$. Therefore, Eq. 3.16 specifies the qualified affirmative action fairness-promoting intervention by setting $\beta_0 = 0$, $\beta_1 = 1$, $\mathcal{E}_{i,1} = 1\{A_i = 1\}$, and $g(X_i, Y_i) = \hat{\mu}(X_i)$. This more general definition (Eq. 3.16) is only required for predictive disparity measures that condition on events $\mathcal{E}$ depending on both Y^* and A; it is straightforward to compute disparities based on events $\mathcal{E}$ that only depend on A over the full population. To compute disparities based on events $\mathcal{E}$ that also depend on Y^*, we find the saddle point of the following Lagrangian: $L(h_f, \lambda) = \hat{\mathbb{E}}\left[\mathbb{E}_{Z_\alpha}\left[c_\lambda(\underline{\hat{\mu}}_i, A_i, Z_\alpha) h_f(X_i, Z_\alpha)\right]\right] - \lambda\hat{\epsilon}$, where we now use case weights $c_\lambda(\underline{\hat{\mu}}_i, A_i, Z_\alpha) := \frac{\beta_0}{\hat{p}_0} g(X_i, Y_i)(1 - A_i) + \frac{\beta_1}{\hat{p}_1} g(X_i, Y_i) A_i + \lambda c(\underline{\hat{\mu}}_i, Z_\alpha)$ with $\hat{p}_a = \hat{\mathbb{E}}[g(X_i, Y_i) 1\{A_i = a\}]$ for $a \in \{0, 1\}$. Finally, as before, we find the saddle point using the exponentiated gradient algorithm.

3.6 Theoretical Results for Optimizing over the Set of Good Models under Selective Labels

Define $\operatorname{loss}_\mu(f) := \mathbb{E}[l(\mu(X_i), f(X_i))]$ for $f \in \mathcal{F}$ with $\operatorname{loss}_\mu(Q)$ defined analogously for $Q \in \Delta(\mathcal{F})$. The error analysis of the exponentiated gradient algorithm continues to hold in the presence of selective labels under oracle access to the true outcome regression function μ.

Theorem 3.6.1 (Selective Labels)**.** Suppose Assumption 3.5.1 holds and the exponentiated gradient algorithm is given as input the modified training data $\{(X_i, A_i, \mu(X_i)\}_{i=1}^n$.

Under the same conditions as Theorem 3.4.1, the exponentiated gradient algorithm returns $\hat{Q}_h$, which when viewed as a distribution over $\mathcal{F}$, satisfies with probability at least $1 - \delta$ either one of the following: 1) $\hat{Q}_h \neq null$, $\operatorname{loss}_\mu(\hat{Q}_h) \leq \epsilon + \tilde{O}(n^{-\phi})$ and $\operatorname{disp}(\hat{Q}_h) \leq \operatorname{disp}(\tilde{Q}) + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi})$ for any $\tilde{Q}$ that is feasible in (3.4); or 2) $\hat{Q}_h = null$ and (3.4) is infeasible.

Proof. Under oracle access to $\mu(x)$, the bound on cost hold immediately from Theorem 3.4.1. The bound on disparity holds immediately for choices $\mathcal{E}_{i,0}, \mathcal{E}_{i,1}$ that depend on only A. For

[4] Note that we state this general form of g to allow g to use Y_i, for doubly-robust style estimates.

choices of $\mathcal{E}_{i,0}, \mathcal{E}_{i,1}$ that depends on Y_i, such as the qualified affirmative action fairness-enhancing intervention, we rely on Lemma 3.6.1. We first observe that under oracle access to $\mu(x)$, we can identify any disparity as

$$\frac{\beta_1 \mathbb{E}[f(X)g(\mu(X)) \mid A=1]}{\mathbb{E}[g(\mu(X)) \mid A=1]} - \frac{\beta_0 \mathbb{E}[f(X)g(\mu(X)) \mid A=0]}{\mathbb{E}[g(\mu(X)) \mid A=0]}, \tag{3.17}$$

where $g(x) = x$ for the balance for the positive class and qualified affirmative action criteria; $g(x) = (1-x)$ for balance for the negative class; and $g(x) = 1$ for the statistical parity and the affirmative action criteria (see proof of Lemma 3.6.1 below proof for an example). We define the shorthand

$$\begin{aligned}
\omega_1 &:= \mathbb{E}[f(X)g(\mu(X)) \mid A=1] \\
\bar{\omega}_1 &:= \mathbb{E}[g(\mu(X)) \mid A=1] \\
\omega_0 &:= \mathbb{E}[f(X)g(\mu(X)) \mid A=0] \\
\bar{\omega}_0 &:= \mathbb{E}[g(\mu(X)) \mid A=0],
\end{aligned}$$

and we use $\hat{\omega}_1$, $\hat{\bar{\omega}}_1$, $\hat{\omega}_0$, and $\hat{\bar{\omega}}_0$ to denote their empirical estimates. Lemma 3.6.1 gives the following bound on the empirical estimate of the disparity:

$$\begin{aligned}
&P\left[\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \geq \epsilon\right] \\
&\leq 4\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] + 2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \\
&+ 2\exp\left[\frac{-n\bar{\omega}_\wedge^2}{4}\right],
\end{aligned}$$

where $\omega_\vee = \max(\omega_1, \omega_0)$, $\bar{\omega}_\wedge = \min(\bar{\omega}_1, \bar{\omega}_0)$ and $\beta = \max(|\beta_1|, |\beta_0|)$.

We now proceed to relax and simplify the bound. For $\epsilon \leq 4\frac{\beta\omega_\vee}{\bar{\omega}_\wedge}$, we have

$$2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \geq 2\exp\left[\frac{-n\bar{\omega}_\wedge^2}{4}\right].$$

Case 1: We first consider the likely case that $\bar{\omega}_\wedge \geq \omega_\vee$. Then we have

$$2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \leq 2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^2}{64\beta^2}\right].$$

1a) If

$$\frac{\epsilon\bar{\omega}_\wedge}{8\beta} \geq 4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} \tag{3.18}$$

then

$$\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^2}{64\beta^2}\right] \leq \exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right].$$

Then we have

$$P\left[\left|\tfrac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \tfrac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\tfrac{\beta_1\omega_1}{\bar{\omega}_1} - \tfrac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \geq \epsilon\right] \tag{3.19}$$

$$\leq 8\exp\left[-\tfrac{n}{2}\left(\tfrac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \tfrac{2}{\sqrt{n}}\right)^2\right]. \tag{3.20}$$

Inverting this bound yields the following: with probability at least $1-\delta$,

$$\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \leq$$
$$\frac{8\beta}{\bar{\omega}_\wedge}\left(4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2}{n}\log\left(\frac{8}{\delta}\right)}\right).$$

1b)

$$\frac{\epsilon\bar{\omega}_\wedge}{8\beta} < 4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} \tag{3.21}$$

implies that

$$\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \leq$$
$$\frac{8\beta}{\bar{\omega}_\wedge}\left(4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}}\right).$$

Case 2: We now consider the unlikely but plausible case that $\bar{\omega}_\wedge < \omega_\vee$. Then we have

$$\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] \leq$$
$$\exp\left[-\frac{n}{2}\left(\frac{\epsilon\omega_\vee}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right]$$

and

$$\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \leq \exp\left[\frac{-n\epsilon^2\omega_\vee^2}{64\beta^2}\right].$$

We proceed with the same steps as in Case 1 to conclude that with probability at least $1-\delta$,

$$\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \leq$$
$$\frac{8\beta}{\bar{\omega}_\wedge}\left(4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2}{n}\log\left(\frac{8}{\delta}\right)}\right).$$

Applying our assumption that

$$R_n(\mathcal{H}) \leq Cn^{-\phi} \text{ and } \hat{\epsilon} = \epsilon - \hat{c}_0 + C'n^{-\phi} - C''n^{-1/2},$$

for $\phi \leq 1/2$ and $C' \geq 2C + 2 + \sqrt{2\ln(8N/\delta)}$ and $C'' \geq \sqrt{\frac{-\log(\delta/8)}{2}}$, then

$$\text{disp}(\hat{Q}_h) \leq \text{disp}(\tilde{Q}) + \tilde{O}(n^{-\phi}), \tag{3.22}$$

which implies

$$\text{disp}(\hat{Q}_h) \leq \text{disp}(\tilde{Q}) + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi}). \tag{3.23}$$

□

□

In practice, estimation error in $\hat{\mu}$ will affect the bounds in Theorem 3.6.1. The empirical analysis in § 3.8 finds that our method nonetheless performs well when using $\hat{\mu}$.

3.6.1 Auxiliary Lemmas for the Proof of Theorem 3.6.1

Concentration result for disparity under selective labels

Lemma 3.6.1.

$$P\left[\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1}-\frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0}-\left(\frac{\beta_1\omega_1}{\bar{\omega}_1}-\frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right|\geq\epsilon\right]$$
$$\leq 4\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta}-4R_n(\mathcal{G})-\frac{2}{\sqrt{n}}\right)^2\right]+2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right]$$
$$+2\exp\left[\frac{-n\bar{\omega}_\wedge^2}{4}\right],$$

where $\omega_\vee=\max(\omega_1,\omega_0)$, $\bar{\omega}_\wedge=\min(\bar{\omega}_1,\bar{\omega}_0)$ and $\beta=\max(|\beta_1|,|\beta_0|)$.

Proof. For exposition, we first show the steps for qualified affirmative action and then extend the result to the general disparity. We can rewrite the qualified affirmative action criterion as

$$\mathbb{E}[f(X)|Y=1,A=1]=\frac{\mathbb{E}[f(X)\mu(X)|A=1]}{\mathbb{E}[\mu(X)|A=1]}\tag{3.24}$$

where $\mu(x):=\mathbb{E}[Y\mid X=x]$.

$\mathbb{E}[f(X)|Y=1,A=1]$

$$=\frac{\mathbb{E}[f(X)1\{Y=1\}|A=1]}{P(Y=1|A=1)}\tag{3.25}$$
$$=\frac{\mathbb{E}[f(X)\mathbb{E}[1\{Y=1\}|X,A=1]|A=1]}{E[P(Y=1|X,A=1)|A=1]}\tag{3.26}$$
$$=\frac{\mathbb{E}[f(X)P(Y=1|X,A=1)|A=1]}{E[\mu(X)|A=1]}\tag{3.27}$$
$$=\frac{\mathbb{E}[f(X)\mu(X)|A=1]}{E[\mu(X)|A=1]}.\tag{3.28}$$

Assuming access to the oracle μ function, we can estimate this on the full training data as

$$\frac{\hat{\mathbb{E}}[f(X)\mu(X,A=1)|A=1]}{\hat{\mathbb{E}}[\mu(X,A=1)|A=1]}.\tag{3.29}$$

Next we will make use of Lemma 2 of Agarwal et al. (2019), which we restate here again for convenience. Under certain conditions on ϕ and g, with probability at least $1-\delta$

$$\left|\hat{\mathbb{E}}\left[\phi(S,g(U))\right]-\mathbb{E}\left[\phi(S,g(U))\right]\right|\leq$$
$$4R_n(\mathcal{G})+\frac{2}{\sqrt{n}}+\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

We invert the bound by setting $\epsilon=4R_n(\mathcal{G})+\frac{2}{\sqrt{n}}+\sqrt{\frac{2\ln(2/\delta)}{n}}$ and solving for δ to get

$$\delta=2\exp\left[-\frac{n}{2}\left(\epsilon-4R_n(\mathcal{G})-\frac{2}{\sqrt{n}}\right)^2\right].\tag{3.30}$$

Now we can restate Lemma 2 of Agarwal et al. (2019) as

$$\mathbb{P}\Big[\Big|\hat{\mathbb{E}}\left[\phi(S, g(U))\right] - \mathbb{E}\left[\phi(S, g(U))\right]\Big| > \epsilon\Big] \tag{3.31}$$

$$\leq 2\exp\left[-\frac{n}{2}\left(\epsilon - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right].$$

Next we revisit the quantity that we want to bound:

$$\left|\frac{\omega}{\bar{\omega}} - \frac{\hat{\omega}}{\hat{\bar{\omega}}}\right| \tag{3.32}$$

where $\omega = \mathbb{E}[f(X)\mu(X, A=1)|A=1]$ and $\bar{\omega} = \mathbb{E}[\mu(X, A=1)|A=1]$ and correspondingly for $\hat{\omega}$ and $\hat{\bar{\omega}}$. We will rewrite Expression 3.32 as a ratio of differences. We have

$$\left|\frac{\hat{\omega}}{\hat{\bar{\omega}}} - \frac{\omega}{\bar{\omega}}\right| \qquad = \left|\tfrac{\hat{\omega}\bar{\omega} - \hat{\bar{\omega}}\omega}{\hat{\bar{\omega}}\bar{\omega}}\right| \tag{3.33}$$

$$= \left|\tfrac{\bar{\omega}(\hat{\omega}-\omega) - \omega(\hat{\bar{\omega}}-\bar{\omega})}{\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega})+\bar{\omega}^2}\right|. \tag{3.34}$$

$$\tag{3.35}$$

By triangle inequality and union bound, we have

$$\mathbb{P}\Big[|\frac{\bar{\omega}(\hat{\omega}-\omega) - \omega(\hat{\bar{\omega}}-\bar{\omega})}{\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega})+\bar{\omega}^2}| \geq \frac{t}{\bar{\omega}^2/2}\Big]$$

$$< \mathbb{P}\Big[|\bar{\omega}(\hat{\omega}-\omega)| + |\omega(\hat{\bar{\omega}}-\bar{\omega})| \geq t\Big] + \mathbb{P}\Big[|(\hat{\bar{\omega}}-\bar{\omega}) + \bar{\omega}^2| \leq \frac{\bar{\omega}^2}{2}\Big]$$

$$< \mathbb{P}\Big[|\bar{\omega}(\hat{\omega}-\omega)| \geq \frac{t}{2}\Big] + \mathbb{P}\Big[|\omega(\hat{\bar{\omega}}-\bar{\omega})| \geq \frac{t}{2}\Big] + \mathbb{P}\Big[|\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega}) + \bar{\omega}^2| \leq \frac{\bar{\omega}^2}{2}\Big]$$

Since $0 \leq \mu(X, A=1) \leq 1$, we can use a Hoeffding bound for the quantity $|(\hat{\bar{\omega}}-\bar{\omega})|$. Note that $0 \leq \omega \leq \bar{\omega} \leq 1$. Then applying Hoeffding's inequality gives us

$$\mathbb{P}\Big[\Big|\omega(\hat{\bar{\omega}}-\bar{\omega})\Big| \geq \frac{t}{2}\Big] \leq 2\exp\left[\frac{-nt^2}{4\omega^2}\right] \tag{3.36}$$

Next we bound the third term:

$$\mathbb{P}\Big[|\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega}) + \bar{\omega}^2| \leq \frac{\bar{\omega}^2}{2}\Big] \quad \leq \mathbb{P}\Big[|\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega})| \geq \tfrac{\bar{\omega}^2}{2}\Big] \tag{3.37}$$

$$= \mathbb{P}\Big[|(\hat{\bar{\omega}}-\bar{\omega})| \geq \tfrac{\bar{\omega}}{2}\Big] \tag{3.38}$$

$$\leq 2\exp\left[\tfrac{-n\bar{\omega}^2}{4}\right] \tag{3.39}$$

where we again used Hoeffding's inequality for the last line.

We bound the first term using the restated Lemma in 3.31:

$$\mathbb{P}\Big[|\bar{\omega}(\hat{\omega}-\omega)| \geq \frac{t}{2}\Big] \leq 2\exp\left[-\frac{n}{2}\left(\frac{t}{2\bar{\omega}} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right]. \tag{3.40}$$

Now we let $\tilde{\epsilon} = \frac{t}{\bar{\omega}^2/2}$ to get

$$P\left[\left|\frac{\hat{\omega}}{\hat{\bar{\omega}}} - \frac{\omega}{\bar{\omega}}\right| \geq \tilde{\epsilon}\right] \tag{3.41}$$

$$\leq 2\exp\left[-\frac{n}{2}\left(\frac{\tilde{\epsilon}\bar{\omega}}{4} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] + \exp\left[\frac{-n\tilde{\epsilon}^2\bar{\omega}^4}{16\omega^2}\right] + \exp\left[\frac{-n\bar{\omega}^2}{4}\right].$$

Now we turn to the general case. Recalling that we define $\beta = \max(|\beta_1, \beta_0|)$, we have

$$P\left[\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \geq \epsilon\right] \leq$$

$$P\left[|\beta_1|\left|\frac{\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\omega_1}{\bar{\omega}_1}\right| + |\beta_0|\left|\frac{\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \frac{\omega_0}{\bar{\omega}_0}\right| \geq \epsilon\right] \leq$$

$$P\left[\left|\frac{\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\omega_1}{\bar{\omega}_1}\right| \geq \frac{\epsilon}{2\beta}\right] + P\left[\left|\frac{\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \frac{\omega_0}{\bar{\omega}_0}\right| \geq \frac{\epsilon}{2\beta}\right] \leq$$

$$2\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_1}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] + \exp\left[\frac{-n\epsilon^2\bar{\omega}_1^4}{64\beta\omega_1^2}\right] +$$

$$\exp\left[\frac{-n\bar{\omega}_1^2}{4}\right] + 2\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_0}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] +$$

$$\exp\left[\frac{-n\epsilon^2\bar{\omega}_0^4}{64\beta\omega_0^2}\right] + \exp\left[\frac{-n\bar{\omega}_0^2}{4}\right] \leq$$

$$4\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] + 2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right]$$

$$+2\exp\left[\frac{-n\bar{\omega}_\wedge^2}{4}\right]$$

where the first inequality holds by triangle inequality, the second inequality holds by the union bound, the third inequality applies (3.41) for $\tilde{\epsilon} = \frac{\epsilon}{2\beta}$, and the final inequality simplifies the bound using the notation $\omega_\vee = \max(\omega_1, \omega_0)$ and $\bar{\omega}_\wedge = \min(\bar{\omega}_1, \bar{\omega}_0)$. □

3.7 Empirical Results on Benchmark Recidivism Data

We use FaiRS to empirically characterize the range of disparities over the set of good models in a recidivism risk prediction task applied to ProPublica's COMPAS data (Angwin et al., 2016a). Our goal is to illustrate (i) how FaiRS may be used to tractably characterize the range of predictive disparities over the set of good models; (ii) that the range of predictive disparities over the set of good models can be quite large empirically; and (iii) how an auditor may use the set of good models to assess whether the COMPAS risk assessment generates larger disparities than other competing good models. Such an analysis is relevant to business-necessity-type audits of disparate impact (ECOA, 1974; CRA, 1964).

COMPAS is a proprietary risk assessment developed by Northpointe (now Equivant) using up to 137 features (Rudin et al., 2020). As this data is not publicly available, our audit makes use of ProPublica's COMPAS dataset which contains demographic information and prior criminal history for criminal defendants in Broward County, Florida. Lacking access to the data used to train COMPAS, our set of good models may not include COMPAS itself (Angwin et al., 2016a). Nonetheless, prior work has shown that simple models using age and criminal history perform on par with COMPAS (Angelino et al., 2018). These features will therefore suffice to perform our audit. A notable limitation of the ProPublica COMPAS dataset is that it does not contain information for defendants who remained incarcerated. Lacking both features and outcomes for this group, we proceed without addressing this source of selection bias. We also make no distinction between criminal defendants who had varying lengths of incarceration before release, effectively assuming a null treatment effect of incarceration on recidivism. This assumption is based on findings in Mishler (2019) that a counterfactual audit of COMPAS yields equivalent conclusions.

We analyze the range of predictive disparities with respect to race for three common notions of fairness (Definitions 3.2.1-3.2.2) among logistic regression models on a quadratic polynomial of the defendant's age and number of prior offenses whose training loss is near-comparable to COMPAS (loss tolerance $\epsilon = 1\%$ of COMPAS training loss).[5] We split the data 50%-50% into a train and test set. Table 3.1 summarizes the range of predictive disparities on the test set. The disparity minimizing and disparity maximizing models over the set of good of models achieve a test loss that is comparable to COMPAS.

For each predictive disparity measure, the set of good models includes models that achieve significantly lower disparities than COMPAS. In this sense, COMPAS generates "unjustified" disparate impact as there exists competing models that would reduce disparities without compromising performance. Notably, COMPAS' disparities are also larger than the maximum disparity over the set of good models. For example, the difference in COMPAS' average predictions for black relative to white defendants is strictly larger than that of any model in the set of good models (Table 3.1, SP). Interestingly, the minimal balance for the positive class and balance for the negative class disparities between black and white defendants over the set of good models are strictly positive (Table 3.1, BFPC and BFNC). For example any model whose performance lies in a neighborhood of COMPAS' loss has a higher false positive rate for black defendants than white defendants. This suggests while we can reduce predictive disparities between black and white defendants relative to COMPAS on all measures, we may be unable to eliminate balance for the positive class and balance for the negative class disparities without harming predictive performance.

In addition to the retrospective auditing considered in this section, characterizing the range of predictive disparities over the set of good models is also important for model development and selection. The next section shows how to construct a more equitable model that performs comparably to a benchmark.

[5] We use a quadratic form following the analysis in Rudin et al. (2020).

Table 3.1: Results showing fairness properties over the set of good models trained on Propublica's recidivism prediction dataset. Our optimization method shows that the redicivism prediction model used in practice, COMPAS, fails to satisfy the “business necessity" defense for disparate impact by race. The set of good models (performing within 1% of COMPAS's training loss) includes models that achieve significantly lower disparities than COMPAS. The first panel (SP) displays the disparity in average predictions for black versus white defendants (Def. 3.2.1). The second panel (BFPC) analyzes the disparity in average predictions for black versus white defendants in the positive class, and the third panel examines the disparity in average predictions for black versus white defendants in the negative class (Def. 3.2.2). Standard errors are reported in parentheses. See § 3.7 for details.

	Min. Disp.	Max. Disp.	COMPAS
SP	−0.060 (0.004)	0.120 (0.007)	0.194 (0.013)
BFPC	0.049 (0.005)	0.125 (0.012)	0.156 (0.016)
BFNC	0.044 (0.005)	0.117 (0.009)	0.174 (0.016)

3.8 Empirical Results on Real-World Consumer Lending Data

Suppose a financial institution wishes to replace an existing credit scoring model with one that has better fairness properties and comparable performance, if such a model exists. We show how to accomplish this task by using FaiRS to find the absolute predictive disparity-minimizing model over the set of good models. On a real world consumer lending dataset with selectively labeled outcomes, we find that this approach yields a model that reduces predictive disparities relative to the benchmark without compromising overall performance.

We use data from Commonwealth Bank of Australia, a large financial institution in Australia (henceforth, "CommBank"), on a sample of 7,414 personal loan applications submitted from July 2017 to July 2019 by customers that did not have a prior financial relationship with CommBank. A personal loan is a credit product that is paid back with monthly installments and used for a variety of purposes such as purchasing a used car or refinancing existing debt. In our sample, the median personal loan size is AU$10,000 and the median interest rate is 13.9% per annum. For each loan application, we observe application-level information such as the applicant's credit score and reported income, whether the application was approved by CommBank, the offered terms of the loan, and whether the applicant defaulted on the loan. There is a selective labels problem as we only observe whether an applicant defaulted on the loan within 5 months (Y_i) if the application was funded, where “funded" denotes that the application is both approved by CommBank and the offered terms were accepted by the applicant. In our sample, 44.9% of applications were funded and 2.0% of funded loans defaulted within 5 months.

Motivated by a decision maker that wishes to reduce credit access disparities across geographic regions, we focus on the task of predicting the likelihood of default $Y_i^* = 1$ based on information in the loan application X_i while limiting predictive disparities across SA4 geographic regions within Australia. SA4 regions are statistical geographic areas defined by the Australian Bureau of Statistics (ABS) and are analogous to counties in the United States. An SA4 region is classified as socioeconomically disadvantaged ($A_i = 1$) if it falls in the top quartile of SA4 regions based on the ABS' Index of Relative Socioeconomic Disadvantage (IRSD), which is an index that aggregates census data related to socioeconomic disadvantage.[6] Applicants from disadvantaged SA4 regions are under-represented among funded applications, comprising 21.7% of all loan applications, but only 19.7% of all funded loan applications.

Our experiment investigates the performance of FaiRS under our two proposed extrapolation-based solutions to selective labels, RIE and IE (See Algorithms 7-8), as well as the Known-Good Bad (KGB) approach that uses only the selectively labelled population. Because we do not observe default outcomes for all applications, we conduct a semi-synthetic simulation experiment by generating synthetic funding decisions and default outcomes. On a 20% sample of applicants, we learn $\pi(x) := \hat{P}(T_i = 1|X_i = x)$ and $\mu(x) := \hat{P}(Y_i = 1|X_i = x, T_i = 1)$ using random forests. We generate synthetic funding decisions $\widetilde{T}_i$ according to $\widetilde{T}_i \mid X_i \sim Bernoulli(\pi(X_i))$ and synthetic default outcomes $\widetilde{Y}_i^*$ according to $\widetilde{Y}_i^* \mid X_i \sim Bernoulli(\mu(X_i))$. We train all models as if we only knew the synthetic outcome for the synthetically funded applications. We estimate $\hat{\mu}(x) := \hat{P}(\widetilde{Y}_i = 1|X_i = x, \widetilde{T}_i = 1)$ using random forests and use $\hat{\mu}(X_i)$ to generate the pseudo-outcomes $\hat{Y}(X_i)$ for RIE and IE as described in Algorithms 7 and 8. As benchmark models, we use the loss-minimizing linear models learned using KGB, RIE, and IE approaches, whose respective training losses are used to select the corresponding loss tolerances ϵ. We use the class of linear models for the FaiRS algorithm for KGB, RIE, and IE approaches.

We compare against the fair reductions approach to classification (fairlearn) and the Target-Fair Covariate Shift (TFCS) method in Coston et al. (2019). TFCS iteratively reweighs the training data via gradient descent on an objective function comprised of the covariate shift-reweighed classification loss and a fairness loss. Fairlearn searches for the loss-minimizing model subject to a fairness parity constraint (Agarwal et al., 2018). The fairlearn model is effectively a KGB model since the fairlearn package does not offer modifications for selective labels.[7] We use logistic regression as the base model for both fairlearn and TFCS. Results are reported on all applicants in a held-out test set, and performance metrics are constructed with respect to the synthetic outcome $\widetilde{Y}_i^*$.

Figure 3.1 shows the AUC (y-axis) against disparity (x-axis) for the KGB, RIE, IE benchmarks and their FaiRS variants as well as the TFCS models and fairlearn models. Colors denote the adjustment strategy for selective labels, and the shape specifies the optimization method. The first row evaluates the models on all applicants in the test set (i.e., the target population). On the target population, FaiRS with reject extrapolation (RIE and IE) reduces disparities while achieving performance comparable to the benchmarks and to the reweighing approach (TFCS). It also achieves lower disparities than TFCS, likely because TFCS optimizes a non-convex objective function and may therefore converge to a local minimum. Reject extrapolation achieves better AUC than all KGB models, and only one KGB model (fairlearn) achieves a lower disparity. The second row evaluates the models on only the *funded* applicants. Evaluation on the funded cases underestimates disparities across the methods and overestimates AUC for the TFCS and KGB models. This highlights that failure to account for the selective labels problem can lead to invalid models and invalid evaluations.

[6] Complete details on the IRSD may be found in Australian Bureau of Statistics (2016).

[7] To accommodate reject inference, a method must support real-valued outcomes. The fairlearn package does not, but the related *fair regressions* does (Agarwal et al., 2019). This is sufficient for SP (Def. 3.2.1), but other parities such as BFPC and BFNC (Def. 3.2.2) require further modifications as discussed in § 3.5.

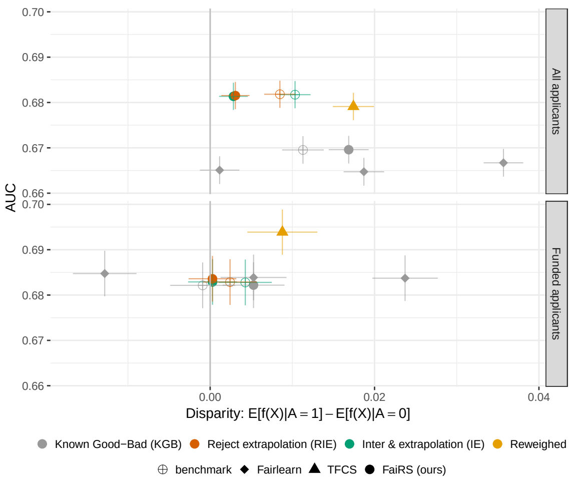


Figure 3.1: Semi-synthetic results for characterizing fairness properties over the set of good models trained on real-world credit lending data. This plot shows the area under the ROC curve (AUC) with respect to the synthetic outcome against disparity in the average risk prediction for the disadvantaged $(A_i = 1)$ vs advantaged $(A_i = 0)$ groups. FaiRS reduces disparities for the RIE and IE approaches while maintaining AUCs comparable to the benchmark models (first row). Evaluation on only funded applicants (second row) overestimates the performance of TFCS and KGB models and underestimates disparities for all models. Error bars show the 95% confidence intervals. See § 3.8 for details.

3.9 Conclusion

This chapter developed a framework, Fairness in the Rashomon Set (FaiRS), to characterize the range of predictive disparities and find the absolute disparity minimizing model over the set of good models. FaiRS is suitable for a variety of applications including settings with selectively labelled outcomes where the selection decision and outcome are unconfounded given the observed features. The method is generic, applying to both a large class of prediction functions and a large class of predictive disparities.

This concludes our discussion, guided by the principles of validity and equity, on methods for constructing and evaluating counterfactual risk assessments. We next turn to the more existential question of whether algorithmic risk assessments are suitable for use in a particular setting. Our investigation into this question will continue to consider aspects of validity and equity alongside our third principle, governance.

4

Framework for Evaluating the Validity of Algorithms in Consequential Decisions

Data-driven algorithmic decision-making, in theory, can afford improvements in efficiency and the benefits of evidence-based decision making. Yet in practice, data-driven decision systems, often taking the form of algorithmic risk assessments, have caused significant adverse consequences in high-stakes settings. Investigators have identified unintended and often biased behavior in algorithmic decision systems used in a variety of applications, from detecting unemployment and welfare fraud to determining pre-trial release decisions and child welfare screening decisions, as well as in algorithms used to inform medical care and set insurance premiums (Eubanks, 2018; Angwin et al., 2016b; Obermeyer et al., 2019b; Vyas et al., 2020; Gilman, 2020; Charette, 2018; Angwin et al., 2017; Fabris et al., 2021). These high-profile incidents have brought into focus key questions such as how we can anticipate these harms before deployment and whether algorithms are suitable for decision-making tasks.

In this chapter, we argue that these questions of governance require that we evaluate the validity of algorithms. We present a framework for using validity considerations to help govern decisions about whether to build and deploy algorithmic decision systems. The contributions in this chapter were originally published in Coston et al. (2023).

To anticipate harms before deployment, researchers and practitioners have proposed a suite of tools and processes to address value-alignment, such as how to promote fairness and establish transparency and accountability (Digital and Office, 2018; Madaio et al., 2020b; Raji et al., 2020; Mitchell et al., 2019a; Gebru et al., 2021). More recently, there have been growing calls to assess the appropriateness of using predictive tools for complex, real-world tasks from a *validity* perspective (Raji et al., 2022). In many cases where algorithms prove unsuitable for real-world use, the problem originates in the initial problem formulation stages (Passi and Barocas, 2019; Barocas and Selbst, 2016b), or in the process of operationalizing latent constructs of interest (e.g., worker well-being, risk of recidivism, or socioeconomic status) via more readily observable measures and indicators (Jacobs and Wallach, 2021; Narayanan, 2019; Recht, 2022).

Without addressing these issues directly, it may be challenging or impossible to align the resulting model with human values after the fact. In some cases, efforts to do so may actually backfire because of unaddressed upstream issues.

Our work seeks to center validity considerations, a crucial criterion for the justified use of algorithmic tools in real-world decision-making (Jacobs and Wallach, 2021; Narayanan, 2019; Recht, 2022). In doing so, we situate our work at the intersection of research that debates algorithm refusal versus repair and research that develops artifacts for responsible ML. Guided by the goal of delivering an accessible tool to promote deliberation and reflection around validity, we propose a structure for a question-and-answer (Q&A)-based protocol.

The main components of this chapter are as follows:

1. We provide a working taxonomy of criteria for the justified use of algorithms in high-stakes settings. We utilize this taxonomy to illuminate two important principles for substantiating/refuting the use of ML for decision making: validity and reliability (§ 4.1).
2. We use this taxonomy to conduct an interdisciplinary literature review on validity, reliability, and value-alignment (§ 4.2).
3. We connect modern validity theory from the social sciences to common challenges in problem formulation and data issues that jeopardize the validity of predictive algorithms in decision making (§ 4.3).
4. We demonstrate how this systematization can inform future work by sketching the structure for a protocol to promote deliberation on validity.

Throughout the chapter we will discuss validity in the context of several high-stakes settings where predictive algorithms are increasingly used to inform human decisions: pre-trial release in the criminal justice system and screening decisions in the child welfare system. In the criminal justice setting, judges must decide whether to release a defendant before trial based on the likelihood that, if released, the defendant will fail to appear for trial as well as the likelihood the defendant will be arrested for a new crime before trial (Kleinberg et al., 2018). For the child welfare screening task, call workers must decide which reports of alleged child abuse or neglect should be screened in for investigation based on an assessment of the likelihood of immediate danger or long-term neglect if no further action is taken (Chouldechova et al., 2018).

4.1 A Taxonomy of Criteria for Justified Use of Data-driven Algorithms

To assess whether the use of data-driven algorithms is adequately justified in a given decision making context, one must account for a wide range of factors. To give structure to this vast array of considerations, we propose a high-level taxonomy: we posit that the justified use of algorithmic tools requires *at minimum* accounting for validity, value-alignment, and reliability. In this section, we offer a precise definition for these terms. § 4.2 offers an overview of existing literature on each of these topics.

The rationale for our taxonomy: To evaluate whether the use of predictive tools is sufficiently justified in a high-stakes decision making domain, at a minimum, we need to answer the following sequence of questions:

- Can we translate (parts of) the decision-making task into a prediction problem where both a measure representing the construct we'd like to predict and predictive attributes are available in the observed data?

- If the answer to the above question is affirmative, does the model we train align with stakeholders' values, such as impartiality and non-discrimination?
- Do we understand the longer-term consequences of deploying the model in decision making processes? For example, how might the deployment setting change over time and can the model be reliably utilized under this changing environment?

The above questions motivate our three high-level categories of considerations for justifying/refuting the use of data-driven algorithms in decision making: validity, value alignment, and reliability.

Before we elaborate on our taxonomy, two remarks are in order. First, we emphasize that a formal, comprehensive taxonomy of considerations around justified-use of algorithms is a formidable research question in itself, and the purpose of our taxonomy is limited to structuring our review of the available literature, tools and resources. We make no claims regarding the comprehensiveness of our taxonomy. We refer the interested reader to treatises on the subject including Fjeld et al. (2020); Floridi and Cowls (2021); Golbin et al. (2020). Additionally, we note that the three categories at the heart of our taxonomy are intimately connected, rather than mutually exclusive.

Our first category of considerations, validity, aims to establish that the system does what it purports to do. As we have seen throughout this dissertation, this quality is much harder to satisfy than one might initially expect. For an additional example, consider the task of predicting which criminal defendants are likely to reoffend. Predictive models are often trained using re-arrest outcomes (Fogliato et al., 2021). Whether a model predicting re-arrest actually predicts reoffense is subject to considerable debate, particularly given that a large body of work has established racial disparities in arrests even for crimes which have little differences in prevalence by race (Alexander, 2011). A model that appears accurate with respect to re-arrests may be quite inaccurate with respect to actual crime. More broadly, the notion of validity requires not only that the system has to predict what it purports to predict, but also must achieve acceptable accuracy both within and outside the training environment (in the real-world deployment). These validity criteria are adapted from validity considerations (e.g., *construct validity*, *internal validity*, and *external validity*) that are widely adopted in social sciences, including psychology, psychometrics, and Human-Computer Interaction (Campbell, 1957; Messick, 1995; Gergle and Tan, 2014).

Definition 4.1.1 (Validity)**.** A measure, test, or model is valid if it closely reflects or assesses the specific concept/construct that the designer intends to measure (Drost, 2011).

We say that a predictive algorithm is valid when it predicts the quantity that we think it does, and similarly we say that an audit or assessment is valid when it evaluates the quantity that we would like to audit or assess. Threats to validity can arise as early as the problem formulation stage where decisions about how to operationalize the problem can induce misalignment between what we intend to predict versus what the model actually predicts (Passi and Barocas, 2019; Jacobs and Wallach, 2021). When validity does not hold, it is quite challenging to assess value-alignment—our next category of considerations. In this sense, we claim that validity is a prerequisite for the more commonly discussed values such as fairness.

Our second category of considerations focuses on the compliance of the system with stakeholders' values.

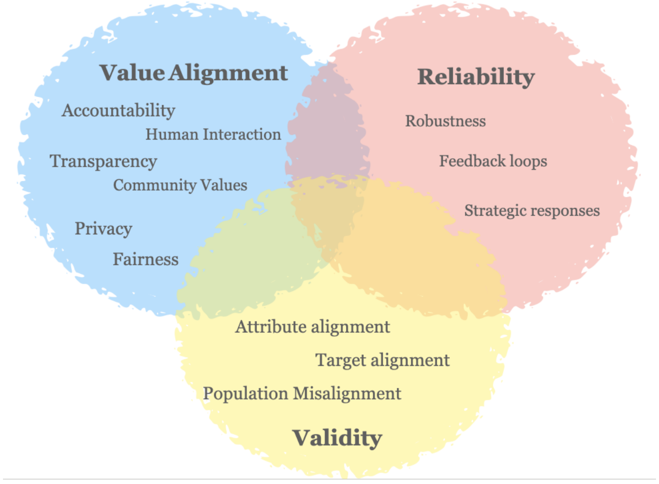


Figure 4.1: Visual representation of our proposed taxonomy for the justified use of algorithms in high-stakes decision making. Validity, reliability and value alignment are required for justified use. These concepts are overlapping and interconnected, encompassing many aspects of responsible machine learning.

Definition 4.1.2 (Value-alignment)**.** Value-alignment requires that the goals and behavior of the system comply with values of relevant stakeholders and communities (Sierra et al., 2021).

Relevant stakeholders might include the communities that will impacted by the algorithmic system or the frontline workers who will work with the system. Commonly discussed values include fairness, privacy, transparency, and accountability. Properties like simplicity and interpretability are often desired as a means to ensure these values (Rudin et al., 2020), and within this taxonomy, we include these properties under the broad umbrella of value-alignment.

The final set of considerations that we will discuss concern reliability over time and context.

Definition 4.1.3 (Reliability)**.** Reliability is the extent to which the output of a measurement/test/model is *repeatable, consistent, and stable* — when different persons utilize it, on different occasions, under different conditions, with alternative instruments that measure the same thing (Drost, 2011).

Reliability concerns in part the dynamical nature of systems in the real world. A system that satisfies our previous two criteria at a given snapshot in time may soon after experience a policy, population, or other notable change that can have profound effects on its validity and value-alignment. Threats to reliability include changes in the population characteristics and/or risk profiles (i.e., distribution shift) or strategic behavior in response to the algorithmic model predictions.

We use this taxonomy to structure a literature review of related work in the following section.

4.2 Background and Related Work

In this section we conduct a structured literature review of prior work in validity, value-alignment, and reliability.

4.2.1 Validity

We begin our literature review with validity. The machine learning literature has vibrant communities addressing validity-related considerations, such as selection bias and representation bias, but, to the best of our knowledge, there is no unifying validity framework around these issues. For this we turn to the theory of validity in the social sciences. In this section we review key concepts from social science research on validity, and in subsequent sections we translate these concepts to the setting of data-driven algorithms.

Construct validity is concerned with whether the measure captures what the researcher intended to measure. Modern validity theory often defines construct validity as the overarching concern of validity research: construct validity integrates considerations of content, criteria, and consequences into a unified construct framework (Messick, 1995; Schotte et al., 1997). Messick (1995) and Gergle and Tan (2014) highlight distinguishable aspects of construct validity. Below we review the definition of different aspects of construct validity, highlighting aspects that are particularly relevant in assessing the validity of data-driven decision-making algorithm.

- ***Face validity*** means that the chosen measure “appears to measure what it is supposed to measure" (Gergle and Tan, 2014). For example, imagine you propose to assess or predict the online satisfaction with a product on an e-commerce website by measuring the proportion of positive comments among all the purchase comments. You feel that the higher the proportion of the positive comments, the more satisfied the customers were, so “on its face” it is a valid measure or prediction target. Face validity is a very weak requirement and should be used analogously to rejecting the null in hypothesis testing: rejecting face validity allows us to conclude that the measure is not valid, but failure to reject face validity does not allow us to conclude it is valid.

- ***Convergent validity*** uses more than one measure for the same construct and then demonstrates a correlation between the two measures at the same point in time. One common way to examine convergent validity is to compare your measure with a gold-standard measure or benchmark. However, Gergle and Tan (2014) warned that convergent validity can suffer from the fact that the secondary variable for comparison may have similar limitations as the measure under investigation.

- ***Discriminant validity*** tests whether measurements of two concepts that are supposed to be unrelated are, in fact, unrelated. Historically researchers have struggled to demonstrate discriminant validity for measures of social intelligence because these measures correlate highly with measures of mental alertness (Campbell and Fiske, 1959).

- ***Predictive validity*** is a validation approach where the measure is shown to accurately predict some other conceptually related variable later in time. For example, in the context of child welfare, Vaithianathan et al. (2020) demonstrated the predictive validity of Allegheny Family Screening Tool (AFST) by showing that the AFST’s home removal risk score *at the time of a maltreatment referral*, was also sensitive to identifying children with a heightened risk of an emergency department (ED) visit or hospitalization because of injury *during the follow-up period*. Therefore, they argued “the risk of placement into foster care as a reasonable proxy for child harm and therefore a credible outcome for training risk stratification models for use by Child Protective Services systems" (Vaithianathan et al., 2020).

Internal validity and **external validity** are important validity considerations in experimental research (Campbell, 1957; Gergle and Tan, 2014). Internal validity is the degree to which the claims of a study hold true for the particular (often artificial) study setting, while external validity is the degree to which the claims hold true for real-world contexts, with varying cultures, different population, different technological configurations, or varying times of the day (Gergle and Tan, 2014). Gergle and Tan (2014) discussed three common ways to bolster external validity in study design: (1) choosing a study task that is a good match for the kinds of activities in the field, (2) choosing participants for the study that are as close as possible to those in the field, and (3) assessing the similarity of the behaviors between the laboratory study and the fieldwork.

Prior work on data-driven decision-making algorithms has probed various aspects of validity threats or concerns, often using the vocabulary of “measurement error", “problem formulation", and “biases". For example, Passi and Barocas (2019) chronicle how the analysts' decisions during problem formulation impacts fairness of the downstream model. Relatedly Jacobs and Wallach (2021) demonstrate that how one operationalizes theoretical constructs into measurable quantities impacts fairness. Suresh and Guttag (2021) also highlight measurement error in their characterization of seven types of harm in machine learning and describe other biases in representation and evaluation that can threaten validity. Representation and evaluation biases can occur when the development sample and evaluation sample, respectively, do not accurately represent who is in the target population. To the best of our knowledge, there is no prior work that proposes tools or processes centered around validity issues. In this chapter, we aim to fill this gap by drawing on the findings in these papers to structure a validity-centered artifact intended for real-world use.

4.2.2 Value Alignment

The literature on value-alignment is vast, and we therefore focus on the works most related to our purpose of developing *artifacts*, such as documents, checklists, and software toolkits, to promote justifying the use of algorithmic systems in decision-making. Documentation artifacts designed to improve transparency and inform trust have been proposed for datasets, machine learning models, and AI products and services (Holland et al., 2020; Gebru et al., 2021; Hutchinson et al., 2021; Mitchell et al., 2019a; Arnold et al., 2019). These artifacts document typical use cases, product/development lineage, and other important specificatons in order to promote proper use as the models, data, and services are shared and re-used across a variety of contexts. Noticing that these documentation products largely represent the perspective of algorithm developers, Krafft et al. (2021b) developed a toolkit designed to engage community advocates and activists in this process.

An increasingly popular mechanism is checklists for fairness and ethics in machine learning. Checklists can provide a structured form for individual advocates to raise fairness or ethics concerns, but a compliance-oriented checklist may fail to capture the nuances of complex fairness and ethical challenges (Madaio et al., 2020b). Recent work has advocated for checklists designed to promote conversations about ethical challenges (Madaio et al., 2020a). However, checklist-style “yes or no" questions may be ill-suited for promoting deliberation. Moreover, in centering around the question “*have we performed all the steps necessary before releasing the model?*”, checklists adopt a “deploy by default" framing that may encourage practitioners to err on the side of brushing concerns aside. To address these issues, we sketch a protocol to promote deliberation centered around the question *“is an algorithmic model appropriate for use in this setting?”*.

Raji et al. (2020) proposed a conceptual framework, SMACTR, for developing an internal audit for algorithmic accountability throughout the machine learning development cycle. The proposed methodology is general-purpose and comprehensive, involving other documentation and checklists discussed in this section (like model cards and datasheets), but this general-purpose methodology may be complicated, expensive and time-consuming to implement, perhaps prohibitively so for teams with limited bandwidth such as the analytics division of a public sector organization. Of note, the SMACTR methodology does not focus on issues of validity. For a given class of problems (e.g., predictive analytics for decision support) there are a set of common validity issues and questions that can be detailed and re-used across contexts. Doing so would complement the SMACTR methodology.

Based on impact assessments in other domains like construction, algorithmic impact assessments (AIAs) require algorithm developers to evaluate the impacts of the proposed algorithm on society at large and particularly on marginalized communities (Reisman et al., 2018; Janssen, 2020; Metcalf et al., 2021). In 2019 the Government of Canada made it compulsory for a government agency using an algorithm to conduct an algorithmic impact assessment (Canada, 2019). A comprehensive AIA will likely need to involve deliberation about validity issues since an invalid algorithm may very well cause adverse impacts. Related to AIA is the UK Government's Data Ethics Framework which asks practitioners to perform a self-assessment of their transparency, fairness, and accountability (Digital and Office, 2018). The framework asks the respondent to identify user needs, consider both the benefits and unintended/negative consequences of the project, and to assess whether historical bias or selection bias may be present in the data. This framework is helpful in its breadth and specificity. However, the framework does not address core validity issues like proxy outcomes.

A number of toolkits are available to visualize the performance metrics and tradeoffs therein of algorithmic models. Yu et al. (2020) propose a two-step method to communicate tradeoffs to algorithm designers that involves first generating a family of predictive models and subsequently plotting their performance metrics. Visualization software has been developed to communicate tradeoffs to algorithm designers (Yu et al., 2020) and to display intersectional group disparities (Cabrera et al., 2019). A number of fairness/ethics toolkits and code repositories are available to help researchers probe model disparities and explore potential mitigations (Adebayo et al., 2016; Bellamy et al., 2018; Saleiro et al., 2018).

A strain of the literature develops pedagogical processes for improving educational instruction of ethics issues in data science curriculum. Shen et al. (2021) proposed a toolkit, Value Cards, to facilitate deliberation among computer science students and practitioners. The Value Cards largely focus on tradeoffs between performance metrics, stakeholder perspectives, and algorithmic impacts. Bates et al. (2020) describes the experience of integrating ethics and critical data studies into a masters of data science program.

Guides for best practices in selecting a predictive algorithm for high-stakes settings have been proposed for public policy and healthcare settings (Kleinberg et al., 2017; Fazel and Wolf, 2018). For instance, Kleinberg et al. (2017) discuss conceptual issues such as target specification, measurement issues, omitted payoff bias, and selective labels. Our work connects these issues, among others, to established concepts of validity from the social sciences.

4.2.3 Reliability

As mentioned earlier, *"reliability is the extent to which measurements are* repeatable *— when different persons perform the measurements, on different occasions, under different* *conditions, with supposedly alternative instruments which measure the same thing"* (Drost, 2011). Reliability encompasses reproducibility. Reliability is also defined as the *consistency* of measurement (Bollen, 1989), and the *stability* of measurement results over a variety of conditions (Nunnally, 1994). Reliability is necessary but not sufficient to ensure validity. That is, reliability of a measure does not imply its validity; however, a highly unreliable measure cannot be valid (Nunnally, 1994).

Drost (2011) enumerates three main dimensions of reliability: equivalence (of measurements across a variety of tests), stability over time, and internal consistency (consistency over time). There are several general classes of reliability considerations:

- **Inter-rater reliability** assesses the degree of agreement between two or more raters in their appraisals. Low inter-rater reliability could be a potential concern in human-in-the-loop designs where human decision-makers receive the predictions of a ML model, and interpret them to reach the final decisions.

- **Test-retest reliability** assesses the degree to which test scores are consistent from one test administration to the next. Population shifts (Quiñonero-Candela et al., 2008), feedback loops (Ensign et al., 2018), and strategic responses (Hardt et al., 2016a) are among the threats to the test-retest reliability of risk assessment instruments.

- **Inter-method reliability** assesses the degree to which test scores are consistent when there is a variation in the methods or instruments used. For example, suppose two different models are independently trained to predict the risk of default by loan applicants. Inter-method reliability assesses whether these models often reach similar predictions for the same loan applicants. Another area in which inter-method reliability is applicable to ML is the extent to which an ML model can reproduce the decisions made by human decision-makers.

- **Internal consistency reliability**, assesses the consistency of results across items within a test. Models that make significantly different predictions for similar inputs may violate this notion of reliability.

Efforts in emerging areas such as MLOps focus on the development of practical tools to assess and ensure the reliability of data-driven predictive analytics (Kreuzberger et al., 2022; Shankar and Parameswaran, 2021; Zaharia et al., 2018). While these efforts are still in their infancy, there is a growing body of work pointing to an urgent need for better tooling (Kreuzberger et al., 2022; Shankar and Parameswaran, 2021). For example, Veale et al. identified key challenges for public sector adoption of algorithmic fairness ideas and methods, highlighting the risks posed by changes in policy, data practices, or organizational structures (Veale et al., 2018). Focusing on the private sector, Holstein et al. (2019b) identified what large companies need to improve fairness in machine learning, highlighting the need for "domain-specific frameworks that can help them navigate any associated complexities." In addition to the above changes, feedback loops and strategic responses can induce population shifts, also known as distribution shift or dataset shift (Moreno-Torres et al., 2012). The literature on data shift concerns the fast detection and characterization of distribution shifts, including distinguishing harmful shifts from inconsequential ones (Rabanser et al., 2019; Ashmore et al., 2021). An active area of research in machine learning aims to design learning algorithms that make accurate predictions even if decision subjects respond strategically to the trained model (see, e.g., (Dong et al., 2018; Hardt et al., 2016a; Mendler-Dünner et al., 2020; Shavit et al., 2020;

Hu et al., 2019)). Generalizing such settings, Perdomo et al. (2020) propose a framework called *performative predictions*, which broadly studies settings in which the act of predicting influences the prediction target.

While our work focuses on validity issues, we hope that it serves as a jumping off point for future work on reliability artifacts for predictive analytics.

4.3 A Taxonomy for Common Threats to Validity of Predictive Models

This section delves into common challenges that jeopardize validity. We organize these challenges into three groups: population misalignment, attribute misalignment, and target misalignment. We connect these groups to notions of validity from the social sciences mentioned in § 4.2.

4.3.1 Attribute Misalignment

To make meaningful predictions, we must have data on pertinent predictive factors, ideally ones for which we can point to evidence supporting the claim that they are relevant to the predictive task at hand. The choice of which features to use in prediction has clear implications for internal, external, and construct validity. If there is no plausible causal path between the target and a feature such that any correlation is entirely spurious, the inclusion of the feature immediately challenges internal and external validity. Additionally, it can fail tests of face validity. A particularly pressing example of a prediction task that lacks face validity is the use of images of human faces to purportedly “predict" criminality (Wu and Zhang, 2016), because an extensive body of research has disproved the pseudoscience of physiognomy and phrenology (Stark and Hutson, 2021).

Note that validity does not require all predictive factors to have a *direct* causal relationship to the target variable. For instance, race is a well-established risk factor for COVID-19 related mortality, although the causal pathways through which race and COVID-19 mortality interact are not well-understood (Tai et al., 2020; Mackey et al., 2021). One plausible pathway is that race is causally associated with access to healthcare, and access has a causal effect on health outcomes (Gray et al., 2020; Mackey et al., 2021). Given the existence of such plausible causal connection, race is often invoked as an important risk factor to weigh in allocation of COVID-19 mitigation resources (Schmidt et al., 2020; Wrigley-Field et al., 2021).

4.3.2 Target Misalignment

In practice there is often considerable misalignment between what humans intended for the algorithm to predict and what the algorithm actually predicts. These issues of construct invalidity can lead to undesirable results after deploying the predictive algorithm.

In many settings, the desired prediction target is not easily observed, and so a proxy outcome is used in its place. For the pre-trial release task in the criminal justice setting, the desired prediction target may be criminal activity, but it is not possible to directly observe all criminal activity. Instead, algorithm designers have used proxy outcomes like re-arrests or re-arrests that resulted in convictions (Fogliato et al., 2021; Bao et al., 2021). The use of proxies in this setting is particularly problematic because there are documented biases in the criminal justice system, such as racial disparities in who is likely to be arrested (Alexander, 2011). These systematic biases mean the predictions are not predicting who may commit a crime but instead are predicting who may be arrested. In healthcare contexts, medical costs are sometimes used to proxy health outcomes. However, due to racial bias in quality of healthcare, these proxies systematically underestimate the severity of outcomes for black patients (Obermeyer et al., 2019b). In other settings further complications arise when the objective of the decision making task is a function of multiple desired prediction targets. For instance, in the child welfare screening setting decision makers may want to reduce both the risk of immediate danger and the long-term risk of neglect. When the algorithm is constructed to only focus on one target, then we may suffer *omitted payoff bias* if the algorithm performs worse in practice on the combined objectives than anticipated from an evaluation on the singular objective (Kleinberg et al., 2018).

Often we only observe outcomes under the decision taken–that is, we have bandit feedback (Swaminathan and Joachims, 2015). Prediction tasks in such settings are counterfactual in nature, in the sense that we would like to predict the outcome under a proposed decision (Coston et al., 2020b). As we saw in Chapter 1, an algorithm trained to predict outcomes that were observed under historical decisions will not provide a reliable estimate of what will happen under the proposed decision if the decision affects the outcomes. Recall the child welfare screening task where the goal is to predict risk of adverse child welfare outcomes if no further action is taken (“screened out" of investigation). Investigation can impact the risk of adverse outcomes if the welfare agency is able to identify family needs and provide appropriate services. A predictive algorithm that is trained on the observed outcomes without properly accounting for the effect of investigation on the outcome will screen out families who are likely to benefit from services. When we have measured all factors jointly affecting the decision and the outcome, we can address treatment effects by training a counterfactual prediction model (Coston et al., 2020b; Schulam and Saria, 2017a). As we saw in Chapter 2, when some confounding factors are unavailable for use at prediction time, as long as we have access to the full set of confounding factors in a batch dataset available for training, then we can properly account for any treatment effects in the bandit feedback setting (Coston et al., 2020a). In settings where we have unmeasured factors in both the training and test data, we can predict bounds on the partially identified prediction target using sensitivity models (Rambachan et al., 2022).

4.3.3 Population Misalignment

Even if we can justify our choice of predictive attributes and target variable, we can still have validity issues if the dataset does not represent the target population due to selection bias or other distribution shifts. This *population misalignment* poses a threat to a valid evaluation of the predictive algorithm because performance on the dataset may not accurately reflect performance on the target population. Notably, fairness properties such as disparities in performance metrics by demographic group can be markedly different on the target population. Chapter 3 discussed this phenomenon in the context of consumer lending, showing how predictive disparities computed on the population of applicants whose loan was approved notably underestimates disparities on the full set of applicants. Similarly in the criminal justice setting, Kallus and Zhou (2018b) demonstrated that significant disparities in New York City Stop, Question, and Frisk error rates persist in the target distribution (all NYC residents) even when there are no disparities in error rates on the data sample (stopped residents). Misalignment between the model's performance during development and performance at deployment pose clear threats to predictive and external validity.

Population misalignment occurs in practice often when the dataset examples are selectively sampled (i.e., not randomly sampled) from the target population. In a number of high-stakes settings, outcomes are only observed for a selectively biased sample of the population. In consumer lending, we only observe default outcomes for applicants whose loan was approved and funded (Coston et al., 2021b). In criminal justice, we only observe re-arrest outcomes for defendants who are released (Kleinberg et al., 2018). In child welfare screening, we only observe removal from home for reports that are screened in to investigation (Chouldechova et al., 2018). A common but potentially invalid approach in such settings is to use the selectively labelled data to both train the predictive model and perform the evaluation, implicitly treating this sample as if it were a representative sample of the target when in reality it is not.

A promising strategy to address selection bias leverages unlabeled samples from the target distribution which are often already available or could be available under an improved data collection practice (Goel et al., 2021). For instance, in consumer lending the features (the application information) are available for all applicants (Coston et al., 2021b). If we believe that we have measured all factors affecting both the selection mechanism and our outcome of interest (i.e., no unmeasured confounding[1]), we can use methods for counterfactual evaluation presented in Chapter 1 to estimate the performance on the full population (including both labelled and unlabelled cases). In settings where we suspect there are unmeasured confounding factors, we can still evaluate a predictive model against the current policy if we can identify an exogenous factor (i.e., an instrumental variable) that only affects the selection mechanism and not the outcome (Lakkaraju et al., 2017; Kleinberg et al., 2018) or if we can specify assumptions that bound the amount of unmeasured confounding (Rambachan et al., 2022).

Another common mechanism under which population misalignment arises is distribution shift due to domain transfer. For example, when expanding credit access to a new international market, a company may want to transfer a model of loan default built on its customer base in one country to the new country (Coston et al., 2019). Because population demographics and other factors may differ between the two countries, the performance of the predictive model in the source country may not be a valid evaluation of the performance we would see in the new (target) country. When unlabeled data is available from the target domain, we may wish to reweigh the source data to make it “resemble" the target data. Under the assumption that there are no unmeasured confounding factors that affect both selection into the source/target domain and the likelihood of the outcome (known as *covariate shift*), we can use the likelihood ratio as weights to estimate the performance on the target population (Bickel et al., 2009; Moreno-Torres et al., 2012). We can also use the weights to reweigh the training data in order to retrain a model.

In practice and even with extreme diligence, it is generally not possible to ensure perfect population, target, and attribute alignment. For instance, nearly all prediction settings will suffer population misalignment due to temporal differences—the training data is observed in the past whereas the prediction task is in the future. A central question concerns the *degree* of this misalignment. As a first step towards characterizing this, we propose a deliberation process to identify and reflect on sources of misalignment in a given setting.

[1] Also known as covariate shift (Moreno-Torres et al., 2012)

4.4 Methodology for Deliberating the Validity of Predictive Models

We propose a series of questions centered around validity to evaluate the justified use of algorithms in a given decision-making context. We next present the top-level questions, discussing them in the context of the child welfare and criminal justice settings. We note that the questions presented in this section are intended purely to illustrate the skeleton of an artifact that is guided by our systematization of concepts from validity theory. Outside the scope of the current contribution, future work designing specific sub-questions must solicit feedback from stakeholders and practitioners to ensure the questions are accessible, comprehensible, and useful.

4.4.1 The High-level Structure of A Validity-Centered Protocol

At a high level, our proposed artifact will consist of five parts. Part 1 prompts the description of the decision-making task and constructs of interest. Part 2, 3, and 4 consists of questions assessing construct validity, internal validity, and external validity. Last but not least, part 5 attempts to contextualize validity concerns within the broader set of considerations around the use of algorithms (e.g., efficiency). In what follows, we briefly sketch each section. For illustrative purposes, we provide hypothetical responses in the child welfare screening setting.

1. Description of the decision-making task. To center the deliberation around validity, the first set of questions require the respondent to describe the key constructs of interest, including the decision making objective(s), the criteria across which the decision is made, and other decision points surrounding this task. For example, in the child welfare screening setting, the answer may be as follows: *The hotline call worker determines whether to screen in a report for investigation based on details in the caller's allegations and administrative records for all individuals associated with the report. The report should be screened in if the call worker suspects the child is in immediate danger or at risk of harm or neglect in the future. Preceding this screening decision was the decision by an individual (e.g., neighbor, mandated reporter, other family member) to report to the child welfare hotline. If a report is screened in for investigation, the next major decision point is whether to offer services to the family. A decision to screen out is successful when the child is not at risk of harm or neglect.*

2. Questions assessing construct validity: At a high level, construct validity requires understanding the constructs involved (e.g., the ideal target outcome and attributes influencing it) and the particular cause and effect relationships among them. To assess construct validity, our protocol will include questions about the following types of validity:

- **Content validity** asks whether the operationalization of each construct of interest serve as a good measure of it. One major approach to assessing content validity is to ask the opinion of experts in the relevant fields.
- **Convergent validity:** To assess convergent validity, one must assess: Is there a standard/ground-truth measure for the construct of interest? If yes, how does that correlate with the new measure on the target population?
- **Discriminant validity:** To assess discriminant validity, one must evaluate the following: Can one think of a concept that is related but theoretically different from the construct of interest? If yes, can the proposed measure distinguish between that concept and the construct of interest?

- **Predictive validity:** refers to the ability of a test to measure some event or outcome in the future. Therefore, to assess predictive validity, we need to ask: Is there high correlation between the results of the proposed measurement and a subsequent related behavior of interest?

One effective way to prompt the respondent to respond to the above questions is to consider what question(s) they would ask an oracle who could answer anything about the future. In our child welfare example, the answer here could be as follows: *We would ask whether the child will suffer harm or neglect in the next year.* Subsequent questions will refer to the outcomes identified in this question block as “oracle outcomes"–that is, the outcomes/events the respondent would like to ask an oracle to predict.

We follow the oracle question with questions about available outcomes in the data, how these available outcomes differ from the oracle outcome(s), and whether any of the previously stated goals are not addressed by the available outcome. These questions direct the respondent to consider for which segments of the population will the oracle and available outcomes be most likely to align and for which segments of the population will the available outcome likely diverge from the oracle outcome. A key question is when the available outcomes are observed. The answer to these questions may illuminate whether measurement error, bandit feedback, or other forms of missingness pertain to this outcome. An example answer in the child welfare screening context can be the following: *Available candidate outcomes in the data include re-referral to the hotline at a later point (e.g., within six months) or removal of the child from home within a timeframe (e.g., two years). Re-referral is a noisy proxy for the oracle outcome of harm/neglect because a re-referral can occur in the absence of any harm/neglect and, on the flip side, a child may be experiencing harm or neglect even when no re-referral is made. We expect on average a child that is re-referred to be more likely to experience harm/neglect than a child whose case is not re-referred. Re-referral is more likely to occur, regardless of underlying true risk of harm/neglect, for families of color and limited socioeconomic means (Eubanks, 2018; , Ed.; Roberts, 2019). Re-referral (or lack thereof) is observed for all reports, including those that are screened in and those that are screened out. By contrast, removal from home is only observed for reports that are screened in for investigation (Coston et al., 2020b).*

A subset of the construct validity questions will direct the respondent to focus on issues of bandit feedback and treatment effects. These questions ask the respondent to consider how the decision relates to the outcome, including whether the outcome is observed under all decisions and whether the decision affects the outcome (and in what ways). For example, the respondent may describe the relationship between the decision and outcome in the child welfare screening setting as follows: *The decision is whether to screen in or screen out a case for a child maltreatment investigation. The outcome that is observed for all decisions is whether the child was later re-referred to the child welfare hotline. If the case is screened in, there are additional observed outcomes: Whether the allegations are substantiated upon investigation by a caseworker, whether the family is offered support in the form of public services, and whether the child is later placed out-of-home. These outcomes are observed under screen out only when a later report is screened in for investigation. The call screener's screening decision affects the outcome. For example, the decision to screen in a case may decrease the likelihood of observing adverse outcomes if the family receives public services that lead to improved parenting practices.*

3. **Questions assessing internal validity:** At a high level, internal validity is concerned

with the existence of defensible *causal* relationship between features and the target label. To hone in on issues of internal validity, the respondent must identify available data features that one can plausibly claim are risk factors or protective factors for the ideal oracle outcome. The respondent must additionally provide evidence to support the claim that these are valid risk factors or protective factors for the oracle outcome. For instance, a respondent in the child welfare screening setting may identify the following as risk factors and protective factors in the data: *The data contains the results of any prior child welfare investigations, and we may suspect that a child in a case that was previously found to have child neglect may be at risk for future neglect. The data also contains information on how often extended members of the family (such as the grandmother) interact with or care for the child, and regular supervision from a stable guardian may mitigate risk of child harm or neglect.*

4. Questions assessing external validity: External validity is concerned with the generalizablity of the model across persons, settings, and times. The question block focusing on external validity contains questions that require the respondent to describe the population for which data is available (*training population*), including provenance, the locale and time period for which data was observed, and whether any of the observations were filtered out of the dataset (e.g., due to missing data issues). The questions similarly direct the respondent to describe the population on which the predictive algorithm will be used (*target population*), including the anticipated time frame and geographies for which the predictive algorithm will be deployed. The respondent will also be asked to specify in what ways the training population differs from the target population. In our running child welfare example, the answer may be: *The training population is all reports to the state's child welfare hotline from 2015-2020 that were recorded in the state records system. No reports were knowingly filtered out of the dataset. The target population is all reports to the state's hotline at least for the next five years. The target population likely differs from the training population because of a change in mandatory reporting in mid 2019 that expanded the definition of mandated reporter to include teachers and sports coaches. As a result, the volume of calls to the hotline increased after the policy change and likely includes some reports that would not have been made absent the policy change.*

5. Tradeoffs between validity and competing considerations: To prompt deliberation on how to weigh misalignments threatening validity against other considerations (such as efficiency or standardization), the next set of questions requires the respondent to articulate why a predictive algorithm may support decision making and to describe how they anticipate the predictive algorithm to complement the existing tools and information available. To ground this reflection in specifics, this section will ask respondents to precisely identify the expected benefits of the algorithm (e.g., improvements in efficiency or uncovering new patterns of risk). Continuing the child welfare example, the answer may be: *We intend for the predictive algorithm to summarize the information in the administrative records which the call screeners typically do not have sufficient time to fully parse. If the administrative records contain additional patterns of risk not captured in the allegations reported by the caller, then we anticipate the predictive algorithm may be able to flag reports that should be screened in but would otherwise be screened out.*

Target respondent: The respondent(s) we expect to deliberate and document answers to these questions are the individual(s) involved in the process of bringing data-driven algorithms into the decision-making process. These may include (but are not limited to) algorithm developers, data scientists and analysts, those responsible for algorithm procurement, management, frontline decision makers, and community members.

4.4.2 Protocol as a Mechanism for Transparency, Oversight, Conversation, & Translation

We next discuss how we envision a protocol reflecting the above structure, potentially in combination with questions from other existing protocols (e.g., focused around value alignment), can serve as a mechanism for transparency, oversight, conversation, and translation.

1. **Protocol as a mechanism of transparency.** A growing body of literature discusses the need to find better ways to empower impacted community members to shape algorithm design (Krafft et al., 2021a; Zhu et al., 2018; Martin Jr et al., 2020). However, community members struggle to do this without sufficient insight into the internal deliberation processes. The protocol can help lower these barriers. For example, without the protocol, community members may be limited to assessing the face validity of models. Publicly shared responses to protocol questions may extend community members' knowledge to encompass a wider range of validity issues that would otherwise be inaccessible or unknown to them.

2. **Protocol as a mechanism for oversight.** If the protocol is reviewed by an independent review board, deliberations around model validity in decision-making could be guided by standards that may reflect and align expectations across practitioners, policymakers, and community members. We draw an analogy to the research Institutional Review Board (IRB), which has a goal of "protecting [the rights and welfare of] research subjects" (for the Protection of Human Subjects of Biomedical and Research, 1978). An independent review board for this protocol could serve to protect impacted community members, as opposed to 'research subjects.'

3. **Protocol as a mechanism for conversation between multiple stakeholders.** If a diverse set of stakeholders are involved in deliberating and discussing the protocol questions, the protocol could help these conversations reach those who may not typically be involved in making model-level design decisions. For example, in some public sector agencies that use algorithmic decision support tools, frontline decision-makers, organizational leaders, and model analysts may develop beliefs and goals around the use of decision-making algorithms in silo (Kawakami et al., 2022; Saxena et al., 2021). The process of responding to the protocol questions can introduce opportunities for structured, proactive modes of interactions across workers who otherwise typically work in isolation. Engaging diverse perspectives in collaborative discussions surrounding the protocol may open opportunities for better understanding and mitigating inter-organizational value misalignments (Holten Møller et al., 2020) that would otherwise get embedded and reinforced through the model itself.

4. **Protocol as a mechanism of translation to bridge academic-practitioner divide.** Recent research suggests that many of the concepts under the purview of our envisaged protocol may be less deliberately scrutinized by practitioners developing algorithms for decision-making in the real-world (Passi and Barocas, 2019; Veale et al., 2018). The protocol may help bridge this divide between the research community and real-world practitioners. For example, this protocol could be a means for the research community to operationalize concerns related to model validity into practical questions intended to guide internal deliberation processes in real-world organizations.

4.5 Conclusion

This chapter translates theoretical validity concepts into considerations for evaluating the justified use of predictive algorithms in practice. We showed how validity is a useful lens for evaluating the use of algorithms because it foregrounds issues in problem formulation, data challenges, and latent construct operationalization that jeopardize the suitability of algorithms. We sketch a structure for a validity-centered deliberation protocol, targeted to guide multi-stakeholder conversations regarding whether or not to develop and use a predictive algorithm.

We emphasize that a validity-focused deliberation protocol is *not* sufficient on its own to justify the use of a predictive algorithm. Rather, we see the primary value of such a protocol as a means to structure and scaffold critical conversations among relevant decision-makers. Moreover, validity is just one component of evaluating the justified use of algorithms, alongside considerations related to reliability, value alignment, and beyond. Last but not least, organizations deploying algorithms should iteratively and constantly re-evaluate whether a predictive algorithm's use is justified, as the conditions for a given algorithm's justification may evolve with time.

This framework comprises a validity perspective on evaluating the justified use of data-driven decision-making algorithms. This perspective unites concepts of validity from the social sciences with data and problem formulation issues commonly encountered in machine learning and clarifies how these concepts apply to algorithmic decision making contexts. We situate the role of validity within the broader discussion of responsible use of machine learning in societally consequential domains. We illustrate how this perspective can inform and enhance future research by sketching a validity-centered artifact to promote and document deliberation on justified use.

We hope that this protocol could enable practitioners to identify and mitigate validity issues before model deployment. We envision practitioners using this protocol in conjunction with other tools designed to align values and promote equity, such as fairness checklists (Madaio et al., 2020b) and bias audits. We next consider how targeted, domain-specific analyzes of racial bias in key decision points can facilitate a context-aware approach to algorithmic auditing. We focus on decision points that determine what data is available for training predictive algorithms and for whom.

5

Assessing Racial Bias in Police Stops

Recidivism prediction instruments (RPIs) are widely used to inform judicial decisions like whether to grant defendants pre-trial release (Chouldechova, 2017). The RPIs typically take the form of statistical risk assessments that have been constructed on administrative datasets. Key decisions made by criminal justice professionals like police officers, judges, and parole boards affect who is in these datasets and what information is available about them. In this chapter, we conduct an in-depth assessment to rigorously assess one of these key decisions points – police stops – for racial bias. Racial biases, whether implicit or explicit, in officers' decision will impact the likelihood that these individuals are observed in administrative datasets later used to construct RPI's.

In this chapter, we develop a method to assess racial bias in police traffic stops that formalizes prior work by casting the question of racial bias in police stops as a *counterfactual* one: *Would the officer have made a different stop decision had they been unable to see the race of the driver?*

We formalize this question using the potential outcomes framework in § 5.2. We formulate an an ideal but infeasible experiment which we then connect to a natural experiment used by prior work that leverages variation in the time of day when the dark of night descends over the year. We propose an estimand to measure racial bias that is both interpretable as a ratio of risk ratios and identifiable as an odds ratio under sampling bias we observe in administrative police records. Our measure is fully identified under several causal assumptions that we provide in § 5.3. We next address an undesirable property of odds ratios – its purported non-collapsibility – by showing this is avoidable if we use an alternative aggregation to the standard arithmetic mean. We provide a new definition of collapsibility that makes this choice of aggregation method explicit, and we demonstrate that the odds ratio is collapsible under geometric aggregation in § 5.3.2.

We analyze the efficiency theory of our estimand in § 5.4 and use these theoretical results to develop nonparametric estimators. In particular, we propose a doubly robust-style estimation approach for the geometric aggregated odds ratio and describe conditions under which the estimator is $\sqrt{n}$-consistent and asymptotically normal in § 5.5. We additionally provide a flexible plug-in approach to estimating the covariate-conditional measure of racial bias. We use these estimators to assess racial bias in police traffic stops on the Stanford Open Policing Project data (Pierson et al., 2020). Our empirical section (§ 5.6) presents the results of our assessment for five cities in the U.S. and discusses implications for policy-making. Several contributions in this chapter were first introduced in the preprint Coston and Kennedy (2022).

The purpose of this chapter is two-fold: First, we aim to provide a rigorous, context-aware method for bias auditing in key decision points in systems where algorithms are used. Second, we hope this method also serves as example for how we can invert the usual paradigm whereby machine learning is used by those in power to assess and predict behavior of people with relatively less power (O'neil, 2016; Barabas et al., 2020). Here, we use machine learning and statistical methods to assess the behavior of those in power. Assessments of this sort can inform policy decisions around where to target police reform initiatives, and if conducted over time, they provide a way to evaluate the effectiveness of reforms.

5.1 Background and Related Work

One of the most common points of entry into the criminal justice system is motor vehicle traffic stops (Davis et al., 2018). Each year nearly 20 million drivers are stopped by the police (Baumgartner et al., 2018). Police officers have considerable discretion in whether to stop a motorist for a minor traffic violation. In this chapter, we build on a long research tradition that considers whether racial bias (implicit or explicit) plays a role in the officer's decision to stop. Prior research indicates that officers may use minor traffic violations as an excuse to search for contraband, and these "pretexual" stops are believed to disproportionately target people of color, particularly black and LatinX communities (Alexander, 2011).

In prior work, Grogger and Ridgeway (2006) posited that darkness casts a veil that makes it harder for officers to observe the driver's race. Then, if police officers are more likely to stop black drivers during daylight all else equal, one might conclude there is racial bias in traffic stops. In order to yield a binary *darkness* variable, they filter out stops between sunset and dusk when the ambient sunlight falls between light and dark. This influential paper initiated a line of work using the *veil of darkness* hypothesis to analyze for racial bias in police stops (e.g., Horrace and Rohlin, 2016; Ritter, 2017; Pierson et al., 2020).

Of particular note is work by Pierson et al. (2020), who compiled a large-scale dataset of police stops in 35 municipal police departments and 21 state patrol agencies on which they performed the *veil of darkness* test. Our empirical analysis (§ 5.6) uses this dataset. The problem formulation and estimation strategy in Pierson et al. (2020) is similar to much of the *veil of darkness* literature, modelling the odds that a stopped driver is black as log linear in the darkness random variable and the confounding factors. This design assesses for bias by testing whether the odds-ratio is significantly different from 1. They include as confounding factors the location, time of the stop, season, and whether a municipal or state police officer made the stop. In order to question whether these variables include all confounding factors and more generally to assess whether their model form is well-specified, we develop a precise definition of what the intervention and potential outcomes are in the next section.

5.2 Problem Formulation and Additional Notation

Police traffic stop data contains the date, time and location of each stop as well as information about the driver's demographics including race.[1] Let $B = \mathbb{I}\{Race = black\}$ denote whether the stopped driver was perceived as black by the officer. $X \in \mathbb{R}^d$ contains other covariates about the traffic stop: time of day, day of year, location, and whether the stop was made by municipal or state police. We also define $S = \mathbb{I}\{Driver\ stopped\}$ and potential outcome

[1] Data from Stanford Open Policing Project (Pierson et al., 2020)

$S^{T=t}$ to denote whether the driver would be stopped under intervention t. We will use the shorthand S^t to denote $S^{T=t}$. Before introducing the interventions, we discuss the sampling bias in our data.

Definition 5.2.1 (Outcome-dependent sampling)**.** Data observed under outcome-dependent sampling contains samples from the conditional distribution $(B_i, X_i) \sim P(B, X \mid S = s)$ for either $s = 1$, $s = 0$, or possibly both.

Administrative police records only contain samples from $P(B, X \mid S = 1)$. In other words, only stopped drivers are observed in the data.

To motivate our intervention, it is helpful to consider what estimands assess racial bias. Racial bias could be cast as a contrast of the race-conditional stop rates $P(S \mid B = 1)$ and $P(S \mid B = 0)$. However, there may be legitimate differences in these quantities if, for instance, driving behavior or schedules differ by race. In the ideal setting, we would ask whether an officer would make the same stopping decision if they could observe driver race as they would if they could not observe driver race. We now consider this ideal (but likely infeasible) experiment which will guide our specification of a feasible experiment.

Ideal experiment: Suppose we have virtual reality googles that can obfuscate the race of the driver. Further suppose we can randomly assign police officers to wear these googles. In this ideal experiment, the intervention variable T indicates whether the officer wore the race-obfuscating googles. If we could record data about stops as well as non-stops, then we could assess racial bias by contrasting $P(S^{T=1} \mid X, B = 1)$ with $P(S^{T=0} \mid X, B = 1)$. However, in practice it would be difficult to implement this experiment due to a variety of political, economic, and technical reasons.[2]

Feasible experiment: Daylight (or lack thereof) can serve as a proxy for race observability. Our feasible experiment defines the intervention T as the indicator for whether the stop occurred during the dark of night.

Additional notation: Denote our data samples as $Z = (B, T, X)$ where $B = 1$ if the officer reports perceiving the stopped driver as black, $T = 1$ if the stop occurred during the dark of night, and X are other covariates. For a random variable Z, we use $P_n(f(Z)) := \frac{1}{n}\sum_{i=1}^{n} f(Z_i)$. Because of the outcome-dependent sampling, we will define the nuisance functions as conditional on a stop having occurred. We define

$$\mu_t(x) := P(B = 1 \mid X = x, T = t, S = 1). \tag{5.1}$$

For clarity we will denote the propensity to assign intervention t as $\pi_t(x) := P(T = t \mid X = x, S = 1)$. As before, when the subscript is omitted, the propensity refers to the treatment $T = 1$. Our analysis will make use of the logit function where $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$.

[2] More realistically, we might be able to conduct this experiment during training sessions where some police departments are already using virtual reality googles (Apex). A key question under such a design is to what extent external validity holds. If officers would respond differently in training simulations than they would in the real-world, then findings under this design may not be informative.

5.3 Identification of the Measure of Racial Bias

In this section, we will consider how to specify and identify our target estimand to assess racial bias. As usual in causal inference, we will use assumptions to write a hypothetical measure of bias in terms of observable quantities (i.e., *causal identification*). Additionally, our identification must resolve the bias in the outcome-dependent sampling of our data (i.e., *sampling identification*).

Ideally we would specify as our target causal estimand the risk difference $\mathbb{E}[S^1 - S^0 \mid X, B = 1]$ or the risk ratio $\mathbb{E}[S^1 \mid X, B = 1]/\mathbb{E}[S^0 \mid X, B = 1]$. These estimands describe how much more likely a black driver is to be stopped under the dark of night intervention versus daylight, in an absolute or relative sense, respectively. However, because we do not observe data about those who were not stopped, we cannot identify these estimands unless we make very strong assumptions. Instead we will target a different estimand that is observable under outcome-dependent sampling. Our target causal estimand will take the form of a ratio of race-conditional risk ratios:

Definition 5.3.1 (Covariate-conditional target causal estimand)**.**

$$\psi(x) = \frac{P(S^1 = 1 \mid X = x, B = 1)}{P(S^0 = 1 \mid X = x, B = 1)} \Big/ \frac{P(S^1 = 1 \mid X = x, B = 0)}{P(S^0 = 1 \mid X = x, B = 0)}$$

.

When this value is significantly different from one, the race-conditional risk ratios are significantly different. This estimand satisfies sampling identifiability because it can alternately be expressed as the odds ratio,

$$\psi(x) = \frac{\text{Odds}(B = 1 \mid S^1 = 1, X = x)}{\text{Odds}(B = 1 \mid S^0 = 1, X = x)}$$

where notably every term is conditional on a stop having occurred. For the remainder of this chapter we will generally work with the odds ratio form of the target estimand as that is more amenable to our setting with outcome-dependent sampling. However, for substantive interpretation of results, we encourage readers to recall the ratio of risk ratios form.

5.3.1 Causal identification

Proposition 5.3.1. We can identify $\psi(x)$ as

$$\psi(x) = \frac{\text{Odds}(B = 1 \mid S = 1, T = 1, X = x)}{\text{Odds}(B = 1 \mid S = 1, T = 0, X = x)}$$

under the assumptions that

1. $T \perp B \mid S^t = 1, X$ for $t = 0, 1$
2. $P(0 < \pi(X) < 1) = 1$
3. $S = TS^1 + (1 - T)S^0$

The second and third assumptions are the usual positivity and consistency assumptions commonly made in causal inference settings. The first assumption is analogous to the usual ignorability assumption that assumes all confounding factors have been measured. However, the form of our ignorability-type assumption is non-standard since it conditions on the potential outcomes. Our condition requires that the racial composition of at-risk drivers not vary with darkenss in ways that are not explained by X. This is implied by the stronger assumption that $T \perp S^t, B \mid X$.

Remark 5.3.1. For the remainder of this chapter, we directly define $\psi(x)$ as the identified odds ratio in Proposition 5.3.1.

To recap, we show how to fully identify the covariate-conditional odds ratio $\psi(x)$. We use the odds ratio because it is identifiable under outcome-dependent sampling. The downside is the odds ratio poses challenges for aggregation, which we consider in the next section. We will propose an alternate method of aggregation to address this issue.

5.3.2 Aggregation and collapsibility

An aggregated measure can be substantively useful to summarize measures across a meaningful unit, including administrative units (e.g., police precinct or department) or temporal units (e.g., year). Aggregation is also helpful for statistical reasons because estimation is generally easier for aggregated measures. We can easily obtain Central Limit Theorem-type results for aggregated measures that describe how to construct confidence intervals. However, estimating the entire $\psi(x)$ curve is more challenging since we hit the curse of dimensionality, and therefore inference requires careful under-smoothing or bias-correction. In this sub-section we consider how to do aggregation of our target estimand.

We consider a desirable formal property for aggregation – collapsibility – that informs our choice of aggregation method. The arithmetic mean is so commonly used for aggregation that it almost seems synonymous, but to achieve collapsibility here we make the case for using the geometric mean to aggregate odds ratios.

A desirable quality for a measure of effect is that the marginal effect describes the effect for a representative unit. The property of collapsibility (Def. 5.3.2) formalizes this quality (Whittemore, 1978; Greenland et al., 1999). In this section, we take a detour to discuss collapsibility in detail.

Remark 5.3.2. In a departure from standard usage, we use the term "marginal" to generically refer to a summary measure via an aggregation method that must be explicitly specified. As an example, the "marginal odds" of B in standard usage unambiguously refers to $\frac{P(B=1)}{P(B=0)} = \frac{\mathbb{E}[P(B=1|X)]}{\mathbb{E}[P(B=0|X)]}$. However, we denote this quantity as the marginal odds *with respect to arithmetic aggregation*, differentiating it from other marginal measures such as, for example, the marginal odds with respect to geometric aggregation $\frac{\prod P(B=1|X=x)^{dP(x)}}{\prod P(B=0|X=x)^{dP(x)}}$.

Collapsibility is often discussed with respect to the arithmetic mean, under which collapsibility requires that we can specify weights for conditional effects such that the marginal effects equals their weighted average (Hernán and Robins, 2021). For instance, the risk ratio, $RR = \mathbb{E}[S^1]/\mathbb{E}[S^0]$, is collapsible with respect to the arithmetic mean with weights $\frac{P(X)}{\mathbb{E}[S^0]}\mathbb{E}[S^0 \mid X]$. The odds ratio, however, is not collapsible with respect to the arithmetic mean (Greenland et al., 1999). The population odds ratio is generally not equal to the conditional odds ratio, even if the conditional odds ratio is a constant (Coston and Kennedy, 2022). Often it is not possible to specify a weighted average of the conditional odds ratio that equals the population odds ratio.

While the odds ratio is not collapsible under the arithmetic mean, it is collapsible under the geometric mean. To demonstrate this, we introduce additional notation. Let $f(a,b)\colon \mathbb{R}^2 \mapsto \mathbb{R}$ denote an effect contrast and let $g_{w(x)}(P)\colon \mathcal{P} \mapsto \mathbb{R}$ denote a statistical functional that aggregates $X \sim P$ with weighting function $w(x)$. For example, letting $p(x)$ denote the density of random variable $X \sim P$, we describe the average risk difference (commonly referred to as average treatment effect) by specifying $g_{p(x)}(P) = \int xp(x)dx$ and $f(a,b) = a - b$. We consider aggregations that can be written as a Fréchet mean–that is, there is an associated distance function d such that

$$g_{w(x)}(P) = \arg\min_{z \in \mathcal{X}} \int_{\mathcal{X}} w(x) d^2(z,x)dx.$$

For ease of notation, we will write $g(X)$ to indicate $g(P)$ for the distribution P over X.

Definition 5.3.2. A contrast f is collapsible with respect to aggregation method g if

$$f\Big(g_{p(x)}(\mu_1(X)), g_{p(x)}(\mu_0(X))\Big) = g_{w(x)}\Big(f(\mu_1(X), \mu_0(X))\Big) \tag{5.2}$$

for weights $w(x)$ in the probability simplex and where $p(x)$ denotes the density or pmf of $X \sim P$.

The left hand side describes the marginal effect–that is, the contrast of the aggregations of $\mu_t(x)$ for $t \in \{0,1\}$. The right hand side describes a weighted aggregation of the conditional contrasts.

For example, the average risk difference is collapsible with respect to the arithmetic mean using as weights the density of x. We briefly remark on the weights $p(x)$ and $w(x)$. While $p(x) = w(x)$ for the average risk difference, this need not be the case. The risk ratio has contrast $f(a,b) = \frac{a}{b}$ and is collapsible under aggregation $g_{w(x)}(P) = \int xw(x)dx$ with $w(x) = \frac{p(x)\mathbb{E}[S^0|X=x]}{\mathbb{E}[S^0]}$.

As far as we are aware, we were the first to introduce this expanded definition of collapsibility that explicitly incorporates the aggregation method (Coston and Kennedy, 2022). This enabled us to make the novel observation that the odds ratio is collapsible under geometric aggregation

Proposition 5.3.2. The odds ratio is collapsible with respect to the geometric mean with weights $p(x)$.

If the conditional OR is a constant c, then the geometric odds ratio also equals c. Recall that this was not necessarily the case for the arithmetic odds ratio, which could take on a value other than c. Since the geometric mean exhibits the desirable property of collapsibility, we propose the geometric mean of $\psi(x)$ as the aggregated target estimand.

Definition 5.3.3 (Aggregated target estimand)**.**

$$\Psi := \prod_{x \in \mathcal{X}} \Big\{\psi(x)\Big\}^{dP(x|S=1)} \quad = \quad \exp\Big(\mathbb{E}\big[\log(\psi(X)) \mid S=1\big]\Big). \tag{5.3}$$

5.4 Efficiency Theory for our Measure of Racial Bias

In this section we provide theoretical analysis that will be informative in constructing efficient estimators for the aggregated targest estimand. Specifically, we derive the von Mises-type expansion for our aggregated target estimand. Functioning as a distributional analog to the Taylor expansion for real-valued functions, the von Mises-type expansion of a target parameter describes two key elements for efficiency theory: the influence function and a remainder term. Influence functions enable us to construct estimators with desirable properties, such as second-order bias, which can achieve fast convergence rates even in nonparametric settings. We will use the influence function presented in this section to construct our estimator in the subsequent section. The second piece, the remainder term, plays an important role in characterizing the error of such estimators (see § 5.5). In a fully nonparametric model, the singular influence function is called the efficient influence function because it characterizes the efficiency bound in a local asymptotic minimax sense. The efficient influence function is therefore instructive for constructing optimal estimators. We direct the interested reader to Bickel et al. (1993); Tsiatis (2006); Kennedy (2022); Hines et al. (2022) for more information on influence functions. We will first provide efficiency theory for $\log(\Psi)$ and subsequently provide theory for Ψ. In this section we implicitly condition on $S = 1$.

Lemma 5.4.1. We have the following von Mises expansion:

$$\begin{aligned}\log\Big(\Psi(P)\Big) &= \log\Big(\Psi(\bar{P})\Big) + \int \varphi(\bar{P})d(P-\bar{P}) + R_2^L(\bar{P},P)\\ R_2^L(\bar{P},P) &= \int \frac{(\bar{\pi}(x)-\pi(x))(\mu_1(x)-\bar{\mu}_1(x))}{(\bar{\mu}_1(x))(1-\bar{\mu}_1(x))\bar{\pi}(x)} - \frac{(\bar{\pi}(x)-\pi(x))(\mu_0(x)-\bar{\mu}_0(x))}{(\bar{\mu}_0(x))(1-\bar{\mu}_0(x))(1-\bar{\pi}(x))}\\ &\quad - \frac{(\mu_1^*(x)-1/2)(\mu_1(x)-\bar{\mu}_1(x))^2}{\mu_1^*(x)^2(1-\mu_1^*(x))^2} + \frac{(\mu_0^*(x)-1/2)(\mu_0(x)-\bar{\mu}_0(x))^2}{\mu_0^*(x)^2(1-\mu_0^*(x))^2} dP(x)\end{aligned}$$

where $\mu_t^*(x)$ lies between $\bar{\mu}_t(x)$ and $\mu_t(x)$ for $t \in \{0,1\}$ and

$$\varphi(Z;P) = \log\left(\frac{\text{odds}(\mu_1(X))}{\text{odds}(\mu_0(X))}\right) - \log(\Psi) + \frac{T(B-\mu_1(X))}{\mu_1(X)(1-\mu_1(X))\pi(X)} - \frac{(1-T)(B-\mu_0(X))}{\mu_0(X)(1-\mu_0(X))(1-\pi(X))}.$$

Since the remainder term $R_2(\bar{P},P)$ is a product of the nuisance functions, we can apply Lemma 2 of Kennedy et al. (2021) to conclude that $\log\Big(\Psi(P)\Big)$ is pathwise differentiable with efficient influence function $\varphi(z;P)$.

Proof. We write our target as

$$\log\Big(\Psi(P)\Big) = \int_X \log\left(\frac{\mu_1(x)}{1-\mu_1(x)}\right)dP(x) - \int_X \log\left(\frac{\mu_0(x)}{1-\mu_0(x)}\right)dP(x)$$

Then, let $f(\mu_t(x)) = \log\left(\frac{\mu_t(x)}{1-\mu_t(x)}\right)$ and apply Lemma 5.4.2 to each term to get the result.

□

Efficiency Theory for Ψ

Theorem 5.4.1. We have the following von Mises-type expansion of our target Ψ:

$$\Psi(P) = \Psi(\bar{P}) + \Psi(\bar{P}) \int \varphi(\bar{P}) d(P - \bar{P})) + R_2(\bar{P}, P)$$
$$\text{for} \quad R_2(\bar{P}, P) = \Psi(\bar{P}) R_2^L(\bar{P}, P) + \frac{1}{2}\Big(\log\big(\Psi(P)\big) - \log\big(\Psi(\bar{P})\big)\Big)^2 \Psi^*$$

where $\varphi(P)$ and $R_2^L(\bar{P}, P)$ are given in Lemma 5.4.1 and Ψ^* lies between $\Psi(\bar{P})$ and $\Psi(P)$.

Then, by Lemma 2 of (Kennedy et al., 2021), our target $\Psi(P)$ is pathwise differentiable with influence function $\Psi(P)\varphi(z; P)$.

Proof. For a Ψ^* that lies between $\Psi(\bar{P})$ and $\Psi(P)$, applying Taylor's Theorem yields

$$\Psi(P) = \Psi(\bar{P}) + \Psi(\bar{P})\Big(\log\big(\Psi(P)\big) - \log\big(\Psi(\bar{P})\big)\Big) + \frac{1}{2}\Big(\log\big(\Psi(P)\big) - \log\big(\Psi(\bar{P})\big)\Big)^2 \Psi^*. \tag{5.4}$$

For the first order expression $\log(\Psi(P)) - \log(\Psi(\bar{P}))$ we apply Lemma 5.4.1 to obtain

$$\log(\Psi(P)) - \log(\Psi(\bar{P})) = \int \varphi(\bar{P}) d(P - \bar{P}) + R_2^L(\bar{P}, P)$$

where each term in the $R_2^L(\bar{P}, P)$ is a second-order nuisance function error.

Substituting back into Eq. 5.4 yields

$$\Psi(P) = \Psi(\bar{P}) + \Psi(\bar{P}) \int \varphi(\bar{P}) d(P - \bar{P})) + R_2(\bar{P}, P) \quad \text{where}$$
$$R_2(\bar{P}, P) = \Psi(\bar{P}) R_2^L(\bar{P}, P) + \frac{1}{2}\Big(\log\big(\Psi(P)\big) - \log\big(\Psi(\bar{P})\big)\Big)^2 \Psi^*$$

. □

5.4.1 Auxiliary lemma used for theoretical results

Second-Order Result for Functions of Regression Functions

We give a generic result for the von Mises-type expansion of smooth functions of the regression functions that is useful for our theoretical analysis.

Lemma 5.4.2. [Generic Von Mises Expansion] For $t \in \{0, 1\}$, let $\psi_t(P) := \int f(\mu_t(x)) dP$ for any twice differentiable function f of the regression function $\mu_t(x) = P(Y = 1 \mid T = t, X)$. Then we can expand $\psi_t(P)$ as

$$\psi_t(P) = \psi_t(\bar{P}) + \int \varphi_t(\bar{P}) d(P - \bar{P}) + R_2(\bar{P}, P)$$
$$R_2(\bar{P}, P) = \int f'(\bar{\mu}_t(x)) \frac{(\bar{\pi}_t(x) - \pi_t(x))(\mu_t(x) - \bar{\mu}_t(x))}{\bar{\pi}_t(x)} + f''(\mu_t^*(x)) \frac{(\mu_t(x) - \bar{\mu}_t(x))^2}{2} dP(x)$$

where $\mu_t^*(x)$ is between $\bar{\mu}_t(x)$ and $\mu_t(x)$ and φ_t is

$$\varphi_t(Z;P) = f(\mu_t(X)) - \psi_t + f^{'}(\mu_t(X))\frac{\mathbb{I}\{T=t\}(Y-\mu_t(X))}{\pi_t(X)}$$

Since the remainder term $R_2(\bar{P}, P)$ is a product of the nuisance function errors, we can apply Lemma 2 of Kennedy et al. (2021) to conclude that $\psi_t(P)$ is pathwise differentiable with efficient influence function $\varphi_t(z;P)$.

Proof of Lemma 5.4.2

Proof. We provide the proof for $t=1$. Similar steps yield the result for $t=0$. The posited influence function of ψ_1

$$\varphi_1 = f(\mu_1(X)) - \psi_1 + f^{'}(\mu_1(X))\frac{T(Y-\mu_1(X)}{\pi(X)}$$

gives $\psi_1(P) - \psi_1(\bar{P}) - \int \varphi_1(\bar{P})d(P-\bar{P})$

$$\begin{aligned}
&= \int f(\mu_1(x)) - f(\bar{\mu}_1(x)) - f^{'}(\bar{\mu}_1(x))\frac{T(Y-\bar{\mu}_1(x))}{\bar{\pi}(x)}dP \\
&= \int f^{'}(\bar{\mu}_1(x))(\mu_1(x)-\bar{\mu}_1(x)) + \frac{f^{''}(\mu_1^*(x))(\mu_1(x)-\bar{\mu}_1(x))^2}{2} - f^{'}(\bar{\mu}_1(x))\frac{T(Y-\bar{\mu}_1(x))}{\bar{\pi}(x)}dP \\
&= \int f^{'}(\bar{\mu}_1(x))(\mu_1(x)-\bar{\mu}_1(x)) + \frac{f^{''}(\mu_1^*(x))(\mu_1(x)-\bar{\mu}_1(x))^2}{2} - f^{'}(\bar{\mu}_1(x))\frac{\pi(x)(\mu_1(x)-\bar{\mu}_1(x))}{\bar{\pi}(x)}dP(x) \\
&= \int \frac{f^{''}(\mu_1^*(x))(\mu_1(x)-\bar{\mu}_1(x))^2}{2} - f^{'}(\bar{\mu}_1(x))\frac{(\pi(x)-\bar{\pi}(x))(\mu_1(x)-\bar{\mu}_1(x))}{\bar{\pi}(x)}dP(x)
\end{aligned}$$

where the first line makes use of the fact that $\int \varphi_1(\bar{P})d\bar{P} = 0$. The second line applies a Taylor expansion with the mean-value form remainder. The third line applies iterated expectation, and the fourth line simplifies. □

5.5 Methodology for Estimating our Measure of Racial Bias

5.5.1 Methodology for estimating the covariate-conditional measure

We propose a flexible plug-in approach for estimating the covariate-conditional odds ratio $\psi(x)$ that can take advantage of modern methods. We propose a two-step approach as follows: First, we use a flexible regression method to estimate the two outcome regression functions, $\mu_1(x)$ and $\mu_0(x)$ (defined in Eq. 5.1). The second step plugs in the outcome regression estimates to obtain our estimate

$$\hat{\psi}(X) = \frac{\hat{\mu}_1(X)}{1-\hat{\mu}_1(X)} \Big/ \frac{\hat{\mu}_0(X)}{1-\hat{\mu}_0(X)}. \tag{5.5}$$

5.5.2 Aggregated measure

We now turn to how to estimate the aggregated measure Ψ. We will build up to our estimator by first proposing a bias-corrected estimator for $\log(\Psi)$ based on the efficient influence function:

$$P_n(\hat{\varphi}) := \frac{1}{n}\sum_{i=1}^{n}\hat{\varphi}(T_i, X_i, B_i) \tag{5.6}$$

where

$$\varphi(T, X, B; \mu_1, \mu_0, \pi) = \text{logit}(\mu_1(X)) - \text{logit}(\mu_0(X)) + \frac{T(B - \mu_1(X))}{\mu_1(x)(1-\mu_1(x))\pi(x)} - \frac{(1-T)(B-\mu_0(X))}{\mu_0(x)(1-\mu_0(x))\pi(x)}$$
$$\hat{\varphi}(T, X, B) = \varphi(T, X, B; \hat{\mu}_1, \hat{\mu}_0, \hat{\pi}).$$

Before providing the error analysis of our proposed estimator, we briefly discuss sample splitting. Sample splitting enables us to avoid overfitting without relying on empirical process conditions.

Definition 5.5.1. $\hat{P}$ denotes a sample that is independent of the sample denoted by P_n and that has sample size $O(n)$.

With iid data, we can obtain these independent samples simply by randomly partitioning the data into two or more folds. More generally, one can use cross-fitting, a procedure which swaps the samples and averages the results, to regain sample efficiency (Robins et al., 2008; Zheng and van der Laan, 2010; Chernozhukov et al., 2018b). For simplicity we present our analysis under single sample splitting. $\hat{P}$ is used to estimate the nuisance functions.

Theorem 5.5.1. The estimator for $\log(\Psi)$

$$\widehat{\log(\Psi)} := \frac{1}{n}\sum_{i=1}^{n}\hat{\varphi}(T_i, X_i, B_i) \tag{5.7}$$

(where φ is given in Eq. 5.6) satisfies

$$\widehat{\log(\Psi)} - \log(\Psi) = O_P\Big(\sum_{t=0}^{1}\|\hat{\pi}_t - \pi_t\|\,\|\hat{\mu}_t - \mu_t\| + \|\hat{\mu}_t - \mu_t\|^2\Big)$$
$$+ (P_n - P)\Big(\varphi(Z; P)\Big) + o_P\Big(\frac{1}{\sqrt{n}}\Big)$$

assuming the following conditions hold:

1. *Convergence in probability in $L_2(P)$ norm:* $\|\varphi - \hat{\varphi}\| = o_P(1)$.
2. *Sample splitting:* Nuisance functions $\hat{\pi}$, $\hat{\mu}_1$, and $\hat{\mu}_0$ are estimated on $\hat{P}$.

And for some $\epsilon \in (0, 1)$,

3. *Strong overlap:* $P(\epsilon < \pi(X) < 1-\epsilon) = 1$ and $P(\epsilon < \hat{\pi}(X) < 1-\epsilon) = 1$.
4. *Outcome variance:* $P(\mu_t(X)(1-\mu_t(X)) > \epsilon) = 1$ and $P(\hat{\mu}_t(X)(1-\hat{\mu}_t(X)) > \epsilon) = 1$ for $t \in \{0,1\}$.

Theorem 5.5.1 demonstrates that our proposed estimator has second-order errors in the nuisance estimation errors, yielding "doubly-fast" rates. That is, we obtain a faster rate for our estimator even when estimating the nuisance function at slower rates. For example, to obtain $n^{-1/2}$ rates for our estimator, it is sufficient to estimate the nuisance functions at $n^{-1/4}$, allowing us to use flexible machine learning methods to nonparametrically estimate the nuisance functions under smoothness or sparsity assumptions. Since our error involves squared terms, this is not the usual double-robustness property that guarantees fast rates when *either* of the propensity or regression function is estimated at fast rates.

Theorem 5.5.1 also suggests that under a nonparametric model, our estimator is locally minimax optimal if $\|\hat{\pi} - \pi\| = O_P(n^{-1/4})$ and $\|\hat{\mu}_t - \mu_t\| = o_P(n^{-1/4})$ for $t \in \{0,1\}$.

Proof. We begin by decomposing the estimator as

$$\widehat{\log(\Psi)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}_1(T_i, X_i, B_i) - \hat{\phi}_0(T_i, X_i, B_i) \tag{5.8}$$

where

$$\hat{\phi}_t(T, X, B) = \text{logit}(\hat{\mu}_t(X)) + \frac{\mathbb{I}\{T=t\}(B - \hat{\mu}_t(X))}{\hat{\mu}_t(X)(1-\hat{\mu}_t(X))\hat{\pi}_t(X)} \tag{5.9}$$

.

Recall that we can similarly decompose the target as

$$\log\Big(\Psi(P)\Big) = \mathbb{E}[\text{logit}(\mu_1(X) - \text{logit}(\mu_0(X)].$$

Then we can write the error

$$\widehat{\log(\Psi)} - \log(\Psi) = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\phi}_1(T_i, X_i, B_i) - \mathbb{E}[\text{logit}(\mu_1(X)]\right) - \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\phi}_0(T_i, X_i, B_i) - \mathbb{E}[\text{logit}(\mu_0(X)]\right) \tag{5.10}$$

Then, let $f(\mu_t(x)) = \text{logit}(\mu_t(x))$ and apply Lemma 5.5.1 for $t=0$ and $t=1$ to get the result.

□

Corollary 5.5.1. The estimator $\widehat{\log(\Psi)}$ is $\sqrt{n}$-consistent and asymptotically normal under the assumptions in Theorem 5.5.1 and the following conditions:

1. $\|\hat{\pi} - \pi\| = O_P(n^{-1/4})$
2. $\|\hat{\mu}_t - \mu_t\| = o_P(n^{-1/4})$ for $t \in \{0,1\}$

The limiting distribution is $\sqrt{n}(\widehat{\log(\Psi)} - \log(\Psi)) \rightsquigarrow \mathcal{N}\Big(0, \text{var}\Big(\varphi(X)\Big)\Big)$.

5.5.3 Estimation of Ψ

Our proposed estimator for Ψ naturally follows as

$$\hat{\Psi} = \exp\left(\widehat{\log(\Psi)}\right) \tag{5.11}$$

Corollary 5.5.2. The estimator $\hat{\Psi}$ is $\sqrt{n}$-consistent and asymptotically normal under the assumptions in Theorem 5.5.1 and in Corollary 5.5.1.

In practice we recommend first estimating $\log(\Psi)$, constructing confidence interval on the $\log$ scale, and subsequently taking the $\exp$ transform to obtain the estimate of Ψ. Because Ψ is non-negative, computing sample averages on the $\log$ scale where the measure is unbounded can provide a better asymptotic normality approximation.

5.5.4 Auxiliary Lemmas for Theoretical Error Analysis

Lemma 5.5.1. [Generic Error Term] Define our target estimad $\psi_t = \mathbb{E}[f(\mu_t(X))]$ where f is any twice differentiable function of the regression function μ_t with bounded second derivative. Then ψ_t has uncentered influence function $\phi_t(Z) = f(\mu_t(X)) + f'(\mu_t(X))\frac{\mathbb{I}\{T=t\}(Y-\mu_t(X))}{\pi_t(X)}$. Under the following conditions,

1. $\left\|\hat{\phi}_t - \phi_t\right\| = o_P(1)$
2. *Sample splitting:* $\hat{\pi}_t$ and $\hat{\mu}_t$ are estimated on $\hat{P}$.
3. $P(\pi_t(X) > \epsilon) = 1$ and $P(\hat{\pi}_t(X) > \epsilon) = 1$ for some $\epsilon > 0$

then our estimator $\hat{\psi}_t := P_n(\phi_t(Z; \hat{\mu}_t, \hat{\pi}_t))$ satisfies

$$\begin{aligned}\psi_t - \hat{\psi}_t = (P - P_n)(\phi_t(Z; P)) + O_P\Big(\|\hat{\pi}_t(X) - \pi_t(X)\|\,\|\mu_t(X) - \hat{\mu}_t(X)\| + \|\mu_t(X) - \hat{\mu}_t(X)\|^2\Big)\\ + o_P\Big(\frac{1}{\sqrt{n}}\Big).\end{aligned}$$

Proof of Lemma 5.5.1

$$\begin{aligned}\psi_t(P) - P_n(\phi_t(Z; \hat{\mu}_t, \hat{\pi}_t) = \overbrace{\psi_t(P) - P(\phi_t(Z; \hat{P}))}^{A} + \overbrace{(P - P_n)(\phi_t(Z; \hat{P}) - \phi_t(Z; P))}^{B}\\ + \overbrace{(P - P_n)(\phi_t(Z; P))}^{C}\end{aligned}$$

For term A, we apply Lemma 5.4.2:

$$\begin{aligned}&= \int f'(\hat{\mu}_t(x))\frac{(\hat{\pi}_t(x) - \pi_t(x))(\mu_t(x) - \hat{\mu}_t(x))}{\hat{\pi}_t(x)} + f''(\mu_t^*(x))\frac{(\mu_t(x) - \hat{\mu}_t(x))^2}{2}dP(x)\\ &\lesssim \int (\hat{\pi}_t(x) - \pi_t(x))(\mu_t(x) - \hat{\mu}_t(x)) + (\mu_t(x) - \hat{\mu}_t(x))^2 dP(x)\\ &\leq \|\hat{\pi}_t(x) - \pi_t(x)\|\,\|\mu_t(x) - \hat{\mu}_t(x)\| + \|\mu_t(x) - \hat{\mu}_t(x)\|^2\end{aligned}$$

The second line applies the conditions given in the lemma statement. Line 3 uses the Cauchy-Schwarz inequality.

For term B, since P_n is the empirical measure on an independent sample from $\hat{P}$, we can apply Lemma 2 of Kennedy et al. (2020) with our assumption that $\left\|\phi_t(Z;\hat{P}) - \phi_t(Z;P)\right\| = o_P(1)$:

$$(P - P_n)(\phi_t(Z;\hat{P}) - \phi_t(Z;P)) = O_P\left(\frac{\left\|(\phi_t(Z;\hat{P}) - \phi_t(Z;P)\right\|}{\sqrt{n}}\right) = o_P\left(\frac{1}{\sqrt{n}}\right).$$

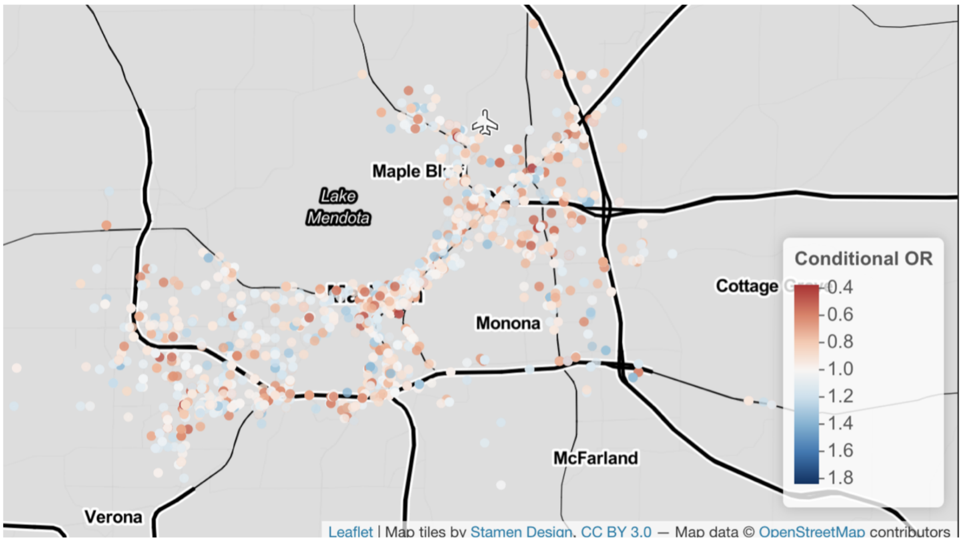
5.6 Empirical Results on Real-World Traffic Stop Data

We apply our estimators to assess racial bias in police traffic stop data from the Stanford Open Policing Project (Pierson et al., 2020). We use traffic stop data from 2010-2017 for two 60-day periods, each centered around the stop and start of daylight savings time. Following prior work, we restrict our analysis to stops in the evening (between 5-8 p.m.) and to stops where the driver was perceived as either black or white. We control for confounding factors the time of the stop, the day of the week of the stop, and the season. We used generalized additive models (GAMs) to estimate the regression nuisance functions. Specifically, we used the default hyper-parameters in the `mgcv` package in R and applied a spline to the time of stop. We used cross-fitting on $k = 10$ folds to produce the estimates of the aggregated bias, using $k - 1$ folds for nuisance function estimation and the kth fold for target parameter estimation. Restricting the analysis to stops occurring between 5-8 p.m. ensures overlaps for some geographies but not all. To address this, we filtered out stops whose estimated propensity score lied outside the range $\epsilon < \hat{\pi}(X) < 1 - \epsilon$ for $\epsilon = \frac{1}{10^6}$.

We present our estimates of the covariate-conditional measure (§ 5.5.1) for 8000 stops in Madison, Wisconsin, as a dot map in Figure 5.1. This map shows the variation in bias across districts in Madison's police department. Red suggests racial bias under the assumptions of our design; white suggests there is no effect; blue suggests that daylight would reduce the risk of a stop for black drivers more than for white drivers. We see a concentration of red stops in the downtown area as well as near the regional airport. This map could inform decisions about where to allocate resources for bias-mitigation, and the measure could be tracked over the roll-out of bias-mitigation efforts to assess their impacts. To rigorously assess the effectiveness, we recommend aggregating the measures into a meaningful unit, such as precinct or district or department, and relying on our asymptotic normality results to quantify uncertainty.

We present the aggregated measure of bias for the city of Madison as well as four other cities in Figure 5.2. These estimates give a summary statistic of our assessment of racial bias and they admit uncertainty quantification. The dots show our bias-corrected point estimates and the bars show 95% CI. Values under the horizontal dashed line indicates evidence of racial bias. This plot presents evidence of racial bias for two cities: St. Paul, Minnesota, and Fayetteville, North Carolina. While we do not see evidence of racial bias in terms of this measure for the other cities, we should not interpret this as indicating there is no such racial bias.

Interpreting a null result A null result should not be interpreted as indicating the absence of racial bias. There could still be bias that was masked by strategic behavior if officers use other visual cues from the vehicle as a proxy for race. Or, we may fail to detect bias if the dark of night is too weak of a proxy for the ideal intervention that obfuscates the driver's race. There may also be bias in upstream decisions about where the officers choose to patrol that will not be reflected in these metrics.



Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors

Figure 5.1: Real-world results of our counterfactual method to estimate racial bias in police officers' decisions to stop drivers in Madison, WI. Red suggests racial bias under our experimental design; white no effect; blue that daylight would reduce the risk of a stop for black drivers more than for white drivers.

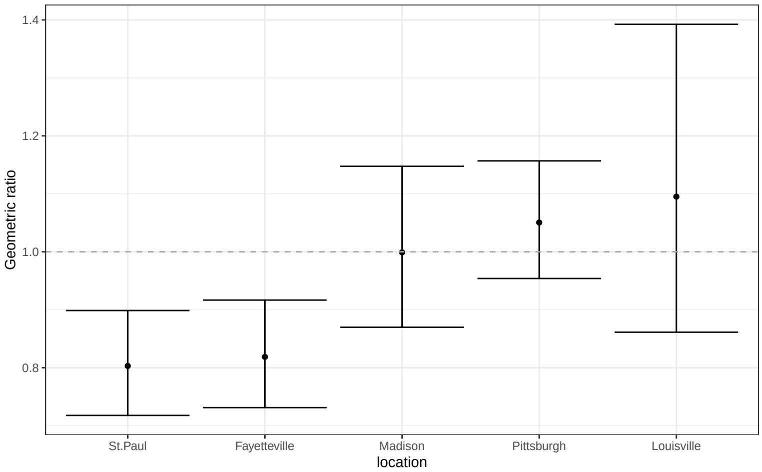


Figure 5.2: Real-world results of our aggregated measure of racial bias in the officer's decision to stop drivers for five cities. Values below the dotted line suggest racial bias. Dots indicate the point estimate of our bias-corrected method, and the bars indicate the 95% confidence intervals.

Additionally, our proposed methodology will not capture any downstream biases, such as bias in the decision to use force, make an arrest, prosecute, etc. In order to understand racial bias in the criminal justice system, empirical results from our methodology should be interpreted alongside the rich literature on bias through the criminal justice pipeline (See e.g. (Alexander, 2011; Doleac, 2021)).

5.7 Conclusion

In this chapter we developed a counterfactual formulation of the *veil of darkness* test for racial bias in police traffic stops. Our counterfactual formulation clarifies the assumptions needed to identify racial bias in officers' decisions to stop when standard measures are not identifiable due to sampling bias. We provided a flexible method to estimate an identified covariate-conditional measure of bias. Analyzing conditional measures can help direct anti-bias resources to the jurisdictions most in need. Additionally, we proposed double machine learning-style estimators that have fast rates of convergence, as the overall error is second order in the nuisance estimation error. We used these estimators to assess racial bias in a number of cities across the US. Evidence of racial bias in police stops can inform the debate on important policy questions like how to reform or rescind the use of pretextual stops and predictive policing.

Our method is also relevant to concerns raised in the growing literature on post-treatment bias in audits for discrimination in police behavior (e.g., use of force) after the stop (Knox et al., 2020; Gaebler et al., 2022). This literature alleges that standard approaches to audit for bias that consider race as the “treatment" may be unreliable in this setting because the data is conditioned on the police stop which may have been affected by race. Our findings can help contextualize this research by illuminating when and where we find evidence of such racial bias.

The methodology developed in this chapter illustrates how machine learning and statistical techniques can be used to probe biases in key decision points in the sociotechnical systems embedding algorithms. Such analyzes are important for their immediate policy implications for police reform, but also more generally because they shift the focus from those traditionally surveilled by algorithms to those in power. Finally, by assessing bias in decisions that influence administrative data, our methodology can serve as a key piece of context-aware audits on decision-making algorithms.

6

Conclusion

This dissertation proposes statistical methodologies and a deliberation framework to align decision-making algorithms toward validity and equity. We identify and address data problems and complex sociotechnical biases that challenge the validity and equity of predictive algorithms used in societally consequential decision making. Failure to recognize and address these issues can cause misalignment between an algorithm's purported purpose versus what it actually does. We showed how missing data and selection bias can threaten validity of standard approaches to model learning, evaluation, and fairness corrections and assessments. As solutions, we developed counterfactual techniques for model construction, evaluation, and fairness assessment. We demonstrated that these methods have real-world benefits by showing how our approach identifies and fixes systematic prediction mistakes in a way that standard practice fails to do on real-world child welfare screening data.

We presented the conditions required for the validity of our learning and evaluation techniques, key among them the condition that we measure all confounding factors. For the setting where some confounding factors are unavailable at prediction time, we proposed an efficient two-stage learning technique that yields valid counterfactual predictions when there is offline data describing the full set of confounders.

We proposed optimization methods to characterize the range of fairness properties over the set of similarly comparable models for both the fully supervised setting and the setting suffering from selective labels. We used this method to efficiently find a more equitable credit lending model that improves access to historically underbanked applicants while maintaining comparable accuracy to the benchmark model.

We explored issues of governance through the lens of validity, developing a framework for structuring deliberation on the justified use of algorithms in high-stakes settings that translates concepts from validity theory to the algorithmic decision-making context. We used this to structure a protocol for deliberating the validity of algorithms that we hope will be instructive in practice in governing responsible use. Broadening our view to the contexts in which algorithms are used, we demonstrated how techniques from causal inference can be used to audit for bias in key decision points. To assess racial bias in police traffic stops, we proposed counterfactual techniques that can accommodate challenging settings with outcome-dependent sampling.

A central theme throughout this work is how causal inference is useful for promoting and evaluating the responsible use of machine learning in high-stakes settings. More generally, our work makes use of methods from statistics, machine learning, and the social sciences to address key questions of responsible use, particularly around the validity, equity, and governance of decision-making algorithms. The methods we proposed are grounded in theory that provides guarantees on efficiency and accuracy.

This dissertation was guided by real-world problems in child welfare, consumer lending, and criminal justice. Throughout the dissertation we illustrated the application of our methods on real-world datasets on child welfare screening from Allegheny County, consumer lending from Commonwealth Bank of Australia, and police traffic stops from the Stanford Open Policing project.

Bibliography

Julius A Adebayo et al. *FairML: ToolBox for diagnosing bias in predictive modeling*. PhD thesis, Massachusetts Institute of Technology, 2016.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69, 2018.

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 2019. URL http://proceedings.mlr.press/v97/agarwal19d.html.

Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

Michelle Alexander. The new jim crow. *Ohio St. J. Crim. L.*, 9:7, 2011.

Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18:1–78, 2018.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. there's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016a.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23, 2016b.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Minority neighborhoods pay higher car insurance premiums than white areas with the same risk. *ProPublica*, 2017. URL https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk.

Apex. Rochester police department case study, 2022. URL https://www.apexofficer.com/case-studies/rochester-police-department.

Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1, 2019.

Rob Ashmore, Radu Calinescu, and Colin Paterson. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)*, 54(5):1–39, 2021.

Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

Australian Bureau of Statistics. Socio-economic indexes for areas (seifa) technical paper. Technical report, Department of the Treasury, 2016.

Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*, 2021.

Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 167–176, 2020.

Solon Barocas and Andrew Selbst. Big data's disparate impact. *California Law Review*, 104: 671–732, 2016a.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104: 671, 2016b.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Jo Bates, David Cameron, Alessandro Checco, Paul Clough, Frank Hopfgartner, Suvodeep Mazumdar, Laura Sbaffi, Peter Stordy, and Antonio de la Vega de León. Integrating fate/critical data studies into data science curricula: Where are we going and how do we get there? In *Conference on Fairness, Accountability, and Transparency*, FAT* '20. Association for Computing Machinery, 2020. ISBN 9781450369367. doi: 10.1145/3351095.3372832. URL https://doi.org/10.1145/3351095.3372832.

Frank R Baumgartner, Derek A Epp, and Kelsey Shoub. *Suspect citizens: What 20 million traffic stops tell us about policing and race*. Cambridge University Press, 2018.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

Kenneth A Bollen. *Structural equations with latent variables*, volume 210. John Wiley & Sons, 1989.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215, 2001.

Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 46–56. IEEE, 2019.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

Donald T Campbell. Factors relevant to the validity of experiments in social settings. *Psychological bulletin*, 54(4):297, 1957.

Donald T Campbell and Donald W Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81, 1959.

Canada, 2019. Directive on automated decision-making, 2019. URL https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015.

Lindsay Cattell, Julie Bruch, et al. Identifying students at risk using prior performance versus a machine learning algorithm. Technical report, Mathematica Policy Research, 2021.

Robert N. Charette. Michigan's midas unemployment system: Algorithm alchemy created lead, not gold, 2018. URL https://spectrum.ieee.org/michigans-midas-unemployment-system-algorithm-alchemy-that-created-lead-not-gold.

Sourav Chatterjee. Assumptionless consistency of the lasso. *arXiv preprint arXiv:1303.5817*, 2013.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and causal parameters. *The Econometrics Journal*, 2018a.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research, 2018b.

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

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. PMLR, 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd Conference on Knowledge Discovery and Data Mining*, 2017.

Amanda Coston and Edward H Kennedy. The role of the geometric mean in case-control studies. *arXiv preprint arXiv:2207.09016*, 2022.

Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 91–98, 2019.

Amanda Coston, Edward Kennedy, and Alexandra Chouldechova. Counterfactual predictions under runtime confounding. In *Advances in Neural Information Processing Systems*, 2020a.

Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 582–593, 2020b.

Amanda Coston, Neel Guha, Derek Ouyang, Lisa Lu, Alexandra Chouldechova, and Daniel E Ho. Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for covid-19 policy. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184, 2021a.

Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, pages 2144–2155. PMLR, 2021b.

Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. A validity perspective on evaluating the justified use of data-driven decision-making algorithms. In *Proceedings of the IEEE Conference on Secure and Trustworthy Machine Learning (forthcoming)*, 2023.

Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. *30th International Conference on Algorithmic Learning Theory*, pages 1–33, 2019.

1964 CRA. Civil rights act of 1964. *Title VII, Equal Employment Opportunities*, 1964.

Jonathan Crook and John Banasik. Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance*, 28(4):857–874, 2004.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women, Oct 2018. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Elizabeth Davis, Anthony Whyde, and Lynn Langton. Contacts between police and the public, 2015. *US Department of Justice Office of Justice Programs Bureau of Justice Statistics Special Report*, pages 1–33, 2018.

Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency. *arXiv preprint arXiv:1807.00905*, 2018.

Alan J Dettlaff, Stephanie L Rivaux, Donald J Baumann, John D Fluke, Joan R Rycraft, and Joyce James. Disentangling substantiation: The influence of race, income, and risk on the substantiation decision in child welfare. *Children and Youth Services Review*, 33(9): 1630–1637, 2011.

Central Digital and Data Office. Data ethics framework, June 2018. URL https://www.gov.uk/government/publications/data-ethics-framework.

Jennifer L Doleac. Racial bias in the criminal justice system. *A Modern Guide to Economics of Crime (to appear)*, 2021.

Jiayun Dong and Cynthia Rudin. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824, 2020.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

Michele Donini, Luca Oneto, Shai Ben-David, John R Shawe-Taylor, and Massimiliano A. Pontil. Empirical risk minimization under fairness constraints. In *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 2796–2806, 2018.

Ellen A Drost. Validity and reliability in social science research. *Education Research and perspectives*, 38(1):105–123, 2011.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012a.

Cynthia Dwork, Toniann Pitassi Moritz Hardt, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012b.

ECOA, 1974. Equal credit opportunity act, 1974. 15 U.S.C. § 1691.

Children's Bureau (Ed.). Child maltreatment 2017. Technical report, US Department of Health and Human Services, 2017.

Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pages 160–171. PMLR, 2018.

Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

Alessandro Fabris, Alan Mishler, Stefano Gottardi, Mattia Carletti, Matteo Daicampi, Gian Antonio Susto, and Gianmaria Silvello. Algorithmic audit of italian car insurance: Evidence of unfairness in access and pricing. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 458–468, 2021.

Qingliang Fan, Yu-Chin Hsu, Robert P Lieli, and Yichong Zhang. Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, pages 1–15, 2020.

Seena Fazel and Achim Wolf. Selecting a risk assessment tool to use in practice: a 10-point guide. *Evidence-based mental health*, 21(2):41–43, 2018.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

Andrew Guthrie Ferguson. Policing predictive policing. *Wash. UL Rev.*, 94:1109, 2016.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Technical report, arXiv preprint arXiv:1801.01489, 2019.

Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*, 2020.

Luciano Floridi and Josh Cowls. A unified framework of five principles for ai in society. In *Ethics, Governance, and Policies in Artificial Intelligence*, pages 5–17. Springer, 2021.

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.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The belmont report: Ethical principles and guidelines for the protection of human subjects of research, 1978.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Johann Gaebler, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel, and Jennifer Hill. A causal framework for observational studies of discrimination. *Statistics and Public Policy*, pages 1–61, 2022.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Darren Gergle and Desney S Tan. Experimental research in hci. In *Ways of Knowing in HCI*, pages 191–227. Springer, 2014.

Talia B Gillis. *False dreams of algorithmic fairness: The case of credit pricing*. SSRN, 2020.

Michele Gilman. Ai algorithms intended to root out welfare fraud often end up punishing the poor instead, 2020. URL https://theconversation.com/ai-algorithms-intended-to-root-out-welfare-fraud-often-end-up-punishing-the-poor-instead-131625.

Naman Goel, Alfonso Amayuelas, Amit Deshpande, and Amit Sharma. The importance of modeling data missingness in algorithmic fairness: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7564–7573, 2021.

Ilana Golbin, Anand S Rao, Ali Hadjarian, and Daniel Krittman. Responsible ai: A primer for the legal community. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 2121–2126. IEEE, 2020.

Darrell M Gray, Adjoa Anyane-Yeboa, Sophie Balzora, Rachel B Issaka, and Folasade P May. Covid-19 and the other pandemic: populations made vulnerable by systemic inequity. *Nature Reviews Gastroenterology & Hepatology*, 17(9):520–522, 2020.

Sander Greenland, Judea Pearl, and James M Robins. Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46, 1999.

Jeffrey Grogger and Greg Ridgeway. Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association*, 101(475):878–887, 2006.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Innovations in Theoretical Computer Science*, ITCS '16, page 111–122, New York, NY, USA, 2016a. Association for Computing Machinery. ISBN 9781450340571. doi: 10.1145/2840728.2840730. URL https://doi.org/10.1145/2840728.2840730.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016b.

James J. Heckman. Varieties of selection bias. *The American Economic Review*, 80(2): 313–318, 1990.

Miguel A Hernán and James M Robins. Causal inference: what if, 2021.

Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, pages 1–13, 2022.

Sarah Holland, Ahmed Hosny, and Sarah Newman. The dataset nutrition label. *Data Protection and Privacy: Data Protection and Democracy (2020)*, 1, 2020.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–16, 2019a.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019b.

Naja Holten Møller, Irina Shklovski, and Thomas T Hildebrandt. Shifting concepts of value: Designing algorithmic decision-support systems for public services. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pages 1–12, 2020.

William C Horrace and Shawn M Rohlin. How dark is dark? bright lights, big city, racial profiling. *Review of Economics and Statistics*, 98(2):226–232, 2016.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 862–870, 2019.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 560–575, 2021.

Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, United Kingdom, 2015.

Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385, 2021.

Heleen L Janssen. An approach for a fundamental rights impact assessment to automated decision-making. *International Data Privacy Law*, 2020.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, pages 9269–9279, 2018a.

Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. *arXiv preprint arXiv:1806.02887*, 2018b.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. Improving human-ai partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. *arXiv preprint arXiv:2204.02310*, 2022.

Danielle Leah Kehl and Samuel Ari Kessler. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. *Berkman Klein Center for Internet & Society*, 2017.

Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer, 2016.

Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

Edward H Kennedy, Wyndy L Wiitala, Rodney A Hayward, and Jeremy B Sussman. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. *Medical care*, 51(3):251, 2013.

Edward H Kennedy, Sivaraman Balakrishnan, Max G'Sell, et al. Sharp instruments for classifying compliers and generalizing causal effects. *Annals of Statistics*, 48(4):2008–2030, 2020.

Edward H Kennedy, Sivaraman Balakrishnan, and Larry Wasserman. Semiparametric counterfactual density estimation. *arXiv preprint arXiv:2102.12034*, 2021.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, page 247–254, 2019.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. A guide to solving social problems with machine learning, Feb 2017. URL https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293, 2018.

Dean Knox, Will Lowe, and Jonathan Mummolo. Can racial bias in policing be credibly estimated using data contaminated by post-treatment selection? *preprint*, 2020.

PM Krafft, Meg Young, Michael Katell, Jennifer E Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, et al. An action-oriented ai policy toolkit for technology audits by community advocates and activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 772–781, 2021a.

PM Krafft, Meg Young, Michael Katell, Jennifer E Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, et al. An action-oriented ai policy toolkit for technology audits by community advocates and activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 772–781, 2021b.

Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. Machine learning operations (mlops): Overview, definition, and architecture. *arXiv preprint arXiv:2205.02302*, 2022.

Amanda Kube, Sanmay Das, and Patrick J Fowler. Allocating interventions based on predicted outcomes: A case study on homelessness services. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Matt Kusner, Chris Russell, Joshua Loftus, and Ricardo Silva. Making decisions that reduce discriminatory impacts. In *International Conference on Machine Learning*, pages 3591–3600, 2019.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, 2017.

Zhiyong Li, Xinyi Hu, Ke Li, Fanyin Zhou, and Feng Shen. Inferring the outcomes of rejected loans: An application of semisupervised clustering. *Journal of the Royal Statistical Society: Series A*, 183(2):631–654, 2020.

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pages 8125–8135, 2018.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Katherine Mackey, Chelsea K Ayers, Karli K Kondo, Somnath Saha, Shailesh M Advani, Sarah Young, Hunter Spencer, Max Rusek, Johanna Anderson, Stephanie Veazie, et al. Racial and ethnic disparities in covid-19–related infections, hospitalizations, and deaths: a systematic review. *Annals of internal medicine*, 174(3):362–373, 2021.

Michael Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Prompting conversations about fairness in ai development with checklists, 2020a.

Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020b.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 349–358. ACM, 2019.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10846–10856, 2018.

Maggie Makar, Adith Swaminathan, and Emre Kıcıman. A distillation approach to data efficient individual treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4544–4551, 2019.

Rogelio A Mancisidor, Michael Kampffmeyer, Kjersti Aas, and Robert Jenssen. Deep generative models for reject inference in credit scoring. *Knowledge-Based Systems*, page 105758, 2020.

Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. Participatory problem formulation for fairer machine learning through community based system dynamics. *arXiv preprint arXiv:2005.07572*, 2020.

Charles T. Marx, Flavio du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. Technical report, arXiv preprint arXiv:1909.06677, 2019.

Marc Mauer. Justice for all-challenging racial disparities in the criminal justice system. *Hum. Rts.*, 37:14, 2010.

Celestine Mendler-Dünner, Juan C Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *arXiv preprint arXiv:2006.06887*, 2020.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 107–118, 2018.

Samuel Messick. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9):741, 1995.

Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 735–746, 2021.

Alan Mishler. Modeling risk and achieving algorithmic fairness using potential outcomes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 555–556, 2019.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019a.

Shira Mitchell, Eric Potash, and Solon Barocas. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. Technical report, arXiv Working Paper, arXiv:1811.07867, 2019b.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1): 521–530, 2012.

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.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Arvind Narayanan. How to recognize ai snake oil. *Arthur Miller Lecture on Science and Ethics*, 2019.

J Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (masters thesis); justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. excerpts english translation (reprinted). *Stat Sci*, 5:463–472, 1923.

Ha-Thu Nguyen et al. Reject inference in application scorecards: evidence from france. Technical report, University of Paris Nanterre, EconomiX, 2016.

Jum C Nunnally. *Psychometric theory 3E*. Tata McGraw-hill education, 1994.

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, 2019a.

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, 2019b.

Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

Samir Passi and Solon Barocas. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 39–48, 2019.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. A large-scale analysis of racial disparities in police stops across the united states. *Nature human behaviour*, 4(7):736–745, 2020.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481, 2020a.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 469–481, 2020b.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44, 2020.

Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. The fallacy of ai functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–972, 2022.

Ashesh Rambachan, Amanda Coston, and Edward Kennedy. Counterfactual risk assessments under unmeasured confounding. *arXiv preprint arXiv:2212.09844*, 2022.

Benjamin Recht. Machine learning has a validity problem., Mar 2022. URL http://www.argmin.net/2022/03/15/external-validity/.

Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now Institute*, pages 1–22, 2018.

Joseph A Ritter. How do police use race in traffic stops and searches? tests based on observability of race. *Journal of Economic Behavior & Organization*, 135:82–98, 2017.

Dorothy E Roberts. Digitizing the carceral state. *Harvard Law Review*, 132, 2019.

James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

James M Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

James M Robins and Andrea Rotnitzky. Inference for semiparametric models: Some questions and an answer-comments, 2001.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

Daniel Rubin and Mark J van der Laan. Extending marginal structural models through local, penalized, and additive learning. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2006.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019a.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019b.

Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1), 2020.

Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

Carl-Erik Särndal, Bengt Swensson, and Jan H Wretman. The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3):527–537, 1989.

Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. A human-centered review of algorithms used within the us child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–41, 2021.

Harald Schmidt, Lawrence O Gostin, and Michelle A Williams. Is it lawful and ethical to prioritize racial minorities for covid-19 vaccines? *Jama*, 324(20):2023–2024, 2020.

CKW Schotte, Michael Maes, Raymond Cluydts, D De Doncker, and P Cosyns. Construct validity of the beck depression inventory in a depressive population. *Journal of Affective Disorders*, 46(2):115–125, 1997.

Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pages 1697–1708, 2017a.

Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pages 1697–1708, 2017b.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, 2019.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. Technical report, arXiv preprint arXiv:1908.01755, 2020.

Vira Semenova and Victor Chernozhukov. Estimation and inference about conditional average treatment effect and other structural functions. *arXiv preprint arXiv:1702.06240*, 2017.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

Shreya Shankar and Aditya Parameswaran. Towards observability for machine learning pipelines. *arXiv preprint arXiv:2108.13557*, 2021.

Yonadav Shavit, Ben Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Hong Shen, Wesley H Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. Value cards: An educational toolkit for teaching social impacts of machine learning through deliberation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–861, 2021.

Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir, and Antoni Perelló. Value alignment: a formal approach. *arXiv preprint arXiv:2110.09240*, 2021.

Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Vernon C Smith, Adam Lange, and Daniel R Huston. Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses. *Journal of Asynchronous Learning Networks*, 16(3):51–61, 2012.

Luke Stark and Jevan Hutson. Physiognomic artificial intelligence. *Available at SSRN 3927300*, 2021.

Megan Stevenson. Assessing risk assessment in action. *Minn. L. Rev.*, 103:303, 2018.

Adarsh Subbaswamy and Suchi Saria. Counterfactual normalization: Proactively addressing dataset shift and improving reliability using causal mechanisms. *Uncertainty in Artificial Intelligence*, 2018.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. *arXiv preprint arXiv:1812.04597*, 2018.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May): 985–1005, 2007.

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, 2021.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16 (1):1731–1755, 2015.

Don Bambino Geno Tai, Aditya Shah, Chyke A Doubeni, Irene G Sia, and Mark L Wieland. The disproportionate impact of covid-19 on racial and ethnic minorities in the united states. *Clinical Infectious Diseases*, 2020:1–4, 06 2020. ISSN 1058-4838. doi: 10.1093/cid/ciaa815.

Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 303–310, 2018.

Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

Administration U.S. Department of Health & Human Services. Child maltreatment 2017., 2019. URL https://www.acf.hhs.gov/cb/research-data-technology/statistics-research/child-maltreatment.

Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. Developing predictive models to support child maltreatment hotline screening decisions: Allegheny county methodology and implementation. *Center for Social data Analytics*, 2017.

Rhema Vaithianathan, Emily Putnam-Hornstein, Alexandra Chouldechova, Diana Benavides-Prado, and Rachel Berger. Hospital injury encounters of children identified by a predictive risk model for screening child maltreatment referrals: evidence from the allegheny family screening tool. *JAMA pediatrics*, 174(11):e202770–e202770, 2020.

Mark J Van Der Laan and Sandrine Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2003.

Mark J van der Laan and Alexander R Luedtke. Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2014.

Mark J Van der Laan, MJ Laan, and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

Neil Vigdor. Apple card investigated after gender discrimination complaints, Nov 2019. URL https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.

Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

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.

Yixin Wang, Dhanya Sridhar, and David M Blei. Equal opportunity and affirmative action via counterfactual predictions. *arXiv preprint arXiv:1905.10870*, 2019.

Alice S Whittemore. Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(3):328–340, 1978.

Elizabeth Wrigley-Field, Mathew V Kiang, Alicia R Riley, Magali Barbieri, Yea-Hung Chen, Kate A Duchowny, Ellicott C Matthay, David Van Riper, Kirrthana Jegathesan, Kirsten Bibbins-Domingo, et al. Geographically targeted covid-19 vaccination is more equitable and averts more deaths than age-based thresholds alone. *Science advances*, 7(40):eabj2099, 2021.

Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*, pages 4038–4052, 2016.

Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 ACM designing interactive systems conference*, pages 1245–1257, 2020.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

Matei Zaharia, Andrew Chen, Aaron Davidson, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, et al. Accelerating the machine learning lifecycle with mlflow. *IEEE Data Eng. Bull.*, 41(4):39–45, 2018.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pages 325–333, 2013.

Guoping Zeng and Qi Zhao. A rule of thumb for reject inference in credit scoring. *Math. Finance Lett.*, 2014:Article–ID, 2014.

Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3): 689–722, 2017.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, pages 3671–3681, 2018a.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making – the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018b.

Wenjing Zheng and Mark J van der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, 2010.

Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–23, 2018.

Michael Zimmert and Michael Lechner. Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv preprint arXiv:1908.08779*, 2019.