Outcome Observability in Electronic Health Record-Based Clinical Prediction Models

by

Mengying Yan

Department of Biostatistics & Bioinformatics
Duke University

Defense Date: March 18th, 2025

Approved:

Benjamin Goldstein, Advisor

Matthew Engelhard

Hwanhee Hong

Michael Pencina

Michael Cary

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics & Bioinformatics in The Graduate School of Duke University
2025

ABSTRACT

Outcome Observability in Electronic Health Record-Based Clinical Prediction Models

by

Mengying Yan

Department of Biostatistics & Bioinformatics
Duke University

Defense Date: March 18th, 2025

Approved:

Benjamin Goldstein, Advisor

Matthew Engelhard

Hwanhee Hong

Michael Pencina

Michael Cary

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics & Bioinformatics in The Graduate School of
Duke University
2025

Abstract

Electronic health records (EHRs) are a rich, real-world data source widely used to develop clinical prediction models (CPMs). However, outcomes in EHR data are often not fully observed, as individuals can receive care at multiple institutions or otherwise fall outside the system's observational reach. This *observability* problem, particularly when it differs by demographic subgroups (*differential observability*), can introduce a critical source of bias in EHR-based CPMs. As a result, models may systematically underestimate risk for vulnerable groups, reinforcing existing health inequities.

While observability is a challenge across all EHR-based studies, this dissertation focuses on outcome observability within CPM development. Specifically, it addresses three key questions: how differential observability can bias CPMs, how to estimate the degree of observability, and how to build robust CPMs despite incomplete outcome information. First, after formally defining observability, we demonstrate how differential observability induces algorithmic bias in CPMs. Next, we propose a novel method to estimate and assess the extent of observability using a fully observed external data. By reweighting the external data to resemble the target EHR population, the approach provides estimates of both overall and differential observability, without requiring direct patient-level linkage. Finally, we address the challenge of constructing CPMs for long-term outcomes with a limited observing window. We frame this as a *positive–unlabeled (PU)* problem, and employ an adversarial domain adaptation method that aligns historical, fully labeled data with more contemporary, partially labeled target cohort to maintain predictive performance despite shifting patient populations.

This dissertation underscores the importance of accounting for observability when leveraging EHR data for CPMs. By defining and measuring outcome observability, as well as adapting modeling strategies to mitigate its impact, this dissertation provides a comprehensive framework for handling observability in EHR data when building clinical prediction models.

Contents

List of Tables

List of Figures

Acknowledgements

I would first like to express my deepest gratitude to my advisor, Ben Goldstein, for his unwavering support, guidance, and mentorship throughout my doctoral journey. His insightful feedback and encouragement have been invaluable in shaping this dissertation. I am also grateful to the members of my dissertation committee—Matt Engelhard, Hwanhee Hong, Michael Pencina, and Michael Cary—for their time, expertise, and constructive suggestions, which greatly strengthened this work.

I want to acknowledge the many individuals who have contributed to my research experience. My sincere thanks go to Chuan Hong, Molei Liu, Brook Alhanti, Siqi Li, and Jonathan Wilson, whose collaboration helped me overcome challenges and explore new areas of research beyond this dissertation.

To my peers in the Biostatistics and Bioinformatics Department, thank you for fostering a supportive and stimulating environment. I am especially grateful to Zigui Wang, Mian Wei, and Boyao Li, whose camaraderie—often shared through board games—brought me light during my most difficult times. I also want to thank Elaona Lemoto, Monique Sparkman, and Achal Awasthi for their social support, both in the department and at conferences, which kept me going. In addition, I wish to recognize Sean O'Brien, Michelle Evans, Tasha Allison, Laura Coutts, and Kendall Mincey for their assistance and encouragement. I am also thankful to the fellows and staff of the AI Health Fellowship Program—Elliot Hill, Scott Sun, Angel Huang, Aditya Parekh, Christine Shen, Riddhiman Bhattacharya, Amanda Bruker, and Andrew Olson—for broadening my perspective with their expertise.

I would like to extend my deepest gratitude to my friends who supported me outside the academic setting. Liangying Chen, my best friend and accountability partner: thank you for studying with me via Zoom all the way from Paris and for sharing both of our PhD journeys—exchanging postcards provided a welcome bright spot in challenging times. I'm glad we were able to explore this journey together. Yao Yang, thank you for always encouraging me and bringing me positive energy. Mingyi Chen, thank you for keeping me company through the pandemic and beyond. Meng Xia, I am grateful for your honest discussions about research questions and life plans. Demy Wang, thank you for going ice skating with me and for inviting me to your wedding. Yanbo Li, I appreciate our friendship, our chats, and sharing both joys and struggles. Sijie Ren, traveling all the way from China to visit me was incredibly meaningful, and I deeply appreciate it. Yukun Zhan, our two decades of friendship have been an incredible gift—thank you for always being there for me.

Finally, to my family, I cannot thank you enough. My parents have always believed in me, no matter which path I chose, and I know they would be there to catch me if I fell. To my partner, Joe, thank you for supporting me in every way possible—your understanding, patience, and unwavering confidence allowed me to pursue my passion wholeheartedly. I also want to pay tribute to the memory of my two grandpas, who could not see me complete this journey but whose love and values continue to guide me.

Thank you, all of you, for helping me reach this significant milestone. This dissertation would not have been possible without your support, wisdom, and understanding.

1. Introduction

Electronic health records (EHRs) are an important data source for healthcare research, as they contain detailed patient-level information—demographics, diagnoses, laboratory results, and medication histories—that may not be accessible elsewhere (Goldstein, Navar, et al., 2017). Collecting data specifically for clinical research can be expensive and time-consuming, whereas utilizing EHRs allows researchers to access extensive patient cohorts and perform large scale retrospective analyses.

A central application of EHR data is the development of clinical prediction models (CPMs), which estimate the likelihood of diagnostic or prognostic events, such as mortality or hospital readmission, and inform patient-specific interventions (Steyerberg, 2009). CPMs derived from EHR data increasingly support clinical decision-making in modern health systems (Alexiuk et al., 2024; Goldstein, Navar, et al., 2017; Sutton et al., 2020). Many institutions employ EHR-based tools that identify high-risk patients for early follow-up or more intensive management, thereby improving patient outcomes and reducing costs. For example, a CPM for predicting unplanned hospital readmission has been implemented in Duke University Health System (DUHS). This model facilitates targeted interventions including case management, medication reconciliation and post discharge follow-up, aiming to reduce the risk of readmission (Gallagher et al., 2020). Another example is the early warning score (EWS) for patient deterioration, which alerts clinicians to potential adverse events and facilitates timely interventions (O'Brien et al., 2020; Smith et al., 2013). These implementations underscore the potential impact of EHR-based CPMs to improve patient outcomes and healthcare delivery.

Despite the abundance of EHR data, limitations arise because EHRs were not designed primarily for research. EHRs only capture patient data from interactions within a health system related to specific encounters, and those interactions tend to be non-random. This phenomenon, known as informed presence, has been shown to bias association estimates in multiple studies (Goldstein et al., 2016, 2019; McCulloch et al., 2016; McGee et al., 2022; Phelan et al., 2017). Fewer studies, have focus on informed presence bias in the context of CPMs (Sisk et al., 2021), where the aim is to achieve good prediction instead of estimate effects. These studies only focus on predictors rather than the presence or absence of outcomes.

In the dissertation, we define the degree to which true patient characteristics are captured in EHR as **observability**. While observability is a challenge across all EHR-based studies, this dissertation focuses on outcome observability within CPM development. For example, in predicting hospital readmission, a patient who is re-hospitalized elsewhere would not appear as having a readmission in the local EHR, even though one occurred. This can lead to informed presence bias into CPM, as the recorded outcomes become systematically skewed by which patients seek or receive care in the observed system.

Observability issues can arise in many ways: patients may relocate, seek care at multiple hospitals, or have outcome that fall outside of a given observation window. **Differential observability** occurs when observability varies across population subgroups. For instance, patients living farther away or from specific racial/ethnic backgrounds may be systematically less likely to have their readmissions or other outcomes recorded. A naive analysis of single-center EHR data may underestimate

outcome rates in already high-risk groups. When the data with observability issues feed into CPMs, the models can underestimate risk for these groups, potentially induce algorithmic bias and exacerbating existing health inequities (Mitchell et al., 2021; Obermeyer et al., 2019).

Although concerns about algorithmic bias in clinical prediction models are well recognized (Cary et al., 2023), the particular role of observability-induced bias has received relatively little direct attention. Within the machine learning life cycle framework introduced by Suresh & Guttag (2021), observability can be seen as a measurement bias as the unobserved outcome may not be measured accurately. It can also be seen as historical bias as the data do not reflect a patient's true current status. While observability can be framed in many ways, this dissertation addresses outcome observability as a distinct source of bias. We formally define outcome observability, illustrate how differential observability can induce differential bias in CPMs, and highlight its significance for clinical risk prediction.

Related concepts exist in other fields. For instance, missing data (Rubin, 1976) and observability are interrelated but differ in focus: missing data concerns how values are absent, while observability focuses on whether an event is truly captured at all. As Weiskopf et al. (2013) emphasize, having complete EHR records does not guarantee a patient's outcome is fully observed. In the typical missing-data paradigm, we often know that a value is missing, whereas in the observability scenario, a non-recorded outcome may reflect either a true negative or a missed positive event. From a missing-data mechanisms perspective (Rubin, 1976), observability can also be viewed as *missing not*

at random (MNAR), since whether an outcome is missing depends on the underlying status (positive vs. negative).

Observability further relates to measurement error in epidemiology, where a true response is measured with some degree of error (Carroll et al., 2006). In particular, if the variable is categorical, measurement error is also referred to as misclassification. Observability can be viewed as a form of misclassification problem unique to EHR data: true positives may appear as negatives (i.e., unobserved), but true negatives typically remain negative. McGee (2022) has discussed how informed presence bias occur as a form of misclassification in association studies. While many studies have addressed measurement error in predictors, fewer have explored measurement error for outcome variables. Even those that do require external validation sets with both the true and misclassified outcomes, and are not designed to estimate the degree of error itself. (Ackerman et al., 2019; Siddique et al., 2019).

Within the machine learning literature, binary outcome observability has been framed as a positive and unlabeled (PU) problem—i.e., only positives (true events) are fully labeled, whereas negatives may be genuine or merely unobserved (Bekker & Davis, 2020; Elkan & Noto, 2008). Although existing PU-learning approaches have been widely studied, many assume *Selected Completely At Random (SCAR)*—meaning that missing outcomes do not depend on individual characteristics. In contrast, *differential observability* aligns more closely with *Selection At Random (SAR)*, for which it is harder to estimate labeling probabilities. Some recent work has explored ways to estimate labeling probabilities (Bekker et al., 2020; Gerych et al., 2022), but these often entail strong assumptions that may not hold in real-world EHR data.

After identifying observability as a source of bias, the next step is to quantify the degree of outcome observability. We propose a novel method to estimate the degree of observability by comparing EHR-based outcome to those in a fully observed external data source without requiring patient-level linkage. By reweighting the external data to match the EHR population, we derive estimates of both overall observability and differential observability across subgroups.

Finally, recognizing that EHR datasets only contain partially observed outcomes, we propose a framework to build CPMs despite outcome observability. We focus on the common scenario of training models for *long-term outcomes* when not enough follow-up time has elapsed to observe all events—yet another manifestation of the observability problem. By aligning a historical, fully labeled cohort with a contemporary, partially labeled cohort using adversarial domain adaptation that accounts for PU labels, we can preserve model performance even as the population shifts over time.

In summary, defining and characterizing observability, estimating its extent, and accounting for it in CPM development are crucial steps toward mitigating bias and improving the reliability of EHR-based predictive modeling. The work presented in this dissertation addresses three core research questions:

1. How does differential observability induce algorithmic bias in EHR-based CPMs?
2. How to estimate and evaluate the degree of observability in a EHR dataset using an external data?
3. What modeling strategies can account for observability to produce robust predictions?

Chapter 2 explores the concept of observability and differential observability and illustrates how it can induce algorithmic bias. Chapter 3 proposes a novel approach to estimate EHR data observability using external data. Chapter 4 proposes and evaluate an adversarial domain adaptation strategy for CPMs under positive-unlabeled conditions, focusing on longer-term outcomes. Finally, Chapter 5 synthesizes the findings, discusses their implications for health informatics, and suggests future directions.

2. Observability and its Impact on Differential Bias for Clinical Prediction Models[1]

2.1 Introduction

Electronic health records (EHRs) are a key data source for developing clinical prediction models (CPMs) (Goldstein, Navar, et al., 2017)**.** However, one challenge of EHR data is that there is incomplete observability of patient outcomes, where observability refers to the degree to which patient characteristics are captured, i.e., observed, within the EHR. Often, what we observe within the EHR is informative, leading to information-based biases collectively referred to as *informed presence bias* (Goldstein et al., 2016; Phelan et al., 2017; Sisk et al., 2021). For example, when studying hospital readmissions, the lack of an observed readmission does not mean that one did not occur. It just means that a person did not readmit to the medical center under study. Here, we are interested in the scenario when the limited observability is different across some factor, e.g., if the capture of readmission differs based on the distance one lives from the medical center. We call this *differential observability*. Typically, we are concerned about differential observability across demographic factors, however, differential observability can occur across any patient-level factor. Herein, we refer to these factors as *sensitive variables* (Boeschoten et al., 2021; Veale & Binns, 2017), though note others have used other terms like *protected attributes* as well (Galhotra et al., 2020; Mitchell et al., 2021). Our concern in this paper is how differential observability can induce biases in the context of CPMs. To date, most studies of EHR based biases have focused on association studies, though

---

[1] This chapter is a published paper: Yan, M., Pencina, M. J., Boulware, L. E., & Goldstein, B. A. (2022). Observability and its impact on differential bias for clinical prediction models. Journal of the American Medical Informatics Association, 29(5), 937–943. https://doi.org/10.1093/jamia/ocac019

Sisk et al.(2021) showed how informed presence can also influence CPMs. In this paper, we further raise the understanding of informed presence, especially the differential observability problem, from a prediction model perspective.

CPMs estimate the probability of a diagnostic or prognostic outcome and provide evidence-based support for clinical decision-making (Steyerberg, 2009). Attention has been paid to algorithmic fairness in prediction models (Mitchell et al., 2021; Obermeyer et al., 2019). When algorithm performance differs across demographic factors (e.g., race, sex, age), this can create and perpetuate health inequities, leading to algorithmic unfairness. We posit that using EHR data with differential observability to develop CPMs can induce such biases. Often these differences are due to structural inequities in how care is provisioned. For example:

- *Health system utilization*. As in the readmission example above, people who live further from the health system may be less likely to readmit to the same hospital. Additionally, multiple studies have detailed racial and ethnic differences in the usage and access to emergency department resources (Brown et al., 2012; Heron et al., 2006; Snowden et al., 2009). If resource utilization serves as an indicator of health status, then there will be differential observability of different groups based on how they seek care.
- *Receipt of differential care*. Inequities within the provision of health care can impact observability. Females are underdiagnosed with myocardial infarction and are less likely to be given proper treatments compared to males. Shah et al. (2015) and Hanchate et al. (2019) demonstrated differences in the types of hospitals patients are transferred to based on one's race/ethnicity. Such inequities in care

and documentation can lead to the under and over-estimation of risk for different groups.

- D*ifferential performance of diagnostic tools*. Pulse oximetry has been found to perform differentially in Black versus White patients (Sjoding et al., 2020). Similarly, breast density can impact the quality of mammograms in detecting breast cancer (Wanders et al., 2017), potentially differentially impacting Asian women (Maskarinec et al., 2001).

In response to these concerns, recent work has questioned how best to handle sensitive variables. In particular, multiple authors have suggested that a better understanding of the relationship between factors such as race and the outcome is needed to decide if it should be incorporated into CPMs (Anderson et al., 2021; Paulus et al., 2018; Vyas et al., 2020). Therefore, before identifying solutions to such biases, it is important to better understand the mechanism under which such biases arise.

In this paper, we seek to formally define algorithmic bias within the context of CPMs and show how it arises. Specifically, we propose that there is a "true" outcome that may not be fully observed within EHR data. If that observability is differential across a sensitive variable, differential bias can ensue. We use simulation methods along with a real data example to show how this differential bias arises and when the inclusion of a sensitive variable can either exacerbate or alleviate such biases.

2.2 Materials and Methods

2.2.1 Definition of Differential Bias

Observing an outcome of interest (e.g. readmissions) within an EHR involves a two-stage process (Figure 2.1). In the first stage, the "true", potentially unobserved, event

occurs and in the second stage, there is the observation, potentially incomplete, of the event in the EHR. For example, if someone has a readmission, but readmits to a different hospital, that information will not be observed within the EHR.

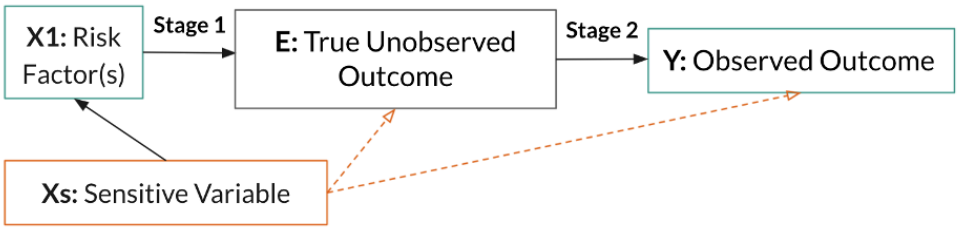

Figure 2.1: Diagram illustrating the differential observing process by sensitive variable X_s

We posit that one mechanism for differential bias is when the observability of such an event differs across a sensitive variable. For example, people who live further away from the hospital under study are less likely to readmit to the same hospital. Therefore, within the data, people who live further away will have lower readmission rates.

A sensitive variable can impact either one of the stages in Figure 2.1 producing different potentials for bias. It can also be correlated with other risk factors. The *differential bias* ($\Delta Bias$) is induced when a model differentially under or over predicts risk for a particular group. We start by considering the statistical definition of bias as the difference between a parameter (i.e., truth) and its estimate. Suppose X_1 is any risk factor for the outcome, X_s is a sensitive variable, E is the true unobserved binary event, and Y is the observed binary outcome. For the previous example, X_s is age, X_1 can be some preexistent health conditions, E=1 is readmission and Y=1 is observed readmission.

The overall bias can be defined as:

$$Overall\ Bias = mean\big[P(E = 1) - \hat{P}(Y = 1)\big] \quad (1)$$

When X_s is binary, the bias within each group is $mean\big[P(E = 1|X_s = 0) - \hat{P}(Y = 1|X_s = 0)\big]$ for group 0 or $mean\big[P(E = 1|X_s = 1) - \hat{P}(Y = 1|X_s = 1)\big]$ for group 1. Based on this definition of subgroup bias, we further define differential bias as the difference between subgroup biases:

$$\Delta Bias = mean\big[P(E = 1|X_s = 1) - \hat{P}(Y = 1|X_s = 1)\big] - mean\big[P(E = 1|X_s = 0) - \hat{P}(Y = 1|X_s = 0)\big] \quad (2)$$

When X_g is continuous, such as age, then we can define differential bias as a difference in biases between two realizations of X_s:

$$\Delta Bias_d = mean\big[P(E = 1|X_s = x + d) - \hat{P}(Y = 1|X_s = x + d)\big] - mean\big[P(E = 1|X_s = x) - \hat{P}(Y = 1|X_s = x)\big] \quad (3)$$

We note a couple of points. First, while it is often of interest to consider differential bias, we emphasize that it is equally important to consider the bias within each sub-group as well. An algorithm can exhibit no differential bias, but still be biased within each subgroup. Second, while the proposed definition is related to miss-calibration, calibration is based on observed factors (outcome and estimated risk), while here we are considering unobserved factors (true and estimated risk).

2.2.2 Simulation Study Framework

To understand the process outlined in Figure 2.1 we conduct a comprehensive simulation study. We have the two stages as:

Stage 1: $logit\big(P(E=1)\big) = \beta_0 + \beta_1 X_1 + \beta_s X_s$

Stage 2: $logit\big(P(Y|E=1)\big) = \gamma_0 + \gamma_s X_s$

Here, the sensitive variable affects the true risk through β_s in stage 1, and affects the probability of observing the outcome through γ_s in stage 2. γ_0 and γ_s together determine the probabilities of observing the outcome. As γ_0 increases, so does the full observability; as $|\gamma_s|$ increases, so does the differential observability.

We start by simulating stage 1, with a binary variable X_s, and a continuous risk factor X_1, where $X_1 \sim N(0,1) + \mu X_s$. In this way, different μ corresponds to different correlation between X_1 and X_s. The true outcome E and the observed outcome Y are generated for different scenarios by varying parameters. For each setting, we learned two CPMs: one with the sensitive variable and one without. We then used the fitted models to predict the risk using new data generated by the same process. The steps of this **general simulation framework** are as follows:

Input: $\beta_0, \beta_1, \beta_s, \gamma_0, \gamma_s$

1. *Generate* X_s *from Bin(1,0.5) and* X_1 *from* $Normal(0,1) + \mu X_s$ *(sample size n=1000).*
2. *Set* $\beta_0, \beta_1, \beta_s$ *and generate true outcome E and true risk P(E=1) using stage 1 logistic model.*
3. *Set* γ_0, γ_s *and generate observed outcome Y using stage 2 logistic model if E=1. Otherwise, Y=0.*

We repeated this process over a grid of parameter values, 1000 times for each setting. For each data set we used the observed Y to train two CPMs: one with just X_1 and with both X_1 and X_s. We then generate a "test" data set under the same initial parameters and compared the predicted risk to the true risk.

For simplicity we first consider when $\mu = 0$, that is, X_1 and X_s are uncorrelated. We then extend the simulations to the cases when X_1 and X_s are correlated (Appendix A1) and when X_s is continuous (Appendix A2).

2.2.3 Simulation Evaluation

We performed 6 sets of simulations (Table 2.1Table 2.1: Simulation Table for binary $\mathbf{X_s}$). The first two simulations are examples of the simple cases where the sensitive variable only impacts either the true outcome ($\beta_s \neq 0$ & $\gamma_s = 0$), i.e. there is no differential observability issue, or only impacts observing the outcome ($\beta_s = 0$ & $\gamma_s \neq 0$). To observe the differential bias, we compared predicted risk to the true risk for each group in each simulation and visualized the results using true versus predicted risk plots.

Table 2.1: Simulation Table for binary $\mathbf{X_s}$

Simulation	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_s}$	$\boldsymbol{\gamma_0}$	$\boldsymbol{\gamma_s}$
1	0	2	1	2	0
2	0	2	0	10	-11
3	0	2	(-3, 3)*	0 ($p_0 = 0.5$)	(-2.5, 3.5)*
4	0	2	(-3, 3)*	1 ($p_0 = 0.73$)	(-3, 2.5)*
5	0	2	(-3, 3)*	2 ($p_0 = 0.88$)	(-5, 2)*
6	0	2	(-3, 3)*	10 ($p_0 = 1$)	(-12, 0)*

*** Simulation parameter was a sequence with steps of 0.25**

Simulations 3 to 6 investigate the more complex scenarios where the sensitive variable affects both the true outcome and observability. We compare the absolute values of $\Delta Bias$ (see equation(2)) including or excluding the sensitive variable X_s under the same parameter settings, i.e., $|\Delta_{in}Bias| - |\Delta_{ex}Bias|$. We fix $\beta_0, \beta_1, \gamma_0$ and vary β_s, γ_s for different magnitudes and directions of stage 1 and stage 2 sensitive variable effects. For each set of simulations, we follow the procedure below.

Set $\beta_0, \beta_1, \& \gamma_0$. For each β_s and γ_s:

1. *Generate the data and obtain the prediction results from the general simulation framework.*
2. *Compute $\Delta_{in}Bias$ and $\Delta_{ex}Bias$.*
3. *Repeat 1000 times and take the mean of $\Delta_{in}Bias$ and $\Delta_{ex}Bias$.*
4. *Compute $|\Delta_{in}Bias| - |\Delta_{ex}Bias|$.*

We generate a heatmap for each simulation (3 to 6) to visualize the changes of $|\Delta_{in}Bias| - |\Delta_{ex}Bias|$ while changing the impacts of X_s on the true outcome and observing the outcome. We report empirical 95% confidence intervals.

2.2.4 Real Data Example

To illustrate this we consider predicting 30-day hospital readmission. We identified all discharges from 2018 – 2019 from one of the hospitals within our health system. We developed a (simple) CPM for 30-day readmission using patient age, length of stay, admission source and five comorbidities. Additionally, we categorized each patient as being local (within the hospital's county) or non-local. We propose that *Local* does not affect 30-day readmission, but the observability of it. We compared model performance with and without this sensitive variable.

All analyses were performed in R version 3.6.3. The simulation code is available in the technical appendix. The data analysis was declared exempt by our institution's IRB.

2.3 Results

2.3.1 Sensitive Variable Impacting *either* the Outcome or Observability

The calibration plots (in probability scale) of the two simple cases where the sensitive variable X_s only impacts either the true outcome or observing the outcome (simulation 1 and 2) are shown in Figure 2.2 and Figure 2.3. The red, green and blue curves are the mean curves for group 0, group 1 and overall, respectively. In Figure 2.2, the data does not have differential observability issues by design. The sensitive variable X_s impacts the true risk of the outcome (stage 1) but not the observability (stage 2). Here, ignoring X_s in the CPM (2a) leads to differential bias as the red and green curves are farther apart, while including X_s leads to less differential bias (2b, two curves overlap). Conversely, in Figure 2.3, the sensitive variable X_s does not impact stage 1 but does impact stage 2 by design. In this case, differential observability issue occurs. Leaving X_s out of the CPM results in less differential bias (3a) where including X_s in the model does induce differential bias (3b).

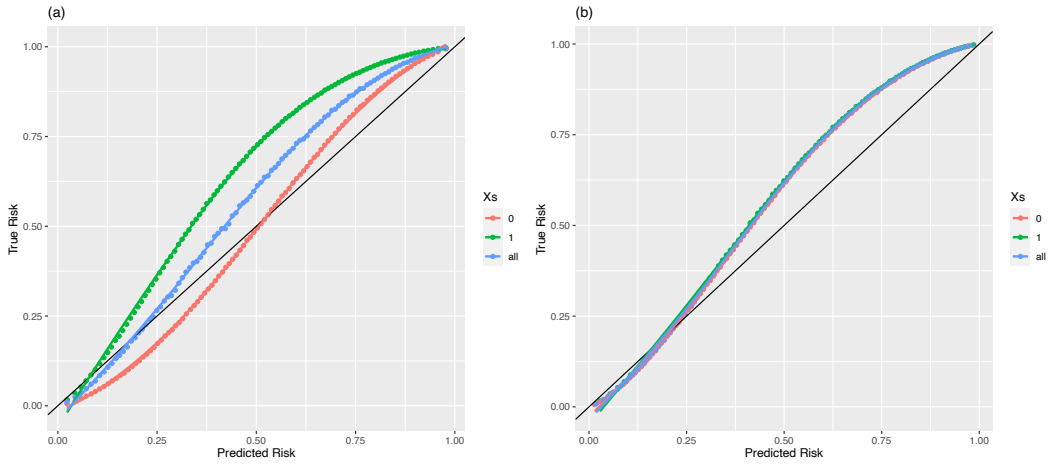

Figure 2.2: True versus predicted risk by the sensitive variable when it only impacts the true outcome

($\beta_s = 1, \beta_0 = 0, \beta_1 = 2, \gamma_s = 0, \gamma_0 = 2$), with no differential observability issue. Left (2a): Excluding sensitive variable in model. Right (2b): Including sensitive variable in model.

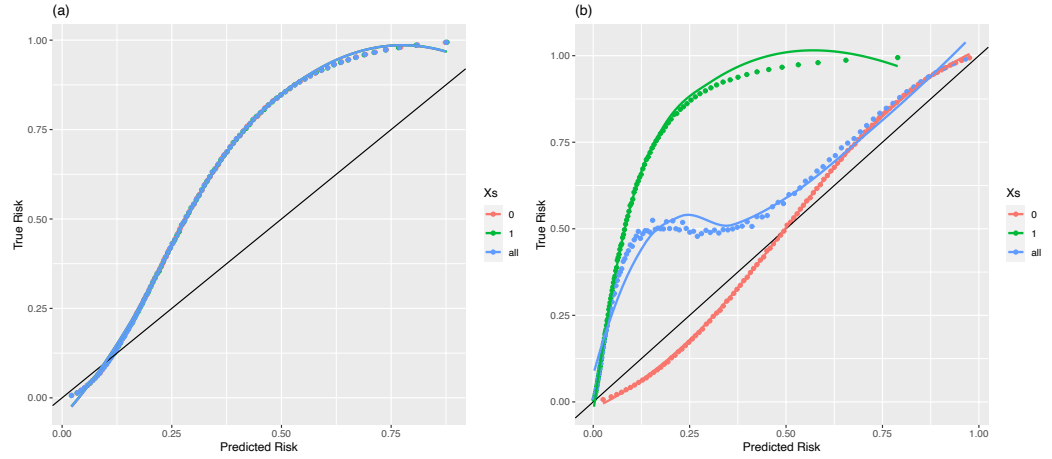

Figure 2.3: True versus predicted risk by the sensitive variable when it only impacts observing outcome

($\beta_s = 0, \beta_0 = 0, \beta_1 = 2, \gamma_s = -11, \gamma_0 = 10$). Left (3a): Excluding sensitive variable in model. Right (3b): Including sensitive variable in model.

We emphasize that the differential bias we are interested in is the difference between subgroup biases. In Figure 2.3a the overall model is biased due to the unobserved outcomes. However, this bias is non-differential, as the curves for each group

and the overall mean curve overlap. Only when we include the sensitive variable into the CPM does it become differential. In particular, the group without observability concerns (red) is almost perfectly calibrated while the group with observability concerns (green) is biased.

Therefore, to mitigate differential bias, the choice of whether to include X_s in the model becomes obvious in the cases where it only impacts one of the two stages. If there is no differential observability issue associated with X_s, then it should be included in the CPM; if X_s only impacts the observation process, it is better to exclude it in the model. When X_s is continuous the story is the same (Appendix A.).

2.3.2 Sensitive Variable Impacting *both* the Outcome and Observability

When X_s affects both the true outcome and observing the outcome the results are less straightforward. Figure 2.4 shows the value of $|\Delta_{in}Bias| - |\Delta_{ex}Bias|$ varies as effect of the sensitive variable on the true outcome (β_s) and on observability (γ_s) varies (simulations 3 and 6). If the value of $|\Delta_{in}Bias| - |\Delta_{ex}Bias|$ is positive (red), it is better to exclude X_s in the model; if the value is negative (blue), it is better to include X_s in the model. The heatmaps are consistent with the results discussed earlier when the sensitive variable only impacts one of the stages (γ_s fixed at 0 or β_s fixed at 0). A more general pattern can also be seen from the figure. The upper left corners of (4a) and (4b) are red indicating that if X_s has a positive effect on the true outcome and a negative effect on observability, then it is better not to include X_s. If the sensitive variable effects are in the same direction (both positive or negative), it is better to include X_s in the model. If we have fully observed data ($\gamma_s = 0$ in Figure 2.4b), we should include X_s in the model. Moreover, the possible correlation between X_1 and X_s does not meaningfully affect the

results under these specific simulation settings (Appendix A1). The main driver of differential bias is the sensitive variable effect on the true outcome and observability.

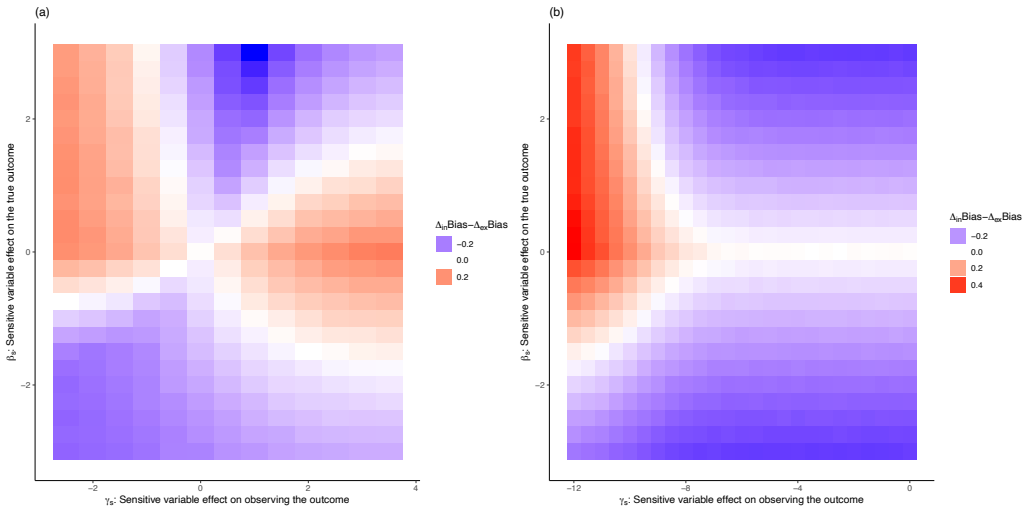

Figure 2.4: Heatmaps of $|\Delta_{in}\text{Bias}| - |\Delta_{ex}\text{Bias}|$ when β_s and γ_s vary.

Left(4a): Probability of observing outcome for reference group is set to be 0.5 ($\gamma_0 = 0$), $P(Y|E=1, X_s=0) = 0.5$. Right(4b): $\gamma_0 = 10$, $P(Y|E=1, X_s=0) = 1$.

2.3.3 Assessment with Observed Data

A follow-up question is whether we can use observed data to assess whether we should include a sensitive variable in a CPM. Table 2.2 shows the values of two commonly used evaluation metrics: AUC and calibration-in-the-large (from simulations 1 and 2). We can see that the AUC between two groups (**Δ**AUC) does not change when X_s is included or excluded. The calibration-in-the-large difference between two groups is always smaller when X_s is included in the model (**Δ**Calibration). This result suggests that metrics that are only based on observed data may not be able to detect the true differential bias induced by observability differences.

Table 2.2: Table of ΔAUC and ΔCalibration under different scenarios over 1000 simulations with 95% confidence interval.

Note that cases with smaller differential bias (bold) are not necessarily corresponding to smaller ΔAUC and ΔCalibration.

	X_s only impacts the true outcome		X_s only impacts observability	
	ΔAUC	**ΔCalibration**	**ΔAUC**	**ΔCalibration**
Exclude X_s	0.025 (-0.028,0.075)	0.130 (0.079,0.183)	***0.152*** (0.091,0.216)	***-0.366*** *(-0.412,-0.219)*
Include X_s	***0.025*** (-0.028,0.075)	***0.0004*** (-0.071,0.069)	0.152 (0.091,0.216)	0.002 (-0.063,0.061)

2.3.4 Real Data Example Results

Table 2.3 shows the model performance for a CPM predicting 30-day hospital readmission, with and without *Local* status. There is not a meaningful difference between the AUCs of the local and non-local groups (ΔAUC). Moreover, the calibration-in-the-large subgroup difference (ΔCalibration) is smaller when *Local* is included. However, since living locally only impacts the observability of the outcome (30-day readmission), according to the simulations we should have smaller *true* differential bias when excluding it from the model. Therefore, as suggested by Table 2.2, we cannot use performance in observed data to determine if there is differential bias.

Table 2.3: Table of AUCs and Calibrations of models include or exclude variable Local with 95% confidence intervals.

ΔAUC and ΔCalibration are the differences between local and non-local AUCs and calibration-in-the-large.

	Exclude *Local*	**Include *Local***
ΔAUC	-0.047 (-0.110,0.015)	*-0.044* *(-0.105,0.020)*
AUC (*Local*=1)	0.638 (0.582.0.693)	*0.642* (0.587,0.697)

AUC (*Local*=0)	0.685 (0.657,0.714)	*0.685* (0.657,0.714)
ΔCalibration	0.018 (0.002,0.036)	*0.010* (-0.007,0.026)
Calibration (*Local*=1)	0.014 (0.003,0.028)	*0.009* (-0.006,0.022)
Calibration (*Local*=0)	-0.004 (-0.014,0.006)	*-0.001* (-0.012,0.009)

2.4 Discussion

In this paper, we propose a formal definition and mechanism for differential bias in CPMs. We first consider how differences in observing an outcome may induce differential bias. Our study attempts to shed some light on the increasingly common question of whether one should include non-clinical factors (i.e. socioeconomic status, race/ethnicity, etc.) in a CPM (Anderson et al., 2021; Paulus et al., 2018; Vyas et al., 2020). Our simulation studies show that the answer is more complex than a simple yes or no.

Our definition of algorithmic bias is based on statistical bias. We extend this to notion of differential bias, a between-group difference of the predicted versus true risk. Similar definitions of bias of subpopulation differences have been considered by others (Barda et al., 2021). This defines algorithmic bias as a difference-in-differences between the "truth" and the "estimand", highlighting that this is based on unobserved factors. In particular, we do not consider differential bias based on differential performance within subgroups because that may be potentially due to predictive heterogeneity as opposed to a true bias.

Using our definition, we propose a two-stage model by which differential bias could arise. Specifically, we posit that differential bias is due to differential observability of patient outcomes. As our opening vignettes discuss, there are a variety of real-world mechanisms through which observability concerns arise: if a particular group has less access to healthcare, less likely to receive the same type/quality of care, if a diagnostic tool performs worse in one group versus another, or if a disease manifests itself differently in different groups. In these cases, patients may have equal rates of the outcome, but our observation of that outcome in the data will differ systematically based on the sensitive variable. As such we posit, that including a sensitive variable in a CPM, under any of these conditions, when it *only* impacts observability, can induce differential bias. A similar mechanism was considered by Boeschoten et al. (2021) as well. In their work, the authors take a latent variable approach considering the observed Y to be a (potentially poor) proxy for the true E. We embed our work in observability concerns that are particularly pertinent to the use of EHR data.

Our simulation results support our theoretical model. When the sensitive variable only impacts the observability of the outcome, in which case differential observability issues may occur, including it in a CPM can induce bias. Conversely, if there is no differential observability issue, then it should be included. For the more complicated cases where the sensitive variable impacts both the true outcome and observability, there is no simple answer to when to include the variable in the model. For example, with COVID-19, age likely impacts *both* the observability of COVID-19 infection (via the development of severe infection (Kronbichler et al., 2020) and general health-seeking behaviors) and the likelihood of being infected. In this case, age is both a true risk factor

(and should be included in a CPM for COVID-19 infection) and a source of differential observability (and should not be included in a CPM). As our simulations show, the differential bias is determined by the direction and magnitude of the effects of that variable on both stages. In this scenario there will always be some bias within the CPM - whether the sensitive variable is included or not - the question then becomes a matter of which is effect is stronger.

This work highlights the importance of asking why we see what we see in EHR datasets. Often the differential access and receipt of medical care observed among individuals categorized according to race or gender is a result of structural racism and gender inequities in society which are not well captured in medical records and related exposures or factors cannot easily be quantified using EHR data ("On Racism: A New Standard For Publishing On Racial Health Inequities," 2020). Further, many variables such as race and gender are inaccurately assigned or incompletely captured within EHR data sources (Paulus et al., 2018). As such, these findings corroborate recent suggestions that race (and similar sensitive variables) should only be included in CPMs if there is a clear rationale for including it. For instance, while race should not be included in CPMs with the intent of using it as a proxy for biological mechanisms, it could be included as a social variable, with the intent of demonstrating how societal racism could produce health effects (Vyas et al., 2020). The rationale for including race and similar variables in CPMs is often not clearly stated, leaving open the possibility that CPMs could be erroneously interpreted. While our work formalizes that a mere statistical association of race with clinical outcomes is not sufficient because the observed statistical association may be due

to differential observability, it also underscores the importance of considering the mechanisms through which such differential observability may occur.

These results leave CPM developers with a difficult choice: include a sensitive variable or not? We assessed whether one can answer this question with observed data by looking at differential AUC and calibration. The results suggested that this might not be possible. Whether a sensitive variable is truly associated with an outcome or only associated with the observability of the outcome, there will always be a *statistical* association between the observed outcome and the sensitive variable. As such, a statistical approach that uses the empirical data to test whether one should or should not include a sensitive variable in a CPM, will always suggest that one should include the sensitive variable. This may contradict the work of others that have suggested looking at sub-group calibration metrics (Barda et al., 2021; Hébert-Johnson et al., 2018). Instead, the analyst needs to consider whether the inclusion of a sensitive variable improves the estimation of the true, unobserved outcome. As such, the question of observability is an implicit assumption that one needs to make when including sensitive variables in a CPM. This is analogous to the well-known assumptions made in association studies such as "no unmeasured confounding". Other potential solutions to this problem, like the latent variable solution proposed by Boeschoten et al. (2021), rely on a missing at random assumption. While this may be appropriate in the latent/proxy variable formulation, most observability concerns are implicitly missing not at random (Wells et al., 2013).

These findings are an example of informed presence bias (Goldstein et al., 2016; Phelan et al., 2017), the fact that the data we observe within EHRs are informative. Moreover, this work fits into a larger body of epidemiological literature focused on

differential missingness and measurement error (Allen et al., 2003; VanderWeele & Hernan, 2012; VanderWeele & Li, 2019). Within the epidemiologic community, it is well recognized that when data are differentially missing or have differential measurement errors, association results can be biased (Copeland et al., 1977; Heid et al., 2002). A lack of observability is one process – unique to EHR data – through which misclassifications are generated (McGee et al., 2022). We extend previous literature by exploring the impact of differential observability, uniquely in the context of CPMs. Ultimately, we show, that like with association studies, there are some unverifiable assumptions researchers need to make when developing EHR based CPMs.

While our results shed important light on the role observability can play in the development of differential bias in CPMs, there are some important limitations. First, our simulations focus on the cases where the sensitive variable affects only one of the two stages. In the more complex scenarios, the conclusions become more subtle, relating to the overall strength of each effect. Moreover, there could be interaction between a differentially observed variable and a true risk factor. We assume that the sensitive variable directly impacts the outcome or the observability, but it is often more complicated. It can have direct effects and indirect effects at the same time. There are also more complexities to be considered where the sensitive variable serves as a proxy for the true causal factor. Finally, we only frame one particular source of differential bias - the observability of the source data. Biases can also arise from the person analyzing data based on, often unstated, assumptions about how the data are coded and processed. It is important for all sources of differential bias to be fully interrogated.

2.5 Conclusion

In conclusion, our study furthers the literature on algorithmic fairness. We propose a formal definition of differential bias which is based on statistical bias. We explore the question of whether one should include a sensitive variable into a CPM and show that the answer depends on whether the sensitive variable truly affects the outcome or just the observability of the outcome. However, we also show that answering this question may not be possible solely on observed data and may instead be an implicit assumption of CPMs. Future work ought to further explore diagnostic measures for determining whether one should include a sensitive variable or not within a real-world dataset as well as further explore other sources of algorithmic bias.

FUNDING STATEMENT

This work was supported by National Institute of Diabetes and Digestive and Kidney Diseases, grant RO1-DK123062 (BAG), and National Institute of Neurological Disorders and Stroke, grant R61-NS120246 (MJP).

3. Estimating the Observability of an Outcome from an Electronic Health Records Dataset Using External Data[1]

One of the key limitations of electronic health records (EHR) data is that not all health care encounters are observed. The degree to which patient information is captured is referred to as observability. Poor observability, and in particular differential observability, can lead to biased estimates and inference. As such, understanding the degree of observability is important in EHR based studies. In this study, we propose using external data with known observability to assess the degree of overall observability in EHR. We also construct a test for differential observability in the target EHR dataset. Using principles from the transportability literature, we show that we can use a balancing score based weight to estimate the observability of our target outcome. We conduct a series of simulation experiments to understand the conditions under which dataset features must be required to obtain proper inference. To illustrate this, we consider hospital readmissions among patients with end stage renal disease as our outcome of interest. We use administrative claims data, where the outcome is fully observed, as our external data.

3.1 Introduction

Electronic health records (EHR) are an important data source for healthcare research as it includes abundant patient-level information about patients who see clinicians in the real world, including patient demographics, diagnoses, medications, vital signs, laboratory

[1] This chapter is a published paper: Yan, M., Hong, H., Wilson, J., & Goldstein, B. A. (2025). Estimating the Observability of an Outcome from an Electronic Health Records Dataset Using External Data. American Journal of Epidemiology, kwaf013. https://doi.org/10.1093/aje/kwaf013

results, and clinical notes, that may not be accessible elsewhere (Goldstein, Navar, et al., 2017; Sharples, 2018). However, since EHR data are not collected for research purposes, there are challenges in using them such as incompleteness and incorrectness (Hersh et al., 2013; Weiskopf et al., n.d.; Sauer et al., 2022). Furthermore, since single-center EHR systems only capture patient data from interactions with a particular health system, the robustness of study results depends on the extent to which a patient receives care from a particular health system. This presents a specific data quality challenge unique to EHR data: incomplete data observability. Observability is defined as the degree to which patient characters are captured within EHR. Taking hospital readmission outcome as an example (Table 3.1), if a patient has a readmission but is not readmitted to the health system under study, then the readmission would not be captured in the observed EHR.

Table 3.1: Example of outcome observability in EHR data

True Outcome (Y)	**Observed Outcome (S)**	**Example**
Present (1)	Present (1)	The patient is readmitted to the hospital under study
Present (1)	Not Present (0)	The patient is readmitted to another hospital
Not Present (0)	Not Present (0)	The patient does not have any readmission

Observability and missing data problems are related, but slightly different. A key differentiator between a lack of observability and missing data is that in the traditional missing data paradigm, we typically have an indicator that a value is missing whereas in the typical observability setup we do not (Phelan et al., 2017; Rubin, 1976). For instance, in the missing data setting, there may be a missingness indicator that a patient should have had a study visit but did not; in the observability setup, a patient may receive care, unbeknownst to the researcher, at another institution. This places observability within an analytic paradigm referred to as "positive and unlabeled data" (Bekker & Davis, 2020; Elkan & Noto, 2008). In this paradigm, data that are observed (e.g., medical encounters, ICD codes) are considered definitively labeled as such, and all other instances are considered not labeled.

Characterizing observability is an important consideration because it can serve as a source of bias in clinical prediction models. In the "sources of harm in the machine learning life cycle framework" (Suresh & Guttag, 2021), observability can be seen as a measurement bias as the unobserved outcome may not be measured accurately. It can also be seen as historical bias as the data do not reflect a patient's true current status. Bias induced by observability also manifests as informed presence bias (Goldstein et al., 2016; McGee et al., 2022; Phelan et al., 2017), a type of bias that arises because interactions with the healthcare system are non-random. As a result, their presence in the EHR is selective. In our example, whether a readmission is recorded in the EHR may depend on factors influencing an individual's likelihood of seeking or receiving care, making it non-random. Moreover, if an outcome is differentially observed across some factors, such as

race and gender, this differential observability can induce algorithmic bias in the clinical risk prediction context (Yan et al., 2022).

Given the importance of observability, it is noteworthy that there has been minimal work to estimate the degree of observability in a EHR dataset. Several studies have used data continuity as a related concept (Lin et al., 2018; Merola et al., 2022), where they link EHR and claims data to estimate the mean proportion captured in EHR data. However, in practice, linking patients between EHR and claims data may not always be feasible. In this paper, we propose a method to estimate the degree of observability by comparing the outcome rates from the EHR dataset under a study to external data source that has known or complete observability. In our setup, the external data is not required to be linked to the target data. Instead, we incorporate a balancing score under a causal inference framework, similar to a propensity score, to reweight the external data to look like the target EHR data and estimate the observability of our target outcome. This allows us to estimate both the degree of observability as well as the potential for differential observability. We conduct a series of simulation experiments to understand the conditions under which this estimation holds. To illustrate our approach, we consider hospital readmissions among patient with end stage renal disease (ESRD) as our outcome of interest.

3.2 Methods

3.2.1 Definition of Observability

We define the observability of a binary outcome from a given data set as the probability of observing the outcome given the true outcome is positive:

$$P(S = 1|Y = 1, Z = z)\ ,$$

where Y is the true outcome, S is the observed outcome, and Z is an indicator for the data source. Our objective is to estimate the observability of a particular outcome in the target EHR data ($Z = 0$), defined as:

$$\pi = P(S = 1|Y = 1, Z = 0).$$

According to the definition, $0 \leq \pi \leq 1$, and $\pi = 1$ indicates that the outcome is fully observed. For target EHR data $Z = 0$, we assume that if $S = 1$ then $Y = 1$ always, while if $S = 0$ then Y could be either 0 or 1. This target EHR outcome structure is illustrated in Table 3.1. Additionally, we suppose an external data source ($Z = 1$) exists, where the true outcome Y is fully observed, i.e., $S = Y$ and $P(S = 1|Y = 1, Z = 1) = 1$.

3.2.2 Estimating Observability

In the target EHR data, observability of outcome Y can be re-expressed as the ratio of the probability of observing an event to the probability of truly having an event as follows:

$$\begin{aligned} \pi &= \mathrm{P}(\mathrm{S} = 1|\,\mathrm{Y} = 1, \mathrm{Z} = 0) \\ &= \mathrm{P}(\mathrm{S} = 1, \mathrm{Y} = 1|\,\mathrm{Z} = 0)/\mathrm{P}(\mathrm{Y} = 1|\,\mathrm{Z} = 0) \qquad (1) \\ &= \frac{\mathrm{P}(\mathrm{S} = 1|\mathrm{Z} = 0)}{\mathrm{P}(\mathrm{Y} = 1|\mathrm{Z} = 0)} \qquad (2) \end{aligned}$$

In the first component of Equation (1), $\mathrm{S} = 1$ implies $\mathrm{Y} = 1$ under $\mathrm{Z} = 0$, so $\mathrm{P}(\mathrm{S} = 1, \mathrm{Y} = 1|\,\mathrm{Z} = 0) = \mathrm{P}(\mathrm{S} = 1|\mathrm{Z} = 0)$. In Equation (2), the numerator $\mathrm{P}(\mathrm{S} = 1|\,\mathrm{Z} = 0)$ can be estimated directly using the target EHR data, while the denominator, $\mathrm{P}(\mathrm{Y} = 1|\,\mathrm{Z} = 0)$, cannot be directly estimated since Y is not observed in the target EHR data.

Let τ be the true outcome probability in the target EHR, with $\tau = P(Y = 1|Z = 0)$. We employ a causal inference framework to estimate τ using weighted

external data, and then further estimate observability, π. The idea of using external data is inspired by the generalizability and transportability literature. These methodologies often aim to estimate results about intervention effects in a target population using results obtained from randomized trials (Cole & Stuart, 2010; Degtiar & Rose, 2021; Stuart et al., 2011). In our observability context, we build a balancing score model on the data source membership and then reweight the external data to make them resemble the target EHR data. However, our objective is not to estimate the causal effect in the target population, but rather to estimate the potential outcomes of being in the fully observed population, which can be seen as the true outcome and further be used to estimate observability.

Let X denote covariates consistently observed and measured in both EHR and external data. Under the Rubin's potential outcome framework (Rubin, 1974), Y_i^z represents the potential outcome for individual i in data source z. We first make the following assumptions.

A1. Consistency of potential outcome: $Y_i^z = Y_i \; if \; Z_i = z$. The potential outcome of a given data source for an individual is the same as true outcome if the individual is from that data source.

A2. Positivity of data source membership: $0 < P(Z = 1|X = x) < 1$. The probability of being in either data source is positive. This implies a good overlap in the covariate distributions between data sources.

A3. Data source exchangeability: $\{Y^{z=0}, \; Y^{z=1}\} \perp Z|X$. Potential outcomes are independent of data source given the observed covariates.

A4. Data source has no causal effect on average on the target data: $E(Y^{Z=0}|\mathrm{Z}=0)$= $\mathrm{E}(Y^{Z=1}|\mathrm{Z}=0)$. This is reasonable as data source should not impact the outcome of the individual. This assumption is important to ensure the results $Y^{Z=1}$ can be transported to $Y^{Z=0}$ on target data.

Under these assumptions, our proposed method for estimating observability follows three steps shown in Figure 3.1.

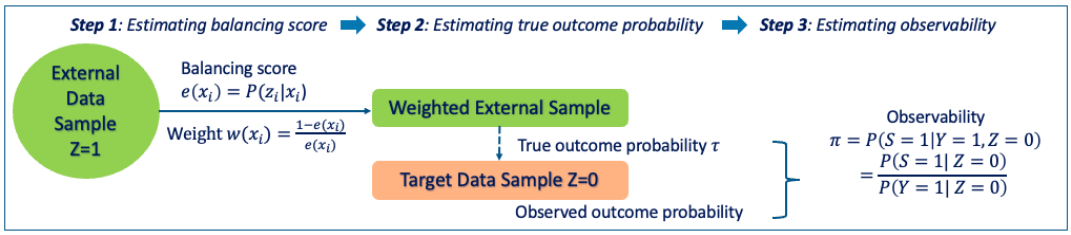

Figure 3.1: Diagrams illustrating the proposed method.

Step 1: *Estimating Balancing Score*

First, we estimate a balancing score to construct weight for the external data. The balancing score is defined as $e(x_i) = P(Z_i = 1|X_i = x_i)$, which is the probability of being in the external dataset given observed covariates. Given assumptions A2 and A3, covariate distributions should be similar between EHR and external data given the balancing score. The balancing score can be estimated using a logistic regression model.

Step 2: *Estimating True Outcome Probability* τ

Under A1, the estimand τ – true outcome probability in the target EHR data – can be written as $\tau = P(Y = 1|\,\mathrm{Z}=0) = P(Y^{Z=0} = 1|\,\mathrm{Z}=0) = \mathrm{E}(Y^{Z=0}|\,\mathrm{Z}=0)$.

Li et al. (2018) showed that under an assumption analogous to A3, the weighted rate $\frac{\sum_i w(x_i)Z_iY_i}{\sum_i w(x_i)Z_i}$ is a consistent estimator of $E(Y^{Z=1}|\,Z=0)$, where $w(x_i) = \frac{1-e(x_i)}{e(x_i)}$. Then, since $Y_i = S_i$ when $Z_i = 1$, the estimator can be re-expressed as $\frac{\sum_i w(x_i)Z_iS_i}{\sum_i w(x_i)Z_i}$. Based on

assumption A4, $E(Y^{Z=1} | Z = 0) = \tau$, therefore, the target data true outcome expectation can be estimated by

$$\hat{\tau} = \frac{\sum_i \widehat{w}(x_i) Z_i S_i}{\sum_i \widehat{w}(x_i) Z_i}, \tag{3}$$

where $\widehat{w}(x_i) = \frac{1-\hat{e}(x_i)}{\hat{e}((x_i)}$.

Step 3: *Estimating Observability π*

Since S is observed in the target EHR data, $\mathrm{P}(\mathrm{S} = 1 | \mathrm{Z} = 0)$ can be directly estimated by the observed sample mean of target data $\frac{\sum_i (1-Z_i) S_i}{\sum_i (1-Z_i)}$. By plugging Equation (3) into Equation (2), the estimator of observability π is:

$$\widehat{\pi} = \frac{\frac{\sum_i (1 - Z_i) S_i}{\sum_i (1 - Z_i)}}{\hat{\tau}}. \tag{4}$$

Subgroup Observability

Knowing subgroup observability is useful for assessing potential differential bias in clinical prediction models based on the target EHR dataset (Yan et al., 2022). The estimation of observability can be easily extended to estimating observability for subgroup $X_g = x_g$:

$$\pi(x_{\mathrm{g}}) = P\big(S = 1 \big| Y = 1, X_g = x_g, Z = 0\big).$$

Then differential observability can be calculated by subtracting the estimated observabilities between subgroups.

Confidence Intervals and Differential Test

Confidence intervals for the estimated observability and subgroup observability can be further obtained via bootstrap. Fixing external data, we create bootstrap samples for the target EHR data and calculate the estimated observability for each bootstrap

sample $\hat{\pi}_{boot,b}$. Using the nonparametric bootstrap samples of estimated observability as $(\hat{\pi}_{boot,1}, \hat{\pi}_{boot,2}, \dots, \hat{\pi}_{boot,B})$, the 95% bootstrap confidence interval is the 0.025 and 0.975 quantiles for bootstrap samples, $(\hat{\pi}_{boot,\alpha=0.025}, \hat{\pi}_{boot,\alpha=0.975})$.

Similarly, 95% bootstrap confidence interval for differential observability $(\hat{\delta} = \hat{\pi}_{groupA} - \hat{\pi}_{groupB})$ is obtained as $(\hat{\delta}_{boot,\alpha=0.025}, \hat{\delta}_{boot,\alpha=0.975})$. Confidence intervals can be used to test if there is an observability difference between two subgroups. If 0 is not in the bootstrap confidence interval for differential observability, then there is a significant observability difference between two subgroups.

3.2.3 Simulation Setup

We conducted a comprehensive simulation study to assess the conditions under which the proposed estimation method performs well. We consider three degrees of overlap in the covariate distributions between data sources following the simulation setup by Zhou et al. (2020). In addition, we study the impact of balancing score model misspecifications on our proposed estimation.

We generate covariates X_1, X_2 and X_3 from a normal distribution $N(0,1)$ and we design the true balancing score model and true outcome model in such way that X_1 is a predictor of outcome with good balance between target and external data, X_2 is not a predictor with poor balance between target and external data, and X_3 is a predictor of outcome with poor balance between data sources (i.e., confounder).

The balancing score model reflects the imbalance of X_2 and X_3 across data sources. Specifically, the data source membership indicator Z is generated based on the true balancing score model: $Z \sim Bernoulli\big(e(X)\big)$, where $logit\big(e(X)\big) = \alpha_0 + \alpha_2 X_2 +$

$\alpha_3 X_3$. Here, smaller α_2 and α_3 values give better covariates overlap between the two data sources. True binary outcome Y is generated from $Bernoulli\big(p(X)\big)$, where $logit\big(p(X)\big) = \beta_0 + \beta_1 X_1 + \beta_3 X_3$ reflects the true predictors $X_1$and X_3. By data structure defined in Table 3.1, the observed outcome S in the target EHR data ($Z = 0$) is always 0 when $Y = 0$. However, when $Y = 1$, S is generated from a Bernoulli distribution with probability π, representing the true observability. For the external data ($Z = 1$), $S = Y$ assuming full observability.

Following the data generating process, we set sample sizes for target and external data as $n_0 = n_1 = 5{,}000$, and set $\beta_0 = \beta_1 = \beta_3 = 1$. We conduct simulations based on 500 iterations on three different true observabilities $\pi = 1, 0.8,$ and 0.6, and three different overlap levels: good overlap under $\alpha_2 = \alpha_3 = 0.1$, moderate overlap under $\alpha_2 = \alpha_3 = 1$, and poor overlap under $\alpha_2 = \alpha_3 = 3$. In addition, six different combinations of covariates are considered in the balancing score models (BSMs) presented in Table 3.2. BSM6 represents the scenario where the balancing score model is correctly specified, while the other BSMs are misspecified models. Additionally, we have included a simulation with 10:1 external-to-target sample sizes ratio in Appendix. We have also evaluated the method in a plasmode simulation (Franklin et al., 2014) based on the real EHR data in Appendix to show that the method works in higher dimensional, correlated cases.

Table 3.2: Balancing score models with different types of covariates for simulation

	BSM1	**BSM2**	**BSM3**	**BSM4**	**BSM5**	**BSM6**

X1 (True outcome predictor)	✓			✓	✓	
X2 (Imbalanced covariate)		✓		✓		✓
X3 (Confounder)			✓		✓	✓

After generating the data, we first fit a logistic regression as balancing score model to obtain the estimated balancing score $\hat{e}(x)$, then calculate the estimated true outcome expectation $\hat{\tau}$ using external data weighted by $\widehat{w}(x)$ and finally calculated the estimated observability $\hat{\pi}$. Then, we calculate the mean absolute error (MAE) and coverage rate of 95% confidence intervals for both estimated true outcome probability $\hat{\tau}$ and estimated observability $\hat{\pi}$.

3.2.4 Real Data Application

As an illustrative data example, we assess 30-day readmission for patients with ESRD, who requires complex care with high rates of service utilization (Johansen et al., 2023; Kolbrink et al., 2023).

Target EHR Data

The target EHR data are from Duke University Health System (DUHS) accessed through a research data mart modeled to the PCORnet Common Data Model (Hurst et al., 2020). DUHS is a medium sized health center in central North Carolina. It consists of three hospitals, including the only two hospitals in Durham County. DUHS serves as both the primary provider for individuals living in Durham County, as well as referral center for individuals across North Carolina (Stolte et al., 2022).

For the target cohort, we include all inpatient ESRD encounters from 2015 to 2022. We define an ESRD encounter as an encounter with a billing ICD-9 code 585.6 or ICD-10 code N18.6 and with a dialysis CPT code 90935 or 90937. Since we have restricted the codes we used to billing codes, it is reasonable to assume that these codes are measured consistently in claims data as well. The analysis is restricted to Medicare patients aged 18+ discharged to home.

External Data

For external data, we use Medicare claims data available from the United States Renal Data System (USRDS) (Johansen et al., 2023). Per federal law, all ESRD patients are eligible for Medicare, regardless of age. These claims data are organized into the USRDS database and include information on health care encounters. We identified all patients' ESRD encounters of which the discharge date is between 2014 and 2018 and discharged to home. We limited encounters based on age, Medicare as primary payer, and ESRD encounter rules as described above. To make the target EHR data and external data comparable, we limited the claims data to North Carolina encounters.

Outcome

Our outcome of interest is 30-day readmission, defined as a return inpatient visit within 30 days of discharge. We presume that not all readmissions will be observed within the DUHS EHR system, because patients may re-hospitalize to a different facility, while Medicare data is presumed to capture all readmissions for its primary payers.

Balancing Covariates

We used covariates available in both data sources including age, sex, race/ethnicity, length of stay of initial hospitalization, and all diagnostic codes (ICD-9-

CM and ICD-10-CM) for each encounter that are grouped into smaller disease categories using Clinical Classifications Software (CCS) (*Clinical Classifications Software (CCS) for ICD-9-CM*, n.d.; *Clinical Classifications Software (CCS) for ICD-10-PCS (Beta Version)*, n.d.; Elixhauser, 1996). This resulted in a total of 270 covariates.

Estimation of Balancing Score

Due to a large numbers of covariates, we used logistic Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) to estimate the balancing score. All the 270 covariates are included in the model. The estimated balancing score was then used to calculate the balancing weight.

Data Description

We compare the mean, median and interquartile range (IQR) of continuous covariates, and counts and percentages of categorical covariates. We used standardized mean difference (SMD) to compare variables in the target EHR dataset and external claims dataset, and SMD for weighted samples can also be calculated using weighted sample mean (Austin, 2011; Austin & Stuart, 2015). SMD <0.1 is commonly seen as balanced.

Estimation of Observability

We first estimate the true readmission rate in EHR $\hat{\tau}$ based on weighted claims data weighting by estimated balancing score. Then according to equation (4), we estimate the observability of readmission in EHR data and the associated 95% confidence intervals. We look at both overall and subgroup readmission rates in EHR and claims data, estimated true outcome probabilities, and observabilities. We consider the following subgroups: Sex (categorized as male/female), Race/Ethnicity (categorized as non-

Hispanic White, non-Hispanic Black, Hispanic) and patient locality (categorized as local to DUHS or not). We also further look at the differential observabilities and its confidence intervals of these subgroups. Since local/non-local variable is not available in claims, the observabilities for local and non-local subgroups are based on the overall estimated true outcome ratio.

3.3 Results

3.3.1 Simulation Results

Simulation results are shown in Figure 3.2 and Figure 3.3. Results show that when a confounder is included in the balancing score model (BSM3, BSM5, BSM6), the MAEs for estimated true outcome probability $\hat{\tau}$ and estimated observability $\hat{\pi}$ are close to 0 even under poor overlap scenarios. The 95% coverage rate for the corresponding cases is also close to 95%. Therefore, our proposed method gives a good estimation of observability even when some imbalanced variables or true outcome predictors are not available as long as confounders are well collected. In cases when the confounder is not included in balancing score models (BSM1, BSM2, BSM4), the worse the overlap degree, the larger the MAE. Thus, when confounders are not available, a good overlap is required to obtain small MAEs for the estimated true outcome probability $\hat{\tau}$ and estimated observability $\hat{\pi}$, though the coverage rates could still be low.

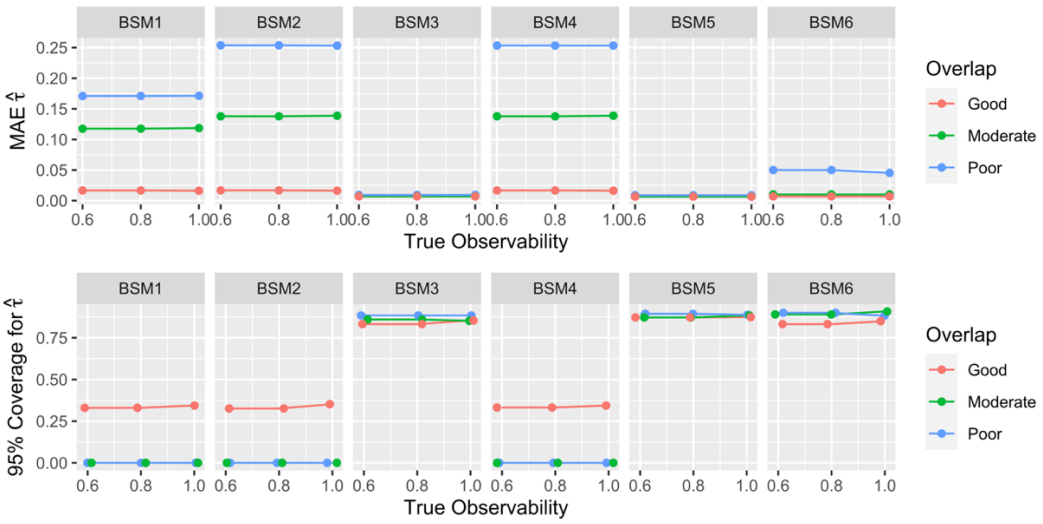

Figure 3.2: Simulation results for estimated true outcome probability using different covariate sets in balancing score models under different overlap levels and different true observabilities.

BSM1: Outcome predictor only model; BSM2: Imbalanced covariate only model; BSM3: "confounder" only model; BSM4:Outcome predictor & imbalanced covariate model; BSM5: Outcome predictor & "confounder" model; BSM6: Imbalanced covariate & "confounder" model.

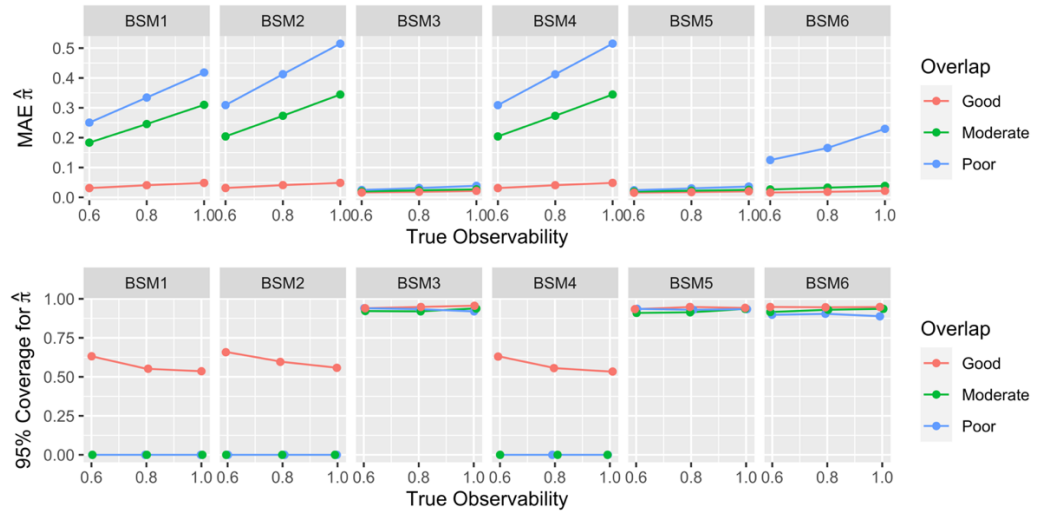

Figure 3.3: Simulation results for estimated observability using different covariate sets in balancing score models under different overlap levels and different true observabilities.

BSM1: Outcome predictor only model; BSM2: Imbalanced covariate only model; BSM3: "confounder" only model; BSM4: Outcome predictor & imbalanced covariate model;

BSM5: Outcome predictor & "confounder" model; BSM6: Imbalanced covariate & "confounder" model.

3.3.2 Real Data Results

Table 3.3 shows the descriptive statistics of selected covariates for the EHR and external samples, and Figure 3.4 shows the SMDs and weighted SMDs. They include five most prevalent CCS categories in the EHR population and the five most imbalanced CCS categories between the two data sources. There are 5,564 and 53,009 encounters for the target EHR and external claims sample respectively. Age and length of stay have similar distributions in the two data sources with small SMDs. Distributions of sex and race/ethnicity are slightly different between the two data sources with SMDs>0.1. Most of the CCS categories are highly imbalanced with large SMDs. After weighting, SMDs for Sex and Race/Ethnicity fell below 0.1, and the most imbalanced CCS categories had significantly reduced SMDs. Small weighted SMDs indicate that the estimated balancing score successfully weights external claims to resemble the target EHR sample.

Table 3.3: Summary of claims and EHR data

	Target EHR Data (n=5,564)	**External Claims Data (n=53,009)**
Age, year, mean (median, IQR)	59.0 (59.7, 18.7)	58.8(60, 20)
Sex, N (%)		
Female	2,403 (43.2)	27,124 (51.2)
Male	3,161 (56.8)	25,885 (48.8)

Race/Ethnicity, N (%)		
Non-Hispanic White	1,292 (23.2)	16,944 (32.0)
Non-Hispanic Black	4,179 (75.1)	34,677 (65.4)
Hispanic	93 (16.7)	1,388 (26.2)
Length of stay, days mean (median, IQR)	6.02 (4, 4)	5.63 (4, 4)
Five most prevalent CCS categories in EHR (%)		
Diseases of the urinary system (CCS 158)	90.7	56.0
Hypertension (CCS 99)	85.7	53.7
Anemia (CCS 59)	77.2	72.8
Factors influencing health care (CCS 257)	75.0	60.7
Residual codes, unclassified (CCS 259)	67.5	53.3
Five most imbalanced CCS categories (%)		
Essential (primary) Hypertension (CCS 98)	39.2	0.8
Other lower respiratory disease (CCS 133)	49.6	7.2

Diseases of the urinary system (CCS 158)	90.7	56.0
Hypertension (CCS 99)	85.7	53.7
Diseases of the heart (CCS 104)	16.7	0.8

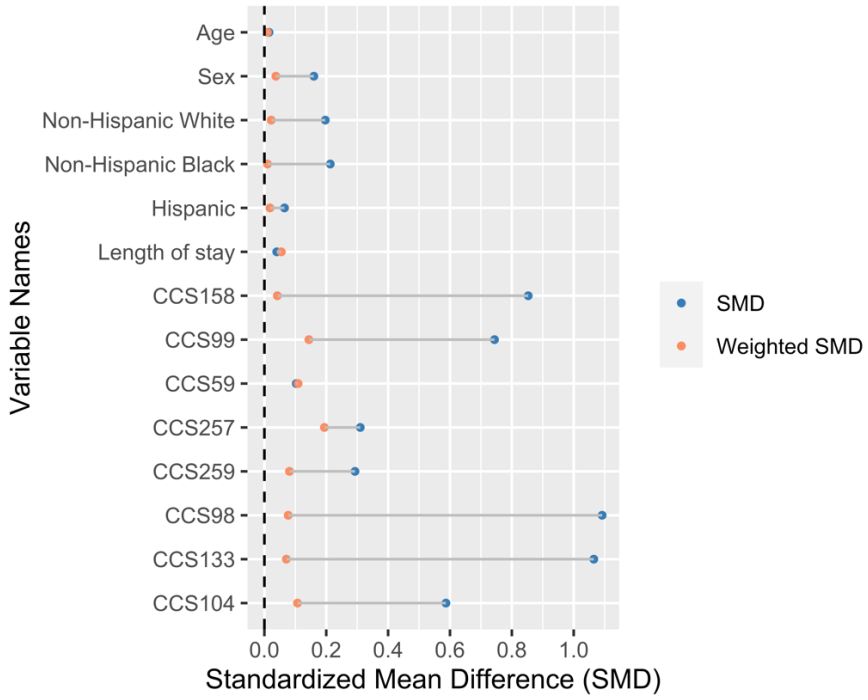

Figure 3.4: Standardized mean difference for top covariates

In Table 3.4, the observed readmission rate in EHR is lower than the observed readmission rate in claims. After weighting the claims data to look like the EHR sample, the expected readmissions with the EHR-based sample rate are slightly higher than before weighting. This allows us to estimate the overall observability for readmission as 0.916,

which means that 91.6% of the true readmissions are observed in the EHR dataset. Estimated subgroup observabilities for sex, race/ethnicity, and locality subgroups are also shown in Table 3.4. Several subgroups show lower observability than the overall observability. In particular, Hispanic individuals have the lowest observability rate of 0.572. Table 3.5 shows tests for differential observability. The 95% confidence interval for differential observability of sex does not include 0 with a point estimation 0.141, which can be interpreted that true readmissions are significantly more likely to be collected in our target EHR data in males than females. For Race/Ethnicity, there were significant differences between Non-Hispanic Blacks and both Non-Hispanic White and Hispanic individuals, with the true readmissions in Non-Hispanic Black individuals significantly more likely to be observed than other groups in our target EHR dataset. Although there is a difference in point estimations of observability between Non-Hispanic White and Hispanic individuals (0.227), and between local and non-local groups, the differences were not statistically significant.

Table 3.4: Readmission outcome rates in EHR and claims cohorts and estimated observability results.

	Observed readmission rate in EHR	**Observed/True readmission rate in claims**	**Estimated true readmission rate for EHR data (based on weighting claims data)**	**Estimated observability (95% CI)**

Overall	0.251	0.305	0.274	0.916 (0.866, 0.980)
Sex				
Male	0.248	0.313	0.253	0.980 (0.913, 1.079)
Female	0.256	0.297	0.305	0.839 (0.774, 0.905)
Race/ethnicity				
Non-Hispanic White	0.236	0.308	0.295	0.799 (0.685, 0.918)
Non-Hispanic Black	0.258	0.303	0.268	0.926 (0.902, 1.035)
Hispanic	0.151	0.294	0.263	0.572 (0.312, 0.841)
Locality				
Local	0.264		0.274	0.964 (0.886, 1.041)
Non-local	0.239		0.274	0.872 (0.825, 0.952)

Table 3.5: Readmission outcome differential observability

Comparison groups	Estimated differential observability (Bootstrap 95% CI)
Male, female	0.141 (0.057, 0.279)
Non-Hispanic White, Hispanic	0.227 (-0.040, 0.528)
Non-Hispanic White, Non-Hispanic Black	-0.127 (-0.291, -0.018)
Non-Hispanic Black, Hispanic	0.354 (0.111, 0.663)
Local, Non-local	0.092 (-0.017, 0.171)

3.4 Discussion

In this paper we propose a novel approach based upon a causal inference framework to estimate observability in a target EHR dataset leveraging information from a fully observed external dataset. To the best of our knowledge, this is the first formal method for estimating observability. The proposed method is broadly applicable, requiring only overlapping covariates between datasets without necessitating a direct linkage between them. Our simulation study showed that the proposed estimator has a low MAE and high coverage rate as long as all confounders are included in the balancing score model. Furthermore, we illustrated our methods using the hospital readmission example and demonstrated how to apply and interpret the differential observability test to assess if two subgroups have a significant difference in observability.

Our approach leverages external data, drawing inspiration from the literature on generalizability and transportability (Cole & Stuart, 2010; Degtiar & Rose, 2021; Stuart et al., 2011). One strategy for generalizing an estimate involves adapting propensity score methods. Instead of modeling the probability of receiving a treatment, this approach

models the propensity score of membership in a randomized trial (Kern et al., 2016; Stuart et al., 2011). By the balancing property of propensity scores (Li et al., 2018; Rosenbaum & Rubin, 1983), we can reweight the external population based on propensity scores to create a pseudo-population with similar covariates. Kern et al. and Westreich et al. have previously introduced approaches using propensity score methods for generalizing or transporting treatment effect estimation (Kern et al., 2016; Westreich et al., 2017). In the observability context, we adopt a similar idea by constructing a balancing score model on the data source membership and then reweighting the external data. For our readmission application, we employed LASSO for the balancing score model, but the method can accommodate any suitable model for estimating balancing score that ensures covariate balance through weighting. While our paper focuses on an outcome, our method can be adapted to address observability issues for any binary variable. The method could also assess observability patterns for continuous variables — whether these values are recorded — but is not intended for imputing or predicting the values themselves.

Many studies have seen incomplete EHR data as a missing data problem. Observability and missing data (Rubin, 1976) are related concepts, but differ in focus: observability concerns what data are observed, while missing data concerns how data are missing. As pointed out by Weiskopf et al., it is important to recognize that complete information on the patient and complete information on the patient encounters in the EHR are not the same thing (Weiskopf et al., 2013). Thus, simply having non missing records of the EHR encounters does not necessarily imply that the patient outcome is fully observed. Although we can treat unobserved outcomes as missing data, the

missingness indicators are unknown, which can make it challenging to distinguish between missing values and actual outcomes of zero. For example, as illustrated in Table 3.1, readmission would be coded as 0 (not present) either when the patient readmission information is missing in EHR, or the patient does not have any readmission. Furthermore, the observability problem would be seen as missing not at random (MNAR) in the missing data mechanisms (Rubin, 1976) since whether the outcome is missing (observed) depends on if the true outcome is positive or negative.

In the real-world data application, we showed that about 10% of the readmission outcome is not observed in the target EHR dataset. We also showed that there is differential observability between male and female groups and different race/ethnicity groups. In our target data of DUHS, readmissions among males have significantly greater observability than females, and readmissions among non-Hispanic Black individuals are more observed than in non-Hispanic White or Hispanic individuals. Our previous work has shown that differential observability can lead to algorithmic bias in clinical prediction models (Yan et al., 2022). Any predictive model based on health system data would need to be thoughtful about how such demographic factors were incorporated.

This approach has some notable limitations. First, the proposed method requires a secondary external dataset with full outcome observability, which may not always be readily available. We used administrative data specific to dialysis patients, however, Medicare or Medicaid are often available to researchers, and will contain information on a broader set of patient populations. Randomized clinical trials (RCTs) are also widely used in secondary data analysis, and similarly will have full adjudication of events. Finally, large epidemiological cohorts, which are available in repositories such as dbGaP,

will also typically have better outcome observability than what we expect in EHR data sources. Another limitation is that the balancing score model can only use variables present in both target and external datasets, which may exclude key predictors available only in EHR. Furthermore, we assume covariates are measured consistently across data sources. In this study, we rely on billing codes in the EHR, which are uniformly reported to claims data, ensuring consistency. However, integrating data from other sources — such as RCTs — may introduce measurement issues due to differing collection standards, requiring careful consideration.

3.5 Conclusion

In conclusion, our study addresses an important challenge in EHR – observability. We propose an approach to estimate the observability of a variable and assess potential differential observability in a given target dataset using external data. This approach can be used before developing clinical prediction models to check if the given data is appropriate to be used for analysis. Further work will continue to focus on observability and explore how to improve clinical prediction models when observability issues occur.

4. Predicting Partially Observed Long-Term Outcomes with Adversarial Positive-Unlabeled Domain Adaptation[2]

Predicting long-term clinical outcomes often requires large-scale training data with sufficiently long follow-up. However, in electronic health records (EHR) data, long-term labels may not be available for contemporary patient cohorts. Given the dynamic nature of clinical practice, models that rely on historical training data may not perform optimally. In this work, we frame the problem as a positive–unlabeled domain adaptation task, where we seek to adapt from a fully labeled source domain (e.g., historical data) to a partially labeled target domain (e.g., contemporary data). We propose an adversarial framework that includes three core components: (1) Overall Alignment, to match feature distributions between source and target domains; (2) Partial Alignment, to map source negatives to unlabeled target samples; and (3) Conditional Alignment, to address conditional shift using available positive labels in the target domain. We evaluate our method on a benchmark digit classification task (SVHN-MNIST), and two real-world EHR applications: prediction of one-year mortality post COVID-19, and long-term prediction of neurodevelopmental conditions (NDC) in children. In all settings, our approach consistently outperforms baseline models and, in most cases, achieves performance close to an oracle model trained with fully observed labels.

---

[2] Yan, M., Xia, M., Huang, W., Hong, C., Goldstein, B. A., & Engelhard, M. M. (2025). Predicting Partially Observed Long-Term Outcomes with Adversarial Positive-Unlabeled Domain Adaptation. Manuscript under review.

4.1 Introduction

Most clinical decision support tools are derived from electronic health records (EHRs) data (Alexiuk et al., 2024; Goldstein, Navar, et al., 2017; Sutton et al., 2020). In recent years, machine learning methods have been increasingly used to perform clinical risk prediction for a variety of outcomes, such as mortality, hospitalization, or disease (Naemi et al., 2021; Shamout et al., 2021; Shickel et al., 2018). Given their dense, granular nature, EHR data are well positioned for predicting near-term outcomes (Goldstein, Pencina, et al., 2017). However, predicting long-term health outcomes has gained interest because early detection of risk can enable timely interventions and potentially improve patient prognosis (Nannan Panday et al., 2017).

One challenge in training models to predict long-term outcomes is the need for sufficient follow-up data to capture events far into the future. As such, researchers have typically utilized large epidemiological cohorts to develop clinical risk models for long-term outcomes (D'Agostino et al., 2008). Within an EHR system, however, long-term follow-up can be challenging to obtain due to the fluid nature of patient engagement with clinical care. Even when long-term follow-up is available, an additional challenge arises from potential data shifts, as models trained on historical data may not fully reflect current clinical practices or patient populations. Given the dynamic nature of clinical care, it is often desirable to use more contemporary data for training. This was highlighted during the COVID-19 pandemic, when there was a pressing need to develop models for longer-term outcomes using newly available data, for which there was limited follow-up.

We illustrate this scenario in Figure 4.1. The left panel shows the typical training scenario for a historical population, predicting one-year outcomes in 2020 requires training data spanning the entire preceding year (2019). In contrast, the right panel shows the target population scenario, where the goal is to initiate training in early 2020 despite having only one month of follow-up data, rather than waiting for a full year of observations.

Our setting shares some similarities with the online learning paradigm, which focuses on incrementally updating models as new data arrives over time (Shalev-Shwartz, 2011). In our scenario, however, the primary challenge is not the sequential arrival of data *per se*, but rather the absence of long-term outcome labels for newer data points (*i.e.,* the target domain). In other words, we focus on updating our model for newer patients whose features have been observed, but whose labels have been only partially observed; whereas online learning focuses on continually updating models with new features and labels as they arrive.

This scenario is a special case of the outcome observability problem (Yan et al., 2022), where the absence of a long-term outcome is simply unobserved. Within the machine learning literature, this can be viewed as a positive and unlabeled (PU) problem (Elkan & Noto, 2008), in which the outcomes that are not observed as "positive" are unlabeled rather than definitively negative.

To address the lack of labeled data in a target population, we can leverage information from a different but related population (*e.g.*, a historical cohort in the same EHR system or an external dataset) for which long-term outcomes are already known. As

illustrated in Figure 4.1, while the target population only has data for up to 30 days, a historical cohort may provide labels for one-year mortality.

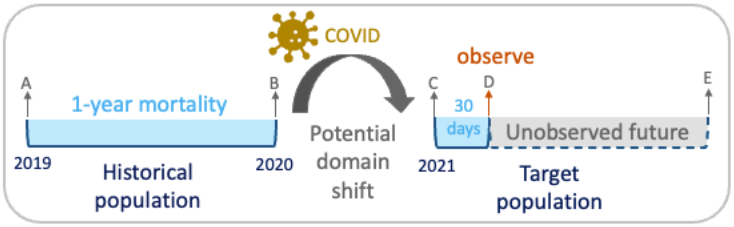

Figure 4.1: Motivating example illustrating the challenges of predicting long-term outcomes with partially observed outcomes.

A-B: outcome observation window for the historical population (fully observed); D: the point at which we aim to deploy the model for the target population; C-D: available outcome observation window for the target population (partially observed); C-E: ideal outcome observation window for the target population.

However, directly applying a model trained on the historical cohort to the target population may yield poor performance due to population shift (Subbaswamy & Saria, 2020). Such shifts can arise over time, across institutions, or among different disease categories, leading to significant bias if the source and target populations differ in their underlying characteristics.

Domain adaptation (DA) strategies are designed to mitigate this problem by adapting a model from a source domain to a related but distinct target domain (Ben-David et al., 2010). By aligning feature distributions or representations across domains, DA can help existing models perform better on the new population of interest. Despite extensive research on domain adaptation in fields such as computer vision and natural language processing, directly applying these approaches to tabular EHR data can be challenging.

For instance, contrastive-learning based DA methods have achieved success in vision tasks (Singh, 2021; Thota & Leontidis, 2021) by learning domain-invariant

representations through drawing positive pairs close in representation space while pushing negative pairs apart. They align distributions between source and target domains by ensuring that similar examples remain close despite domain shifts. However, positive pairs are typically constructed through image augmentations—a practice that does not translate easily to the structure of tabular EHR data.

Adversarial learning-based methods, such as Domain-Adversarial Neural Networks (DANN) (Ganin et al., 2017) and Adversarial Discriminative Domain Adaptation (ADDA) (Tzeng et al., 2017), offer a more promising route for tabular EHR settings. Indeed, several EHR risk prediction studies have successfully explored adversarial techniques (Yu et al., 2021; T. Zhang et al., 2022; Y. Zhang et al., 2019) for domain adaptation in unsupervised settings. However, Zhang et al. (2022) and Zhang et al. (2019) do not account for the positive–unlabeled nature of our problem, and Yu et al. (2021) focus on generating adversarial samples to improve domain adaptation, rather than using adversarial training to align features between the source and target domains.

In this work, we seek a method to adapt a model from a historical source cohort to a contemporary target cohort while accounting for the partial observability of outcomes in the target data. To achieve these goals, we propose an adversarial domain adaptation framework tailored to PU data in EHRs. Our method features three core components: (1) An ADDA-based alignment term that match overall feature distributions between source and target populations; (2) A partial alignment term that ensures source negatives have high likelihood under the target unlabeled distribution; (3) A conditional alignment to align the conditional distribution using available positive labels in the target domain.

Our contributions are as follows:

- We formalize the challenge of adapting EHR-based long-term outcome prediction models to temporal distribution shifts under limited follow-up.
- To address this challenge, we propose a novel adversarial positive-unlabeled domain adaptation approach that aligns a fully labeled source domain with a partially labeled target domain.
- We validate our method on one public benchmark and two real-world EHR datasets, demonstrating its practical effectiveness and robustness across diverse clinical settings.

Domain adaptation methods for PU settings remain sparse; we are aware of only one existing approach (Sonntag et al., 2022). Our method therefore constitutes a new adversarial framework with partial distribution matching, providing a robust solution for predicting long-term outcomes in EHR data when labels are only partially observable.

4.2 Method

In this section, we first define our problem (Section 4.2.1), then describe how we pretrain the model on the source data (Section 4.2.2) Finally, we detail three components we used to align target features and source features (Section 4.2.3).

4.2.1 Problem Definition

We frame our problem as a positive and unlabeled (PU) domain adaptation problem. Let the source domain be $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^{n_s}$, where each sample (x_i, y_i) has a fully observed label $y_i \in \{0,1\}$. The target domain is $\mathcal{D}_t = \mathcal{D}_{tp} \cup \mathcal{D}_{tu}$, where $\mathcal{D}_{tp} = \{(x_i, y_i = 1)\}_{i=1}^{n_{tp}}$ contains observed positive samples, and $\mathcal{D}_{tu} = \{(x_i)\}_{i=1}^{n_{tu}}$ contains the unlabeled samples whose true label may be 0 or 1 but is unobserved (and treated as 0).

Our objective is to learn a predictive model for the target domain by leveraging both the fully labeled source domain data ($\mathcal{D}_s$) and the PU target domain data ($\mathcal{D}_t$).

We build upon the Adversarial Discriminative Domain Adaptation (ADDA) framework (Tzeng et al., 2017), which uses adversarial training to reduce the domain discrepancy. In the standard ADDA setting, a source encoder and a source classifier are first trained on labeled source data, then a target encoder is learned such that its feature representation is aligned with that of the source domain via an adversarial objective. However, ADDA assumes that the target data are unlabeled.

In our PU domain adaptation setting, we do have partial supervision on the target domain (i.e., some target positives are observed). Therefore, we introduce additional alignment steps to effectively utilize these partially labeled target examples.

Our method involves: a source encoder M_s, a classifier $\mathcal{F}$, a target encoder M_t, and two discriminators D_1 and D_2 for different alignment objectives.

4.2.2 Source Model Pre-training

We begin with pre-training the source encoder M_s and the classifier $\mathcal{F}$ on the fully labeled source domain. Formally, we minimize the following to parameterize M_s and $\mathcal{F}$:

$$\min_{M_s,\mathcal{F}} \mathcal{L}_{CE}(\mathcal{F}(M_s)),$$

where $\mathcal{L}_{CE}$ is a cross-entropy loss for supervised learning. After this pre-training, we initialize the target encoder M_t using weights of M_s. In many real-world EHR applications, historical or pre-existing models are available and can be directly used for initialization and fixed thereafter.

4.2.3 Feature Alignment

To adapt the model from the source domain to the partially labeled target domain, we align the target and source feature distributions. As depicted in Figure 4.2, our alignment strategy includes three components: (1) Overall Alignment, (2) Partial Alignment and (3) Conditional Alignment.

These components are motivated by the need to align the entire target feature space with the source feature space (Overall Alignment); align the negative portion of the source distribution with the unlabeled portion of the target in a way that accommodates potential hidden positives in the unlabeled set (Partial Alignment); and ensure that the conditional distribution $P(Y \mid X)$ is also well-aligned by leveraging the known target positives (Conditional Alignment).

We train the target encoder M_t adversarially using two discriminators D_1 and D_2 to achieve overall and partial alignment (Sections 4.2.3.1 and 4.2.3.2), respectively. Then incorporate conditional alignment by optimizing M_t with a supervised loss of $\mathcal{F}(M_t(.))$ on $\mathcal{D}_{tp}$. (Section 4.2.3.3). Finally, we summarize the overall objective in Section 4.2.3.4.

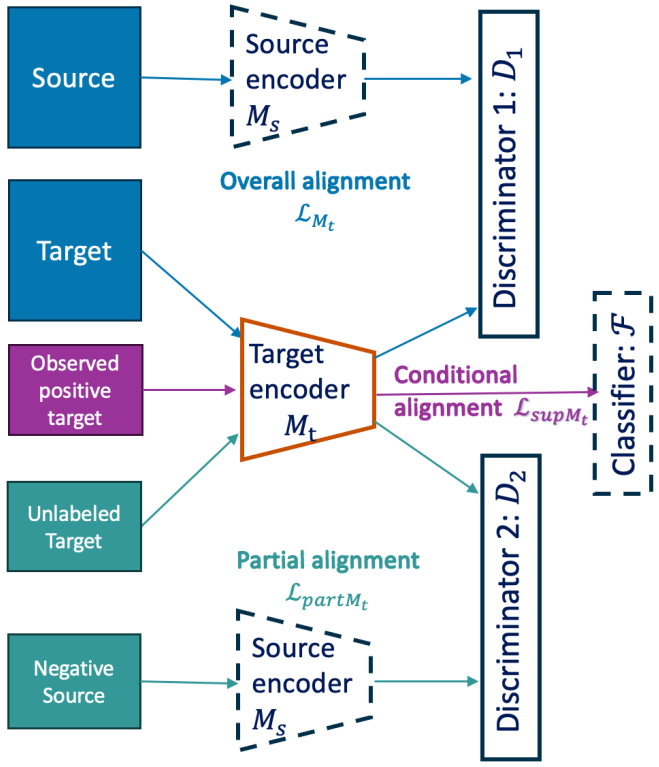

Figure 4.2: Illustration of the proposed framework with three alignment components: overall alignment, partial alignment, and conditional alignment. The source encoder and classifier are from the pre-training step.

4.2.3.1 Overall Alignment

To align the feature distributions, we introduce a discriminator D_1, that learns to distinguish between features from the source encoder M_s and the target encoder M_t. We train D_1 using a binary cross-entropy loss to classify domain indicators (source = 1 vs. target = 0):

$$\mathcal{L}_{D_1} = -\sum_{x \in \mathcal{D}_s} \left[\log D_1\left(M_s(x)\right)\right] - \sum_{x \in \mathcal{D}_t} \left[\log(1 - D_1(M_t(x)))\right].$$

This follows a standard adversarial setup in which D_1 is trained to separate the source and target distributions, while target encoder M_t is trained to prevent D_1 from doing so effectively by matching the source and target distributions in the representation

space. This is done by employing the following inverted-label GAN loss (Tzeng et al., 2015) on the target domain:

$$\mathcal{L}_{M_t} = \sum_{x \in \mathcal{D}_t} [\log D_1(M_t(x))].$$

This is a cross-entropy loss but with inverted domain indicators (target = 1). In other words, we treat the target features as if they were from the source domain to align the overall feature distribution of target with source. Overall alignment helps reduce overall domain discrepancy, ensuring that target features live in a feature space more similar to source features.

4.2.3.2 Partial Distribution Alignment

Although overall alignment brings the entire target distribution closer to the source, it treats all target samples uniformly. In PU learning, however, the unlabeled set $\mathcal{D}_{tu}$ contains both negative and positive samples, but we do not know which are truly positive. Simply forcing all target unlabeled samples to match source negatives can be detrimental, because some unlabeled target samples may be positive.

To address this, we introduce a second discriminator D_2 focusing on aligning source negatives ($\mathcal{D}_{sn}$) with unlabeled target data ($\mathcal{D}_{tu}$), while still permitting positives in the unlabeled set to deviate if needed.

We introduce a KL-divergence-based loss that encourages the negative portion of the source distribution to align with a corresponding portion of the unlabeled target distribution, while allowing the support of the unlabeled target samples, which include positives, to extend beyond that of the source negative samples:

$$\mathcal{L}_{D_2} = KL(p_{sn} \mid p_{tu}) = \mathbb{E}_{x \sim p_{sn}} \left[\log \left(\frac{p_{sn}(x)}{1 - p_{sn}(x)} \right) \right],$$

where p_{sn} and p_{tu} denote the distributions of source negatives and target unlabeled data, respectively.

Ghimire et al. (2021) showed that KL-divergence can be estimated adversarially using a discriminator. In practice, D_2 is parameterized that outputs a log-odds score, which can be used to estimate $\log\left(\frac{p_{sn}(x)}{1-p_{sn}(x)}\right)$. Thus, the KL term can be estimated by

$$\frac{1}{N}\left(\sum_{x\in\mathcal{D}_{sn}} D_2\big(M_s(x)\big) + \sum_{x\in\mathcal{D}_{tu}} D_2\big(M_t(x)\big)\right).$$

We again adopt an adversarial objective for the target encoder M_t, but only on the unlabeled target samples. We use a loss similar to (\ref{eq:GAN_loss}):

$$\mathcal{L}_{partM_t} = \sum_{x\in\mathcal{D}_{tu}} [\log D_2(M_t(x))].$$

Here, we treat the unlabeled target features as if they were from the source negatives, thereby encouraging their representations to match.

4.2.3.3 Conditional Alignment

Even with overall and partial alignment, the conditional distribution $P(Y|X)$ may still differ between source and target domains. This is often referred to as conditional shift, can severely degrade performance if not addressed (Zhao et al., 2019).

To address the conditional shift, we leverage the observed target positives $\mathcal{D}_{tp}$. Specifically, we require M_t to classify these known positives correctly under the source-trained classifier $\mathcal{F}$.

Hence, we introduce a supervised loss on $\mathcal{D}_{tp}$. Since all observed labels $y = 1$, the cross-entropy loss can be written as:

$$\mathcal{L}_{supM_t} = -\sum_{x \in \mathcal{D}_{tp}} [log(\mathcal{F}(M_t(x))]$$

This ensures that the learned target representations produce strong positive predictions in the target domain. Such conditional alignment leverages the label information in the target domain to reduce model shift and improve classification performance.

4.2.3.4 Overall Training Objective

Putting it all together, the overall training process consists of training two discriminators (D_1, D_2) and training the target encoder M_t jointly. We set the parameters of D_1 and D_2 to minimize $\mathcal{L}_{D_1}$ and $\mathcal{L}_{D_2}$, and set the parameters of M_t to minimize the combined loss $\mathcal{L}_{combM_t}$ for M_t that includes an inverted-label GAN loss on all target, an inverted-label GAN loss on target unlabeled, and a supervised loss on target positives:

$$\mathcal{L}_{combM_t} = \lambda_1 \mathcal{L}_{M_t} + \lambda_2 \mathcal{L}_{partM_t} + \lambda_3 \mathcal{L}_{supM_t}.$$

The proposed framework is shown in Figure 4.2. By building upon ADDA and incorporating partial distribution alignment for negative classes and conditional alignment via limited positive target labels, our framework addresses the unique challenges of PU domain adaptation.

4.3 Data Applications

To illustrate our method, we first conduct a *proof of concept* analysis on the SVHN and MNSIT datasets. We then move to our two motivating real-world EHR based examples. We compare the proposed method to: (1) An oracle model that trains on target data with fully observed labels; (2) A baseline model trained exclusively on the (historical) source data without adaptation; (3) A model trained only on the partially observed target data; (4) Unsupervised ADDA.

To assess the contributions of each alignment component, we also conduct an ablation study for each application. We compare several variants of our proposed method: (1) Only using overall alignment $\mathcal{L}_{M_t}$ (ADDA)); (2) Only using overall and partial alignments ($\mathcal{L}_{M_t} + \mathcal{L}_{partM_t}$); (3) Only using overall and conditional alignments ($\mathcal{L}_{M_t} + \mathcal{L}_{supM_t}$); (4) Only using conditional alignments ($\mathcal{L}_{supM_t}$); (5) Using all three components ($\mathcal{L}_{M_t} + \mathcal{L}_{partM_t} + \mathcal{L}_{supM_t}$).

4.3.1 Proof of Concept: Digits Classification

We first illustrate our proposed method on a well-known domain adaptation task involving the SVHN and MNIST datasets. We use the SVHN dataset as the source domain and a stylized variant of MNIST (drawn from MNIST-C (Mu & Gilmer, 2019)) as the target domain. We randomly assign half of the MNIST target samples to the canny-edge style and the other half to the motion-blur style.

4.3.1.2 Data setup

Since we focus on a binary classification task, we restrict our dataset to two digits: 3 (positive) and 5 (negative). We then create a PU setting in the target MNIST data by assigning observed positive labels with probability p_{obs1} for the canny-edge images, and p_{obs2} for the motion-blur images. This design generates different label-selection mechanisms, allowing us to test the robustness of our method under varying degrees of label observability.

4.3.1.2 Results

Table 4.1 shows the AUC averaged over 10 runs. Despite having very few labeled samples in the target domain, our proposed method achieves the highest AUC across all three label observability conditions, significantly outperforming baseline methods.

Table 4.1: AUC for digit classification (3vs.5) task on stylized MNIST data.

Obs1 and Obs2 are labeling prob- abilities for canny edges and motion blur styles, respectively. Oracle model: trained on MNIST with fully observed labels; Source model: trained on SVHN without adaptation; Target model: trained on the partially observed MNIST. ADDA: unsupervised baseline DA method. Bolded values indicate the best performance for each scenario.

Model		Oracle (full obs)	Source (w/o adapt)	Target (part obs)	ADDA	Proposed
Obs1	Obs2					
10%	1%	0.999	0.908	0.916	0.956	**0.980**
10%	5%	0.999	0.908	0.933	0.951	**0.991**
1%	1%	0.999	0.908	0.816	0.935	**0.975**

4.3.2 Real Data Application 1: One-year Mortality Prediction After COVID-19

Our first real-world application aims to predict one-year mortality following emergency department (ED) visits, using data shortly after the COVID-19 pandemic commenced. This is a critical task given that ED patients often exhibit elevated risk of mortality (Gunnarsdottir & Rafnsson, 2006) and early identification of high-risk patients can facilitate timely interventions and potentially improve clinical outcomes (Dundar et al., 2016; Hung et al., 2017). While ample historical is available, given the changing nature of clinical care after COVID-19, it is reasonable to presume that there was meaningful domain-shift, therefore any model implemented in current clinical practice should be trained primarily on *post* COVID-19 data (Figure 4.1).

4.3.2.1 Data setup

We use data from our institution's EHR system to define two cohorts: a historical source cohort (all ED encounters in 2018) with fully observed one-year mortality, and a target cohort (all ED encounters in 2021). Although we have full observability for the target cohort here, we simulate situations such that we only observe shorter outcomes in the target data, as if it were immediately after the COVID-19 outbreak. Specifically, we set up the target data to observe 7-day, 30-day, and 90-day mortality, which correspond to 17.3%, 31.1% and 51.7% actuarial outcomes respectively. Predictors including demographics, vital signs and disease diagnosis, with a total of 107 features. We selected a sample size of 20,000 from each cohort as the target training data and source training data, and an additional 10,000 samples from the 2021 cohort as the target test data, ensuring no overlap in patients between the training and test sets.

4.3.2.1 Results

Table 4.2 summarizes the AUC results for predicting one-year mortality after an ED encounter under various approaches and observation windows. As a reference, if we had access to the full one-year outcome data in the target, the best performing model (i.e., oracle model) would achieve an AUC of 0.907. Directly applying the model trained solely on historical data yields an AUC of 0.869, reflecting the distribution mismatch between source and target. Training on the partially observed target data alone (i.e., using only the first 90, 30, or 7 days of outcomes) gives lower AUC than the oracle model. Nonetheless, performance improves as the outcome observation window increases (from 7 days to 90 days). Using the unsupervised ADDA alignment boosts performance,

demonstrating the benefit of adversarial domain adaptation for this task. Our proposed approach achieves the highest AUC across all observation windows.

Table 4.2: AUC results for one-year mortality prediction on target data, given partial observations of the target outcome (90 days, 30 days, or 7 days).

Event rates and outcome observability levels for each scenario are reported. Oracle model: trained on post-COVID target data with fully observed labels; Historical source model: trained on historical data without adaptation; Target model: trained exclusively on the partially observed target. ADDA: unsupervised baseline DA method. Bolded values indicate the best performance for each scenario.

Model			Oracle (full obs)	Historical (w/o adapt)	Target (part obs)	ADDA	Proposed
Event Rate	Obs	Observe window					
3.6%	51.7%	90 days	0.907	0.869	0.883	0.885	**0.909**
2.2%	31.1%	30 days	0.907	0.869	0.858	0.885	**0.903**
1.2%	17.3%	7 days	0.907	0.869	0.820	0.885	**0.900**

Notably, with just 90 days of outcome data—about half of the full observation period—we already attain an AUC of 0.909, which is close to and even slightly above the 0.907 achieved by the oracle model. This indicates that partial and conditional alignment, in addition to overall alignment, effectively transfers knowledge from the historical source data and leverages the labels in target data.

Overall, these results demonstrate that our proposed method with only limited outcome can be close to the performance of an oracle model trained with full one-year outcome labels in the target data.

4.3.3 Real Data Application 2: Long-term prediction of neurodevelopmental conditions

Neurodevelopmental conditions (NDCs) include a range of conditions such as autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), motor delay (MD), language delay (LD), and developmental delays (DD) that typically manifest

in early childhood (Thapar et al., 2017). Early detection of these conditions is crucial for initiating services to impact long-term outcomes (Cioni et al., 2016). In this application, our goal is to use clinical features available at 18 months old to predict whether a child will have an NDC diagnosis by 5.5 years of age (typical school age and 4 years in the future). Given the long-term nature of the outcome, to properly train a model in 2024, we could only use data as late as 2019. Ideally, we want to incorporate more contemporary data that captures evolving diagnostic standards and care practices. The proposed method will allow us to train the model with more contemporary yet partially observed outcome data.

4.3.3.1 Data setup

Following the cohort definition and feature engineering in Huang et al. (2024), we use embeddings of diagnosis and procedure codes extracted from the first 18 months of life. We divide the dataset into a historical source population (children born in 2014) and a target population (children born in 2017). We assume that, as of the current time, we only observe outcomes for up to 6 months, 1 year, or 2 years from the index date in the target population. These observation windows correspond to 14.3%, 30.4%, and 46.7% of full label availability, respectively.

4.3.3.2 Results

As shown in Table 4.3, models using historical source data without adaptation and model trained on partially observed target data both fall notably short of the oracle performance. Baseline unsupervised ADDA does improve the AUC to 0.786 as it aligns the overall features of target with the historical data. Our proposed method gives the best AUC among all three limited observability scenarios. Even in the most limited scenario

(6 months of available outcome, i.e., 14.3% of the full data), the proposed model obtains an AUC of 0.813, nearly matching the oracle model's 0.818.

Table 4.3: AUC results for long-term NDC prediction on target data, given partial observations of the target outcome (6 months, 1 year, 2 years).

Event rates and outcome observability levels for each scenario are reported. Oracle model: trained on target data with fully observed labels; Historical source model: trained on historical data without adaptation; Target model: trained exclusively on the partially observed target. ADDA: unsupervised baseline DA method. Bolded values indicate the best performance for each scenario.

Model			Oracle (full obs)	Historical (w/o adapt)	Target (part obs)	ADDA	Proposed
Event Rate	Obs	Observe window					
10.0%	46.7%	2 years	0.818	0.779	0.772	0.786	**0.814**
6.5%	30.4%	1 years	0.818	0.779	0.755	0.786	**0.813**
3.1%	14.3%	6 months	0.818	0.779	0.724	0.786	**0.813**

4.3.4 Ablation study

For SVHN-MNIST example, we evaluate different components of the proposed method in Table 4.4. We report overall AUC and subgroup AUC for canny edges and motion blur styles. As shown in Table 4.4 adding partial alignment on top of overall alignment ($\mathcal{L}_{M_t} + \mathcal{L}_{partM_t} + \mathcal{L}_{supM_t}$) generally boosts performance, especially in cases with higher positive-label observability. Adding conditional alignment ($\mathcal{L}_{M_t} + \mathcal{L}_{supM_t}$ or $\mathcal{L}_{M_t} + \mathcal{L}_{partM_t} + \mathcal{L}_{supM_t}$) further increases both the overall and subgroup AUC, leveraging the small pool of known positives. However, in these cases when observability is extremely low, conditional alignment alone may not help because the very small number of positive labels can lead the model to overemphasize the positive class. Overall, the combination of all three components (overall, partial, and conditional alignment) yields the best performance.

Table 4.4: AUC comparison across different components for digits classification.

Subgroup performance is re- ported for canny-edge and motion-blur styles. $\mathcal{L}_{M_t}$: overall alignment; $\mathcal{L}_{partM_t}$: partial alignment; $\mathcal{L}_{supM_t}$: conditional alignment.

Model		$\mathcal{L}_{M_t}$(ADDA)	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{sup}M_t}$	$\mathcal{L}_{\text{sup}M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t} + \mathcal{L}_{\text{sup}M_t}$
Obs	Style					
		0.956	0.965	0.981	0.5	0.980
10%	canny edges	0.934	0.957	0.984	0.5	0.989
1%	motion blur	0.973	0.975	0.969	0.5	0.972
		0.951	0.963	0.987	0.5	0.991
10%	canny edges	0.934	0.960	0.994	0.5	0.995
5%	motion blur	0.971	0.964	0.981	0.5	0.987
		0.935	0.955	0.966	0.5	0.975
1%	canny edges	0.902	0.939	0.969	0.5	0.989
1%	motion blur	0.969	0.973	0.963	0.5	0.967

The AUC results for prediction one-year mortality and predicting long-term NDC in Table 4.5 show that in contrast to the digit classification task, including our partial alignment term in addition to overall alignment does not substantially improve performance. This may be because overall adversarial alignment already reduces much of the distribution mismatch, and in contrast to the classification task, this is a risk prediction setting in which positive and negative distributions are not separable. On the other hand, adding conditional alignment significantly improves performance over overall alignment alone. Enforcing correct classification of known positives in the target domain helps reduce label shift and results in higher AUC. However, using conditional alignment in isolation is not sufficient: Without adversarial alignment to anchor the representations, the model can overemphasize the positive class, harming performance on negatives. Apart from AUC, we observed that adding conditional alignment can increase the calibration-in-the-large in Appendix C1. Because conditional alignment shifts predictions toward positive outcomes, this underscores the importance of balancing all alignment components.

Table 4.5: AUC comparison in the ablation studies for one-year mortality prediction and long-term NDC prediction with different observation windows.

$\mathcal{L}_{M_t}$: overall alignment; $\mathcal{L}_{partM_t}$: partial alignment; $\mathcal{L}_{supM_t}$: conditional alignment.

Model	$\mathcal{L}_{M_t}$(ADDA)	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{sup}M_t}$	$\mathcal{L}_{\text{sup}M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t} + \mathcal{L}_{\text{sup}M_t}$
1-year mortality					
Observe window					
90 days	0.885	0.887	0.908	0.834	0.909
30 days	0.885	0.885	0.905	0.836	0.903
7 days	0.885	0.885	0.903	0.842	0.900
long-term NDC					
Observe window					
2 years	0.786	0.786	0.814	0.587	0.805
1 year	0.786	0.786	0.813	0.599	0.804
6 months	0.786	0.783	0.813	0.604	0.798

4.4 Discussion

In this paper, we address the challenge of developing a clinical prediction model for long-term outcomes before the long-term outcomes are fully observed in the target data. This question is particularly relevant when one wants to account for changing clinical practices over time. We formulated this setting as a positive–unlabeled domain adaptation problem, which incorporates the dual tasks of adaptation across source and target domains to mitigate population shift, and partial label observability to handle unlabeled data that may still be positive in the future.

To tackle these challenges, we proposed a new adversarial framework that extends ADDA with three key components: (1) Overall Alignment to align the overall feature distribution between source and target; (2) Partial Alignment to match source negatives with the unlabeled portion of target data; and (3) Conditional Alignment to directly utilize the limited positive labels in the target domain.

Through experiments on a benchmark digit classification task (SVHN to stylized MNIST) and two real-world EHR applications (one-year mortality prediction and long-

term NDC prediction), we demonstrated that the proposed method consistently outperforms baseline models. Notably, our approach achieves performance comparable to a model trained with fully observed labels (oracle model) in most scenarios.

In practice, this framework can be integrated into existing clinical workflows, allowing health systems to use historical cohorts to train robust models that predict long-term outcomes even when the current population only has short-term outcome data available. By aligning feature representations and considering available positive labels, our method can facilitate long-term risk prediction and potentially improve patient outcomes.

Although our motivating example focuses on using historical data, the proposed method can be generalized to consider any fully labeled *source* dataset, applied to a partially observed *target* dataset. For example, we may have source data emanating from an external cohort such as larger payer databases, clinical trials, or epidemiological cohorts. In these settings, it is reasonable to assume that there is some degree of domain shift to a target health system data context. Moreover, if the outcome can take place over an extended observation window, it is unlikely to be fully observed. This work therefore fits into a larger framework regarding the use of external data sources to address observability challenges in EHR-based datasets (Yan et al., 2025).

Despite promising results, our approach has two primary limitations. First, using conditional alignment can lead to larger calibration error, therefore additional re-calibration strategies may be needed. Second, achieving balance among the three alignment components is non-trivial, and determining the optimal weight requires careful

tuning. Future research could explore more robust and systematic methods for achieving this balance.

Overall, this work highlights the importance of domain adaptation and positive–unlabeled learning as complementary strategies for bridging the gap between historical datasets and current patient populations in need of timely, long-term risk predictions. We hope our findings encourage further exploration of adaptive techniques that can make the most of limited outcome labels in clinical practice.

5. Conclusion

This dissertation has explored how incomplete and differentially captured outcomes in electronic health records (EHRs)—an issue we term as *observability*—can affect the development and fairness of clinical prediction models (CPMs). The central objective was to examine how observability arises, how to estimate and assess it, and how to build robust CPMs even when many outcomes remain unobserved.

We began by proposing a formal definition of differential bias to demonstrate that differential observability, where some subgroups are systematically less likely to have their outcomes captured in EHR, can induce algorithmic bias in prediction models. Through simulations and empirical examples, we showed that this bias depends on whether a *sensitive variable* (e.g., socioeconomic status or race/ethnicity) is included in the model. If that variable directly affects the true outcome, excluding it can worsen the bias; if it primarily affects *observability* rather than the outcome itself, including it may actually exacerbate bias. Thus, understanding the mechanism behind differential observability is critical to deciding whether and how to incorporate sensitive variables into CPMs.

Next, recognizing that observability is not always directly quantifiable in a given EHR system, we proposed a novel method to estimate how much of a target outcome is likely to be unobserved. By comparing EHR-derived outcomes to a fully observed external data source, and using a balancing score to reweight the external data to resemble the EHR population, this method provides practical estimates for both overall and differential observability. Simulation studies showed that the proposed estimator has a low MAE and high coverage rate if all relevant confounders are included in the

balancing score model. An illustrative application to hospital readmission data further showed how to interpret a differential observability test to assess whether two subgroups differ significantly in observability.

Finally, we addressed how to develop robust CPMs in situations where outcomes are partially observed. We focused on specific observability scenario that long-term outcome is not yet fully observed for contemporary cohort, though our method can be generalized to any partially observed target dataset and fully labeled source dataset. Framing the observability problem as positive and unlabeled (PU) data problem, we proposed an adversarial domain adaptation framework that leverages fully labeled historical data. By aligning feature representations of a contemporary, partially labeled population with a historical cohort, our method adapts CPMs to shifting patient populations while accounting for observability. In real-world scenarios, health systems can deploy this framework to train reliable models for long-term outcomes, even when only short-term outcomes are available for newly enrolled patients. Moreover, because the method supports Selection At Random (SAR)—a situation in which outcome observability depends on subgroup membership or time—it can improve performance even under differential observability conditions. Therefore, this approach holds promise for a wide range of EHR-based CPM applications that involve partially observed outcomes.

An important implication of this work is its contribution to fairness and equity in healthcare. By quantifying and addressing differential observability, practitioners can better detect when certain subgroups are underrepresented in the data and thus at risk of poorer model performance. Our proposed methods can be integrated into the clinical

prediction modeling workflow: observability estimation method can be used prior to model development (using external data) to ensure the dataset is fit for robust analysis; and if observability issues arise, adapting model through domain adaptation with historical data or external data so that CPMs remain well performed.

While the methods and results in this work contribute significantly to addressing observability in EHR, several limitations highlight opportunities for further research. First, although this dissertation focuses on outcome observability in CMPs in this dissertation, it is important to recognize that observability constitutes only one of many potential biases in EHR-based research. Other sources, such as differences in measurement of covariates, changes in coding practices, shifting clinical guidelines, and incomplete clinical documentation, can also affect CPM performance.

Second, while we illustrate the proposed observability estimation method in the context of CMP, its derivation does not depend on any predictive farmwork. As such, it could be equally useful for assessing observability in association or causal analyses. Moreover, CPMs are usually deployed in clinical decision making with specific risk thresholds that determine whether a patient is targeted for intervention. Future work could explore how observability affects these decision rules and consequently clinical outcomes.

A third limitation lies in our reliance on having historical or external data with more complete outcome capture, although such data is available in our examples, it may may not hold in all settings. In instances where fully observed data are unavailable or difficult to obtain, alternative strategies such as self-supervised methods can be explored. Additionally, leveraging data from different modalities can also be a promising direction.

Finally, while our adversarial domain adaptation framework proves effective in handling positive–unlabeled (PU) data and adapting to population shifts, real-world healthcare systems often benefit from continuous model monitoring and updating. Extending these methods to dynamically evolving clinical environments remains an important direction for future research. By fostering a bias-aware mindset throughout model development and deployment, health systems can better mitigate observability and related biases, ultimately enhancing both the accuracy and fairness of EHR-based clinical decision support.

In summary, by explicitly defining observability, developing methods to estimate it, and proposing modeling strategies that accommodate partially observed outcomes, this work advances the reliability and equity of EHR-based clinical prediction models. Through careful application of these techniques, healthcare systems can better identify, measure, and mitigate the biases caused by observability inherent in routinely collected EHR data, ultimately improving patient care and health outcomes.

Appendix A.

Observability and its Impact on Differential Bias for Clinical Prediction Models

A.1 Correlation between X_1 and X_s

We conducted simulations when X_1 and X_s are correlated. Figure A1, A2 and A3 show the heatmap with different correlations. With a higher correlation between X_1 and X_s, there is more differential bias when including the sensitive variable (Figure A3). But the general patterns of $|\Delta_{in}Bias| - |\Delta_{ex}Bias|$ are similar to the pattern in Figure 4 when X_1 and X_s have weak to medium correlations (Figure A1 and A2). In these cases, the correlation does not meaningfully affect the results.

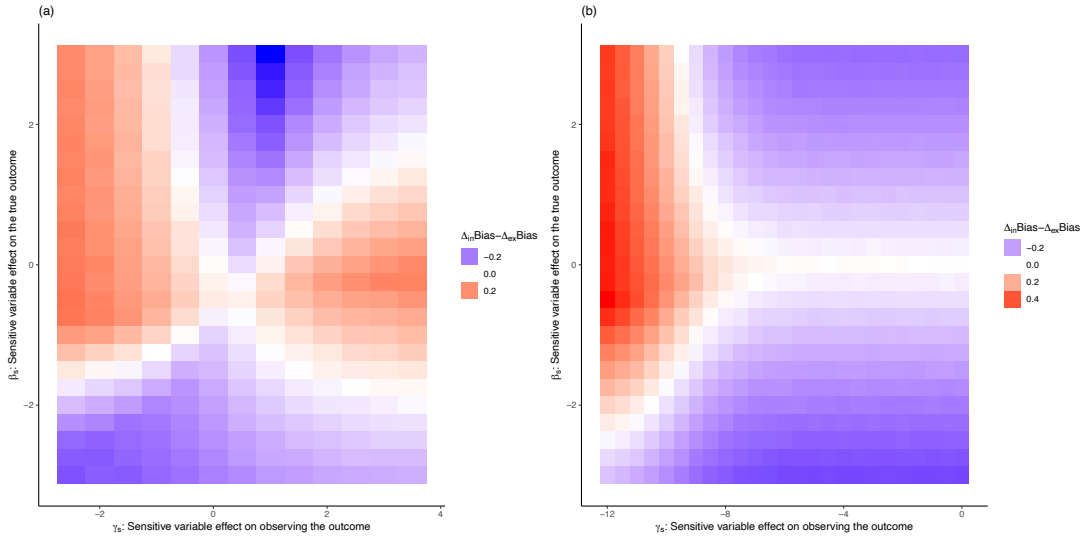

Figure A1: Heatmaps of $|\Delta_{in}\mathbf{Bias}| - |\Delta_{ex}\mathbf{Bias}|$ when $\boldsymbol{\beta_s}$ and $\boldsymbol{\gamma_s}$ vary. $\mathbf{Correlation}(X_1, X_s) = 0.25$. Left(4a): Probability of observing outcome for reference group is set to be 0.5 ($\gamma_0 = 0$), $P(Y|E = 1, X_s = 0) = 0.5$. Right(4b): $\gamma_0 = 10$, $P(Y|E = 1, X_s = 0) = 1$.

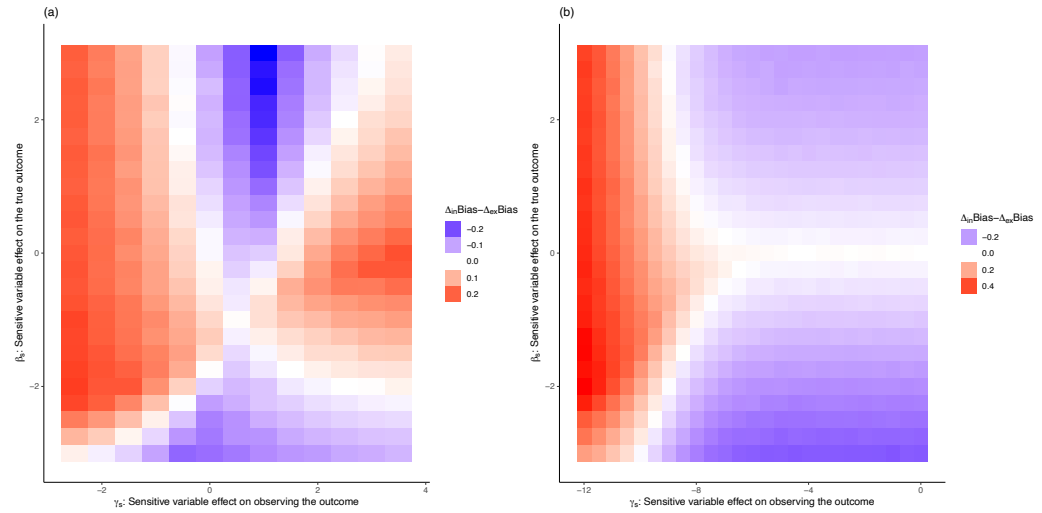

Figure A2: Heatmaps of $|\Delta_{in}\text{Bias}| - |\Delta_{ex}\text{Bias}|$ when β_s and γ_s vary. $\text{Correlation}(X_1, X_s) = 0.5$. Left(4a): Probability of observing outcome for reference group is set to be 0.5 ($\gamma_0 = 0$), $P(Y|E = 1, X_s = 0) = 0.5$. Right(4b): $\gamma_0 = 10$, $P(Y|E = 1, X_s = 0) = 1$.

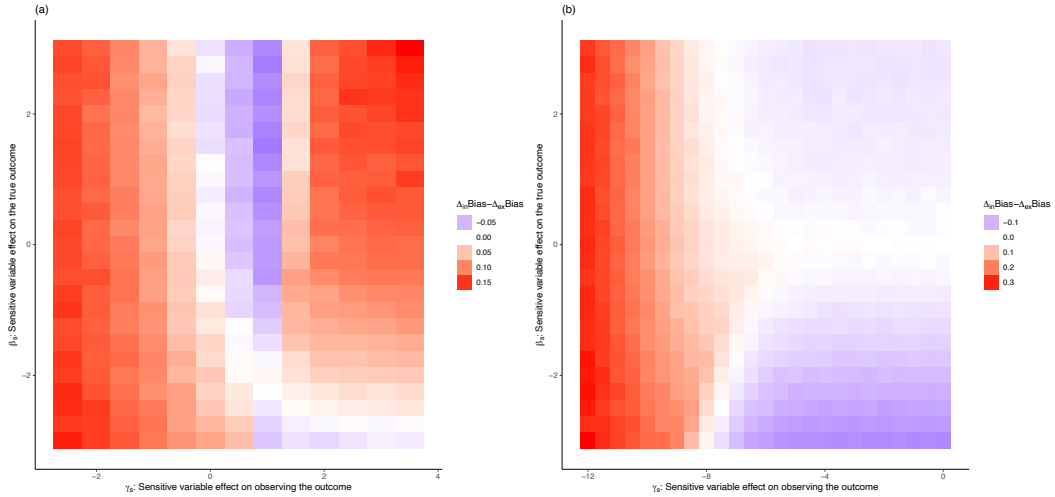

Figure A3: Heatmaps of $|\Delta_{in}\text{Bias}| - |\Delta_{ex}\text{Bias}|$ when β_s and γ_s vary. $\text{Correlation}(X_1, X_s) = 0.8$. Left(4a): Probability of observing outcome for reference group is set to be 0.5 ($\gamma_0 = 0$), $P(Y|E = 1, X_s = 0) = 0.5$. Right(4b): $\gamma_0 = 10$, $P(Y|E = 1, X_s = 0) = 1$.

A.2 Continuous X_s

For continuous cases, the following simulation is conducted. X_s is generated from Uniform(0,1). We compared predicted risk to the true risk as X_s varies.

Table A1: Simulation Table for continuous X_s

Simulation	β_0	β_1	β_s	γ_0	γ_s
7	0	2	1	2	0
8	0	2	0	10	-11

If the sensitive variable X_s is continuous, the results are consistent with the binary case. When there is no differential observability issue (Figure A4a), differential bias is smaller when including X_s in the model. Because in this case, the difference between true and predicted risk changes less as X_s changes and the differential bias is smaller. Conversely, when X_s only impacts observability (Figure A4b), differential bias is smaller when excluding X_s.

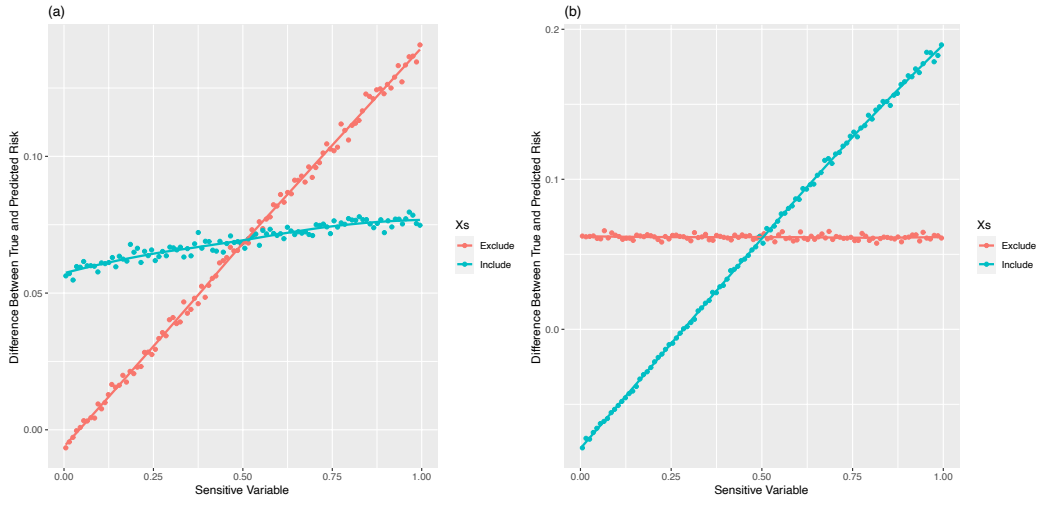

Figure A4: Sensitive variable X_s versus the difference between true risk and predicted, either include or exclude X_s in the model. (a) When there is no differential observability

issue ($\boldsymbol{\beta_s = 1, \beta_0 = 0, \beta_1 = 2, \gamma_s = 0, \gamma_0 = 2}$). (b) When $\boldsymbol{X_s}$ only impacts observing outcome ($\boldsymbol{\beta_s = 0, \beta_0 = 0, \beta_1 = 2, \gamma_s = -11, \gamma_0 = 10}$).

Appendix B.

Estimating the Observability of an Outcome from an Electronic Health Records Dataset Using External Data

B.1 Additional Simulations

In practice, external datasets often have larger sample sizes compared to the target dataset. To reflect this scenario, we conducted additional simulations where the external dataset's sample size is ten times that of the target dataset. The setup is similar to the simulations described in the main text, except that $n_0 = 10{,}000$ and $n_1 = 1{,}000$. Results in Figure B1 and 2 show that, even with different sample sizes, when a confounder is included in the balancing score model (BSM3, BSM5, BSM6), the MAEs are still close to 0 even under poor overlap scenarios. Although the 95% coverage rate for the estimated true observability is slightly lower than in cases with better sample balance, the coverage rate for the estimated observability remains high. These findings suggest that our proposed method continues to provide reliable estimations of observability, even under less ideal conditions.

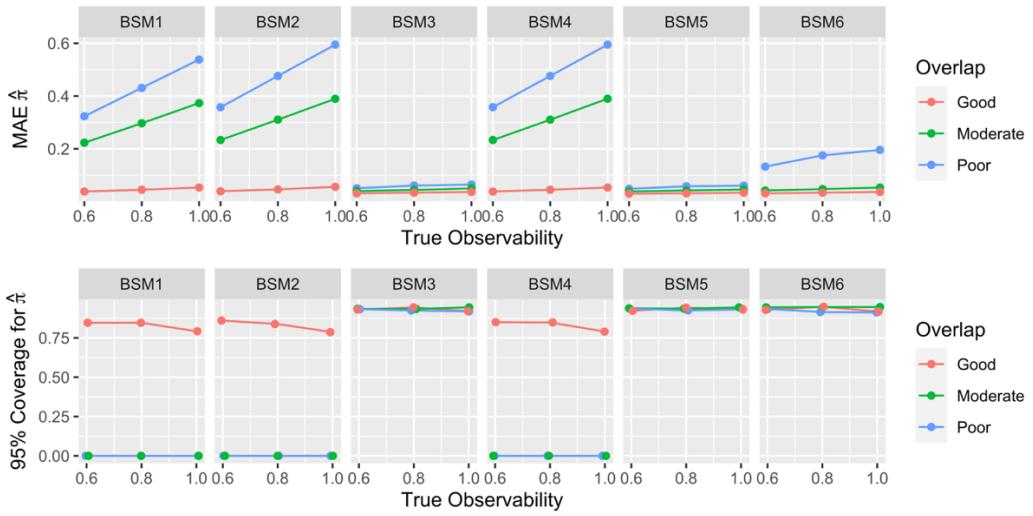

Figure B1: Simulation results for estimated observability using different covariate sets in balancing score models under different overlap levels and different true observabilities with 10:1 sample size ratio between external ($n_0 = 10,000$) and target data ($n_1 = 1,000$).

BSM1: Outcome predictor only model; BSM2: Imbalanced covariate only model; BSM3: "confounder" only model; BSM4: Outcome predictor & imbalanced covariate model; BSM5: Outcome predictor & "confounder" model; BSM6: Imbalanced covariate & "confounder" model.

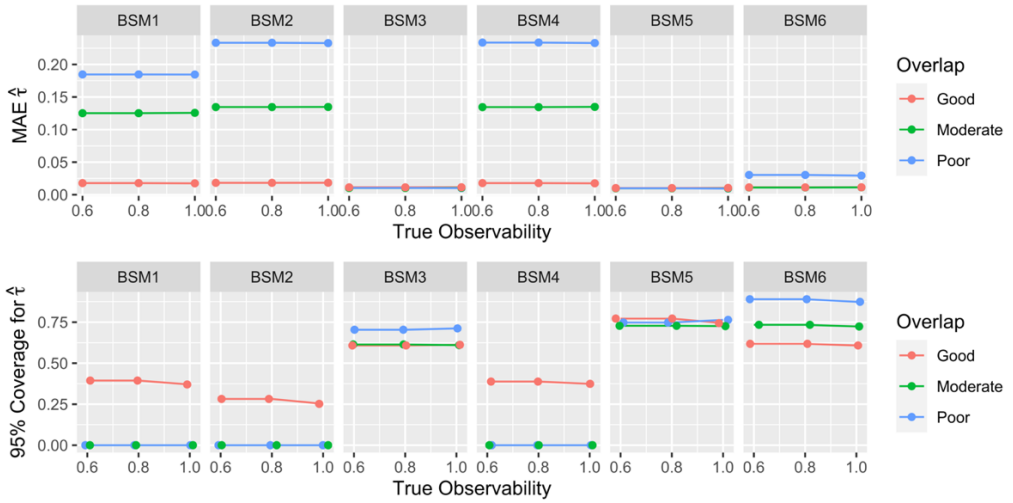

Figure B2: Simulation results for estimated true outcome probability using different covariate sets in balancing score models under different overlap levels and different true observabilities with 10:1 sample size ratio between external ($n_0 = 10,000$) and target data ($n_1 = 1,000$). BSM1: Outcome predictor only model; BSM2: Imbalanced covariate only model; BSM3: "Confounder" only model; BSM4: Outcome predictor & imbalanced covariate model; BSM5: Outcome predictor & "confounder" model; BSM6: Imbalanced covariate & "confounder" model.

B.2 Plasmode Experiment

The plasmode simulation is based on the ESRD dataset used in the Real Data Application section. It preserves the 270 covariates from the real data study, with plasmode data sampled to produce source and target sample sizes of 10,000 and 1,000, respectively, using sampling with replacement. A logistic model was then applied to generate true outcomes in both data sources, followed by setting true observability levels at 1, 0.8, and 0.6 to simulate observed outcomes in the EHR. To estimate observability, we used LASSO model for balancing score model as the one in Real Data Application.

Table B1: presents the mean absolute error (MAE) of the estimated observability and the coverage rate of the confidence intervals. The results indicate that, even in correlated and higher-dimensional cases, the proposed estimation method maintains strong performance with low MAE and high coverage rate.

Table B1: MAE and coverage rate for 95% CI of estimated observability under different true observability based on plasmode simulation data over 100 iterations

True observability	True outcome rate in EHR	Observed outcome rate in EHR	True outcome rate in claims	Mean estimated observability	MAE	Coverage rate for 95% CI
1	0.332	0.332	0.317	1.125	0.124	0.926
0.8	0.332	0.264	0.317	0.898	0.098	0.934
0.6	0.332	0.199	0.317	0.676	0.076	0.922

Appendix C.

C.1 Calibration Results

We also evaluated calibration-in-the-large for both one-year mortality prediction and long-term NDC prediction using different variants of the proposed method. Table C1 shows that adding conditional alignment can increase mean calibration in these two risk prediction tasks. Because conditional alignment shifts predictions toward positive outcomes, these results underscore the importance of balancing all alignment components.

Table C1: Mean calibration-in-the-large comparison for one-year mortality prediction and long-term NDC prediction with different observation windows.

$\mathcal{L}_{M_t}$: overall alignment; $\mathcal{L}_{partM_t}$: partial alignment; $\mathcal{L}_{supM_t}$: conditional alignment.

Model	$\mathcal{L}_{M_t}$(ADDA)	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{sup}M_t}$	$\mathcal{L}_{\text{sup}M_t}$	$\mathcal{L}_{M_t} + \mathcal{L}_{\text{part}M_t} + \mathcal{L}_{\text{sup}M_t}$
1-year mortality					
Observe window					
90 days	0.028	0.029	0.113	0.742	0.125
30 days	0.028	0.027	0.103	0.667	0.118
7 days	0.028	0.029	0.081	0.576	0.085
Long-term NDC					
Observe window					
2 years	0.004	0.006	0.126	0.776	0.144
1 year	0.004	0.005	0.106	0.759	0.120
6 months	0.004	0.005	0.083	0.766	0.098

References

Ackerman, B., Siddique, J., & Stuart, E. A. (2019). *Transportability of Outcome Measurement Error Correction: From Validation Studies to Intervention Trials*. https://doi.org/10.48550/ARXIV.1907.10722

Alexiuk, M., Elgubtan, H., & Tangri, N. (2024). Clinical Decision Support Tools in the Electronic Medical Record. *Kidney International Reports*, *9*(1), 29–38. https://doi.org/10.1016/j.ekir.2023.10.019

Allen, A. S., Rathouz, P. J., & Satten, G. A. (2003). Informative Missingness in Genetic Association Studies: Case-Parent Designs. *The American Journal of Human Genetics*, *72*(3), 671–680. https://doi.org/10.1086/368276

Anderson, M. A., Malhotra, A., & Non, A. L. (2021). Could routine race-adjustment of spirometers exacerbate racial disparities in COVID-19 recovery? *The Lancet. Respiratory Medicine*, *9*(2), 124–125. https://doi.org/10.1016/S2213-2600(20)30571-3

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679. https://doi.org/10.1002/sim.6607

Barda, N., Yona, G., Rothblum, G. N., Greenland, P., Leibowitz, M., Balicer, R., Bachmat, E., & Dagan, N. (2021). Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association: JAMIA*, *28*(3), 549–558. https://doi.org/10.1093/jamia/ocaa283

Bekker, J., & Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Machine Learning*, *109*(4), 719–760. https://doi.org/10.1007/s10994-020-05877-5

Bekker, J., Robberechts, P., & Davis, J. (2020). Beyond the Selected Completely at Random Assumption for Learning from Positive and Unlabeled Data. In U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, & C. Robardet (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 11907, pp. 71–85). Springer International Publishing. https://doi.org/10.1007/978-3-030-46147-8_5

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, *79*(1–2), 151–175. https://doi.org/10.1007/s10994-009-5152-4

Boeschoten, L., van Kesteren, E.-J., Bagheri, A., & Oberski, D. L. (2021). Achieving Fair Inference Using Error-Prone Outcomes. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(5), 9. https://doi.org/10.9781/ijimai.2021.02.007

Brown, L. E., Burton, R., Hixon, B., Kakade, M., Bhagalia, P., Vick, C., Edwards, A., & Hawn, M. T. (2012). Factors influencing emergency department preference for access to healthcare. *The Western Journal of Emergency Medicine*, *13*(5), 410–415. https://doi.org/10.5811/westjem.2011.11.6820

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420010138

Cary, M. P., Zink, A., Wei, S., Olson, A., Yan, M., Senior, R., Bessias, S., Gadhoumi, K., Jean-Pierre, G., Wang, D., Ledbetter, L. S., Economou-Zavlanos, N. J., Obermeyer, Z., & Pencina, M. J. (2023). Mitigating Racial And Ethnic Bias And Advancing Health Equity In Clinical Algorithms: A Scoping Review. *Health Affairs*, *42*(10), 1359–1368. https://doi.org/10.1377/hlthaff.2023.00553

Cioni, G., Inguaggiato, E., & Sgandurra, G. (2016). Early intervention in neurodevelopmental disorders: Underlying neural mechanisms. *Developmental Medicine & Child Neurology*, *58*(S4), 61–66. https://doi.org/10.1111/dmcn.13050

Clinical Classifications Software (CCS) for ICD-9-CM. (n.d.). Retrieved January 4, 2024, from https://hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

Clinical Classifications Software (CCS) for ICD-10-PCS (beta version). (n.d.). Retrieved January 4, 2024, from https://hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp

Cole, S. R., & Stuart, E. A. (2010). Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American Journal of Epidemiology*, *172*(1), 107–115. https://doi.org/10.1093/aje/kwq084

Copeland, K. T., Checkoway, H., McMICHAEL, A. J., & Holbrook, R. H. (1977). Bias Due To Misclassification in the estimation of relative risk. *American Journal of Epidemiology*, *105*(5), 488–495. https://doi.org/10.1093/oxfordjournals.aje.a112408

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care: The Framingham Heart Study. *Circulation*, *117*(6), 743–753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579

Degtiar, I., & Rose, S. (2021). A Review of Generalizability and Transportability. *arXiv:2102.11904 [Stat]*. http://arxiv.org/abs/2102.11904

Dundar, Z. D., Ergin, M., Karamercan, M. A., Ayranci, K., Colak, T., Tuncar, A., Cander, B., & Gul, M. (2016). Modified Early Warning Score and VitalPac Early Warning Score in geriatric patients admitted to emergency department. *European Journal of Emergency Medicine: Official Journal of the European Society for Emergency Medicine*, *23*(6), 406–412. https://doi.org/10.1097/MEJ.0000000000000274

Elixhauser, A. (1996). *Clinical Classifications for Health Policy Research, Version 2: Hospital Inpatient Statistics*. U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research.

Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled data. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*, 213. https://doi.org/10.1145/1401890.1401920

Franklin, J. M., Schneeweiss, S., Polinski, J. M., & Rassen, J. A. (2014). Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex

healthcare databases. *Computational Statistics & Data Analysis*, *72*, 219–226. https://doi.org/10.1016/j.csda.2013.10.018

Galhotra, S., Shanmugam, K., Sattigeri, P., & Varshney, K. R. (2020). Fair Data Integration. *arXiv:2006.06053 [Cs, Stat]*. http://arxiv.org/abs/2006.06053

Gallagher, D., Zhao, C., Brucker, A., Massengill, J., Kramer, P., Poon, E. G., & Goldstein, B. A. (2020). Implementation and Continuous Monitoring of an Electronic Health Record Embedded Readmissions Clinical Decision Support Tool. *Journal of Personalized Medicine*, *10*(3), 103. https://doi.org/10.3390/jpm10030103

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2017). Domain-Adversarial Training of Neural Networks. In G. Csurka (Ed.), *Domain Adaptation in Computer Vision Applications* (pp. 189–209). Springer International Publishing. https://doi.org/10.1007/978-3-319-58347-1_10

Gerych, W., Hartvigsen, T., Buquicchio, L., Agu, E., & Rundensteiner, E. (2022). Recovering the Propensity Score from Biased Positive Unlabeled Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(6), 6694–6702. https://doi.org/10.1609/aaai.v36i6.20624

Ghimire, S., Masoomi, A., & Dy, J. (2021). Reliable Estimation of KL Divergence using a Discriminator in Reproducing Kernel Hilbert Space. *Advances in Neural Information Processing Systems*, *34*, 10221–10233. https://proceedings.neurips.cc/paper_files/paper/2021/hash/54a367d629152b720749e187b3eaa11b-Abstract.html

Goldstein, B. A., Bhavsar, N. A., Phelan, M., & Pencina, M. J. (2016). Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record. *American Journal of Epidemiology*, *184*(11), 847–855. https://doi.org/10.1093/aje/kww112

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health

records data: A systematic review. *Journal of the American Medical Informatics Association: JAMIA*, *24*(1), 198–208. https://doi.org/10.1093/jamia/ocw042

Goldstein, B. A., Pencina, M. J., Montez-Rath, M. E., & Winkelmayer, W. C. (2017). Predicting mortality over different time horizons: Which data elements are needed? *Journal of the American Medical Informatics Association: JAMIA*, *24*(1), 176–181. https://doi.org/10.1093/jamia/ocw057

Goldstein, B. A., Phelan, M., Pagidipati, N. J., & Peskoe, S. B. (2019). How and when informative visit processes can bias inference when using electronic health records data for clinical research. *Journal of the American Medical Informatics Association: JAMIA*, *26*(12), 1609–1617. https://doi.org/10.1093/jamia/ocz148

Gunnarsdottir, O. S., & Rafnsson, V. (2006). Mortality of the users of a hospital emergency department. *Emergency Medicine Journal*, *23*(4), 269–273. https://doi.org/10.1136/emj.2005.026690

Hanchate, A. D., Paasche-Orlow, M. K., Baker, W. E., Lin, M.-Y., Banerjee, S., & Feldman, J. (2019). Association of Race/Ethnicity With Emergency Department Destination of Emergency Medical Services Transport. *JAMA Network Open*, *2*(9), e1910816. https://doi.org/10.1001/jamanetworkopen.2019.10816

Hébert-Johnson, Ú., Kim, M. P., Reingold, O., & Rothblum, G. N. (2018). Calibration for the (Computationally-Identifiable) Masses. *arXiv:1711.08513 [Cs, Stat]*. http://arxiv.org/abs/1711.08513

Heid, I. M., Küchenhoff, H., Wellmann, J., Gerken, M., Kreienbrock, L., & Wichmann, H. E. (2002). On the potential of measurement error to induce differential bias on odds ratio estimates: An example from radon epidemiology: DIFFERENTIAL MEASUREMENT ERROR BIAS. *Statistics in Medicine*, *21*(21), 3261–3278. https://doi.org/10.1002/sim.1252

Heron, S. L., Stettner, E., & Haley, L. L. (2006). Racial and ethnic disparities in the emergency department: A public health perspective. *Emergency Medicine Clinics of North America*, *24*(4), 905–923. https://doi.org/10.1016/j.emc.2006.06.009

Hersh, W. R., Weiner, M. G., Embi, P. J., Logan, J. R., Payne, P. R. O., Bernstam, E. V., Lehmann, H. P., Hripcsak, G., Hartzog, T. H., Cimino, J. J., & Saltz, J. H. (2013). Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Medical Care*, *51*(8 0 3), S30–S37. https://doi.org/10.1097/MLR.0b013e31829b1dbd

Huang, W. A., Engelhard, M., Coffman, M., Hill, E. D., Weng, Q., Scheer, A., Maslow, G., Henao, R., Dawson, G., & Goldstein, B. A. (2024). A conditional multi-label model to improve prediction of a rare outcome: An illustration predicting autism diagnosis. *Journal of Biomedical Informatics*, *157*, 104711. https://doi.org/10.1016/j.jbi.2024.104711

Hung, S.-K., Ng, C.-J., Kuo, C.-F., Goh, Z. N. L., Huang, L.-H., Li, C.-H., Chan, Y.-L., Weng, Y.-M., Seak, J. C.-Y., Seak, C.-K., & Seak, C.-J. (2017). Comparison of the Mortality in Emergency Department Sepsis Score, Modified Early Warning Score, Rapid Emergency Medicine Score and Rapid Acute Physiology Score for predicting the outcomes of adult splenic abscess patients in the emergency department. *PloS One*, *12*(11), e0187495. https://doi.org/10.1371/journal.pone.0187495

Hurst, J. H., Liu, Y., Maxson, P. J., Permar, S. R., Boulware, L. E., & Goldstein, B. A. (2020). Development of an electronic health records datamart to support clinical and population health research. *Journal of Clinical and Translational Science*, *5*(1), e13. https://doi.org/10.1017/cts.2020.499

Johansen, K. L., Chertow, G. M., Gilbertson, D. T., Ishani, A., Israni, A., Ku, E., Li, S., Li, S., Liu, J., Obrador, G. T., Schulman, I., Chan, K., Abbott, K. C., O'Hare, A. M., Powe, N. R., Roetker, N. S., Scherer, J. S., Peter, W. S., Snyder, J., … Wetmore, J. B. (2023). US Renal Data System 2022 Annual Data Report: Epidemiology of Kidney Disease in the United States. *American Journal of Kidney Diseases*, *81*(3), A8–A11. https://doi.org/10.1053/j.ajkd.2022.12.001

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations. *Journal of Research on Educational Effectiveness*, *9*(1), 103–127. https://doi.org/10.1080/19345747.2015.1060282

Kolbrink, B., Schüssel, K., von Samson-Himmelstjerna, F. A., Esser, G., Floege, J., Kunzendorf, U., & Schulte, K. (2023). Patient-focused outcomes after initiation of dialysis for ESRD: Mortality, hospitalization and functional impairment. *Nephrology Dialysis Transplantation*, *38*(11), 2528–2536. https://doi.org/10.1093/ndt/gfad099

Kronbichler, A., Kresse, D., Yoon, S., Lee, K. H., Effenberger, M., & Shin, J. I. (2020). Asymptomatic patients as a source of COVID-19 infections: A systematic review and meta-analysis. *International Journal of Infectious Diseases*, *98*, 180–186. https://doi.org/10.1016/j.ijid.2020.06.052

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, *113*(521), 390–400. https://doi.org/10.1080/01621459.2016.1260466

Lin, K. J., Glynn, R. J., Singer, D. E., Murphy, S. N., Lii, J., & Schneeweiss, S. (2018). Out-of-system Care and Recording of Patient Characteristics Critical for Comparative Effectiveness Research. *Epidemiology (Cambridge, Mass.)*, *29*(3), 356–363. https://doi.org/10.1097/EDE.0000000000000794

Maskarinec, G., Meng, L., & Ursin, G. (2001). Ethnic differences in mammographic densities. *International Journal of Epidemiology*, *30*(5), 959–965. https://doi.org/10.1093/ije/30.5.959

McCulloch, C. E., Neuhaus, J. M., & Olin, R. L. (2016). Biased and unbiased estimation in longitudinal studies with informative visit processes. *Biometrics*, *72*(4), 1315–1324. https://doi.org/10.1111/biom.12501

McGee, G., Haneuse, S., Coull, B. A., Weisskopf, M. G., & Rotem, R. S. (2022). On the Nature of Informative Presence Bias in Analyses of Electronic Health Records. *Epidemiology*, *33*(1), 105–113. https://doi.org/10.1097/EDE.0000000000001432

Merola, D., Schneeweiss, S., Jin, Y., Lii, J., & Lin, K. J. (2022). Advancing an Algorithm for the Identification of Patients with High Data-Continuity in Electronic Health Records. *Clinical Epidemiology*, *14*, 1339–1349. https://doi.org/10.2147/CLEP.S370031

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Mu, N., & Gilmer, J. (2019). *MNIST-C: A Robustness Benchmark for Computer Vision* (arXiv:1906.02337). arXiv. https://doi.org/10.48550/arXiv.1906.02337

Naemi, A., Schmidt, T., Mansourvar, M., Naghavi-Behzad, M., Ebrahimi, A., & Wiil, U. K. (2021). Machine learning techniques for mortality prediction in emergency departments: A systematic review. *BMJ Open*, *11*(11), e052663. https://doi.org/10.1136/bmjopen-2021-052663

Nannan Panday, R. S., Minderhoud, T. C., Alam, N., & Nanayakkara, P. W. B. (2017). Prognostic value of early warning scores in the emergency department (ED) and acute medical unit (AMU): A narrative review. *European Journal of Internal Medicine*, *45*, 20–31. https://doi.org/10.1016/j.ejim.2017.09.027

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

O'Brien, C., Goldstein, B. A., Shen, Y., Phelan, M., Lambert, C., Bedoya, A. D., & Steorts, R. C. (2020). Development, Implementation, and Evaluation of an In-Hospital Optimized Early Warning Score for Patient Deterioration. *MDM Policy & Practice*, *5*(1), 2381468319899663. https://doi.org/10.1177/2381468319899663

On Racism: A New Standard For Publishing On Racial Health Inequities. (2020). *Health Affairs Blog*. https://doi.org/DOI: 10.1377/hblog20200630.939347

Paulus, J. K., Wessler, B. S., Lundquist, C. M., & Kent, D. M. (2018). Effects of Race Are Rarely Included in Clinical Prediction Models for Cardiovascular Disease. *Journal of General Internal Medicine*, *33*(9), 1429–1430. https://doi.org/10.1007/s11606-018-4475-x

Phelan, M., Bhavsar, N. A., & Goldstein, B. A. (2017). Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference. *EGEMS (Washington, DC)*, *5*(1), 22. https://doi.org/10.5334/egems.243

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Sauer, C. M., Chen, L.-C., Hyland, S. L., Girbes, A., Elbers, P., & Celi, L. A. (2022). Leveraging electronic health records for data science: Common pitfalls and how to avoid them. *The Lancet Digital Health*, *4*(12), e893–e898. https://doi.org/10.1016/S2589-7500(22)00154-6

Shah, A. S. V., Griffiths, M., Lee, K. K., McAllister, D. A., Hunter, A. L., Ferry, A. V., Cruikshank, A., Reid, A., Stoddart, M., Strachan, F., Walker, S., Collinson, P. O., Apple, F. S., Gray, A. J., Fox, K. A. A., Newby, D. E., & Mills, N. L. (2015). High sensitivity cardiac troponin and the under-diagnosis of myocardial infarction in women: Prospective cohort study. *BMJ (Clinical Research Ed.)*, *350*, g7873. https://doi.org/10.1136/bmj.g7873

Shalev-Shwartz, S. (2011). Online Learning and Online Convex Optimization. *Foundations and Trends® in Machine Learning*, *4*(2), 107–194. https://doi.org/10.1561/2200000018

Shamout, F., Zhu, T., & Clifton, D. A. (2021). Machine Learning for Clinical Outcome Prediction. *IEEE Reviews in Biomedical Engineering*, *14*, 116–126. IEEE Reviews in Biomedical Engineering. https://doi.org/10.1109/RBME.2020.3007816

Sharples, L. D. (2018). The role of statistics in the era of big data: Electronic health records for healthcare research. *Statistics & Probability Letters*, *136*, 105–110. https://doi.org/10.1016/j.spl.2018.02.044

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

Siddique, J., Daniels, M. J., Carroll, R. J., Raghunathan, T. E., Stuart, E. A., & Freedman, L. S. (2019). Measurement error correction and sensitivity analysis in longitudinal dietary intervention studies using an external validation study. *Biometrics*, *75*(3), 927–937. https://doi.org/10.1111/biom.13044

Singh, A. (2021). CLDA: Contrastive Learning for Semi-Supervised Domain Adaptation. *Advances in Neural Information Processing Systems*, *34*, 5089–5101. https://proceedings.neurips.cc/paper/2021/hash/288cd2567953f06e460a33951f55daaf-Abstract.html

Sisk, R., Lin, L., Sperrin, M., Barrett, J. K., Tom, B., Diaz-Ordaz, K., Peek, N., & Martin, G. P. (2021). Informative presence and observation in routine health data: A review of methodology for clinical risk prediction. *Journal of the American Medical Informatics Association: JAMIA*, *28*(1), 155–166. https://doi.org/10.1093/jamia/ocaa242

Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial Bias in Pulse Oximetry Measurement. *The New England Journal of Medicine*, *383*(25), 2477–2478. https://doi.org/10.1056/NEJMc2029240

Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., & Featherstone, P. I. (2013). The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, *84*(4), 465–470. https://doi.org/10.1016/j.resuscitation.2012.12.016

Snowden, L. R., Catalano, R., & Shumway, M. (2009). Disproportionate use of psychiatric emergency services by african americans. *Psychiatric Services (Washington, D.C.)*, *60*(12), 1664–1671. https://doi.org/10.1176/ps.2009.60.12.1664

Sonntag, J., Behrens, G., & Schmidt-Thieme, L. (2022). Positive-Unlabeled Domain Adaptation. *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, 1–10. https://doi.org/10.1109/DSAA54385.2022.10032409

Steyerberg, E. W. (2009). *Clinical Prediction Models*. Springer New York. https://doi.org/10.1007/978-0-387-77244-8

Stolte, A., Merli, M. G., Hurst, J. H., Liu, Y., Wood, C. T., & Goldstein, B. A. (2022). Using Electronic Health Records to understand the population of local children captured in a large health system in Durham County, NC, USA, and implications for population health research. *Social Science & Medicine (1982)*, *296*, 114759. https://doi.org/10.1016/j.socscimed.2022.114759

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials: Use of Propensity Scores to Assess Generalizability. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(2), 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Subbaswamy, A., & Saria, S. (2020). From development to deployment: Dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, *21*(2), 345–352. https://doi.org/10.1093/biostatistics/kxz041

Suresh, H., & Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *arXiv:1901.10002 [Cs, Stat]*. http://arxiv.org/abs/1901.10002

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digital Medicine*, *3*, 17. https://doi.org/10.1038/s41746-020-0221-y

Thapar, A., Cooper, M., & Rutter, M. (2017). Neurodevelopmental disorders. *The Lancet Psychiatry*, *4*(4), 339–346. https://doi.org/10.1016/S2215-0366(16)30376-5

Thota, M., & Leontidis, G. (2021). Contrastive Domain Adaptation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2209–2218. https://doi.org/10.1109/CVPRW53098.2021.00250

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tzeng, E., Hoffman, J., Darrell, T., & Saenko, K. (2015). Simultaneous Deep Transfer Across Domains and Tasks. *2015 IEEE International Conference on Computer Vision (ICCV)*, 4068–4076. https://doi.org/10.1109/ICCV.2015.463

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial Discriminative Domain Adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2962–2971. https://doi.org/10.1109/CVPR.2017.316

VanderWeele, T. J., & Hernan, M. A. (2012). Results on Differential and Dependent Measurement Error of the Exposure and the Outcome Using Signed Directed Acyclic Graphs. *American Journal of Epidemiology*, *175*(12), 1303–1310. https://doi.org/10.1093/aje/kwr458

VanderWeele, T. J., & Li, Y. (2019). Simple Sensitivity Analysis for Differential Measurement Error. *American Journal of Epidemiology*, *188*(10), 1823–1829. https://doi.org/10.1093/aje/kwz133

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 205395171774353. https://doi.org/10.1177/2053951717743530

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms. *The New England Journal of Medicine*, *383*(9), 874–882. https://doi.org/10.1056/NEJMms2004740

Wanders, J. O. P., Holland, K., Veldhuis, W. B., Mann, R. M., Pijnappel, R. M., Peeters, P. H. M., van Gils, C. H., & Karssemeijer, N. (2017). Volumetric breast density affects performance of digital screening mammography. *Breast Cancer Research and Treatment*, *162*(1), 95–103. https://doi.org/10.1007/s10549-016-4090-7

Weiskopf, N. G., Bakken, S., Hripcsak, G., & Weng, C. (n.d.). A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *eGEMs*, *5*(1), 14. https://doi.org/10.5334/egems.218

Weiskopf, N. G., Hripcsak, G., Swaminathan, S., & Weng, C. (2013). Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*, *46*(5), 10.1016/j.jbi.2013.06.010. https://doi.org/10.1016/j.jbi.2013.06.010

Wells, B. J., Chagin, K. M., Nowacki, A. S., & Kattan, M. W. (2013). Strategies for handling missing data in electronic health record derived data. *EGEMS (Washington, DC)*, *1*(3), 1035. https://doi.org/10.13063/2327-9214.1035

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., & Cole, S. R. (2017). Transportability of Trial Results Using Inverse Odds of Sampling Weights. *American Journal of Epidemiology*, *186*(8), 1010–1014. https://doi.org/10.1093/aje/kwx164

Yan, M., Hong, H., Wilson, J., & Goldstein, B. A. (2025). Estimating the Observability of an Outcome from an Electronic Health Records Dataset Using External Data. *American Journal of Epidemiology*, kwaf013. https://doi.org/10.1093/aje/kwaf013

Yan, M., Pencina, M. J., Boulware, L. E., & Goldstein, B. A. (2022). Observability and its impact on differential bias for clinical prediction models. *Journal of the American Medical Informatics Association*, *29*(5), 937–943. https://doi.org/10.1093/jamia/ocac019

Yu, Y., Chen, P.-Y., Zhou, Y., & Mei, J. (2021). *Adversarial Sample Enhanced Domain Adaptation: A Case Study on Predictive Modeling with Electronic Health Records* (arXiv:2101.04853). arXiv. https://doi.org/10.48550/arXiv.2101.04853

Zhang, T., Chen, M., & Bui, A. A. T. (2022). AdaDiag: Adversarial Domain Adaptation of Diagnostic Prediction with Clinical Event Sequences. *Journal of Biomedical Informatics*, *134*, 104168. https://doi.org/10.1016/j.jbi.2022.104168

Zhang, Y., Yang, X., Ivy, J., & Chi, M. (2019). Time-aware Adversarial Networks for Adapting Disease Progression Modeling. *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, 1–11. https://doi.org/10.1109/ICHI.2019.8904698

Zhao, H., Combes, R. T. D., Zhang, K., & Gordon, G. (2019). On Learning Invariant Representations for Domain Adaptation. *Proceedings of the 36th International Conference on Machine Learning*, 7523–7532. https://proceedings.mlr.press/v97/zhao19a.html

Zhou, Y., Matsouaka, R. A., & Thomas, L. (2020). Propensity score weighting under limited overlap and model misspecification. *Statistical Methods in Medical Research*, *29*(12), 3721–3756. https://doi.org/10.1177/0962280220940334

.