UNIVERSITY OF CALIFORNIA

Los Angeles

Large Scale Observational Analytics for Clinical Evidence Generation

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biomathematics

by

Yuxi Tian

2020

ABSTRACT OF THE DISSERTATION

Large Scale Observational Analytics
for Clinical Evidence Generation

by

Yuxi Tian
Doctor of Philosophy in Biomathematics
University of California, Los Angeles, 2020
Professor Marc A. Suchard, Chair

Longitudinal observational health data are rapidly becoming standardized and consolidated at massive scale. However, the large size and observational nature of this data create infrastructural, statistical, and computational challenges to their utilization for generating reliable clinical evidence. I first review principles of observational research and various methodological and statistical tools used to conduct modern observational studies. I then discuss methodological advancements and their clinical implementations in large scale observational health analytics. I introduce a framework for evaluating propensity score methods that are a central tool in addressing confounding in non-randomized studies. This framework incorporates simulations that model real-world survival data and negative control experiments.I adapt my evaluation framework to probe the real-world prevalence and consequences of "instrumental variables" that unduly dominate propensity score models and bias clinical effect size estimates. I then compare propensity score adjustment methods in research evaluating spline functions for multiple treatment settings. Next, I turn to statistical computing challenges that hinder the application of high-quality methods in large data. I utilize graphics processing unit (GPU) programming to accelerate logistic regression, a staple statistical regression used for propensity score estimation, in the high-dimensional regimes necessitated by the largest health databases. Finally, I conduct clinical studies using tools developed through the Observational Health Data Sciences and Informatics (OHDSI) community that allow large-scale and high-quality observational studies to be conducted with previously

unattainable efficiency. In one study, I analyze the comparative effectiveness of two popular osteoporosis drugs in preventing fractures and in regards to concerning drug-related adverse events. In a second study, I address the highly controversial use of recombinant human bone morphogenetic protein 2 in spinal fusion surgeries. In a third study, I report on the comparative effectiveness of antidepressant treatments in preventing suicide and suicidal ideation within a novel all-by-all paradigm of conducting many hypothesis simultaneously within a medical domain. In a final study, I evaluate the effectiveness of generic vs branded medications of many drugs across three medical domains with regards to death and major cardiovascular events. I conclude with thoughts about future research and progress in observational medical science.

The dissertation of Yuxi Tian is approved.

Damla Senturk

Janet Sinsheimer

Carlos Portera-Cailliau

Marc A. Suchard, Committee Chair

University of California, Los Angeles

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to deeply thank Marc Suchard for being my first instructor of medical school, introducing me to observational health data when I thought no other research field was appropriate, and having remarkable patience with me during some difficult years of graduate school. I would also like to thank Martijn Schuemie for his integral role on so many of my projects, particularly my first one in Chapter 4, and being ever available for support. I also acknowledge Patrick Ryan for welcoming me to the OHDSI community inviting me into his research group; Chapter 9 was borne out of a weeklong "publishathon" workshop in New Jersey.

I thank Trevor Shaddox, who completed his MD/PhD training in the Suchard group, for introducing me to Marc's research and walking me through my first summer rotation project, which eventually lead to the work in Chapter 7. Elande Baro and Rongmei Zhang graciously welcomed me to the FDA during my summer internship and mentored me for the work in Chapter 6. I want to thank Georg Heinze for his hospitality during my time in Vienna, and for inviting me to collaborate on the work in Chapter 11.

The UCLA-Caltech Medical Scientist Training Program saw something in me when almost no other medical school did, and accepted me off the waitlist as I was preparing for an alternative career. I also want to thank the Paul and Daisy Soros Fellowships for New Americans for welcoming me into their community. I am honored to join such a talented and motivated group of individuals who are all striving to make our society and the world a better place. Finally, I want to thank my parents for their lifelong sacrifice to provide me with opportunities they didn't have, including teaching me physics at home and giving up their weekends to take me to math practices. They have stood by me through tough times when no one else was there.

Chapter 4 is a version of [Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. International journal of epidemiology. 2018 Dec 1;47(6):2005-14]. It can be found at `https://doi.org/10.1093/ije/dyy120`.

Chapter 5 is in preparation for publication, as [Tian Y, Pratt Nicole, Hester LL, Schuemie MJ, Suchard MA. Evaluating Instrumental Variables in Propensity Score Models Using Synthetic and Negative Control Experiments]. Its supplementary material can be found at `https://github.com/yuxitian/Dissertation`.

Chapter 6 is a version of [Tian Y, Baro E, Zhang R. Performance evaluation of regression splines for propensity score adjustment in post-market safety analysis with multiple treatments. Journal of biopharmaceutical statistics. 2019 Sep 3;29(5):810-21]. It can be found at `https://doi.org/10.1080/10543406.2019.1657138`.

Chapter 7 is in preparation for publication, as [Tian Y, Shaddox TR, Suchard MA. GPU Parallelization of Cyclic Coordinate Descent for Large Scale Cross Validated Logistic Regression].

Chapter 8 has been submitted to Scientific Reports as [Kim Y, Tian Y, Yang J, Huser V, Jin P, Lambert CG, Park H, Park RW, Rijnbeek PR, Van Zandt M, Vashisht R, Wu Y, You SC, Duke J, Hripcsak G, Madigan D, Reich C, Shah NH, Ryan PB, Schuemie MJ, Suchard MA. Comparative safety and effectiveness of alendronate versus raloxifene in women with osteoporosis: an observational cohort study across nine databases]. Its supplementary material can be found at `https://github.com/yuxitian/Dissertation`.

Chapter 9 is in preparation for publication, as [Tian Y, Park DY, Harden J, Schuemie MJ, Suchard MA. Safety and Effectiveness of Recombinant Human BMP-2 in Spinal Fusion Surgeries: A Large-scale Analysis Across Multiple Observational Data Sources]. Its supplementary material can be found at `https://github.com/yuxitian/Dissertation`.

Chapter 10 is in preparation for publication, as [Tian Y, Miner AS, Schuemie MJ, Suchard MA. Comprehensive Comparative Effectiveness of Antidepressant Treatments in Preventing Suicide and Suicidal Ideation Across National Longitudinal Databases]. Its supplementary material can be found at `https://github.com/yuxitian/Dissertation`.

Chapter 11 is a version of [Tian Y, Reichardt B, Dunkler D, Hronsky M, Winkelmayer WC, Bucsics A, Strohmaier S, Heinze G. Comparative effectiveness of branded vs generic versions of antihypertensive, lipid-lowering and hypoglycemic substances: a population-wide

cohort study. Scientific Reports. 2020 Apr 6;10(1):1-2]. It can be found at `https://doi.org/10.1038/s41598-020-62318-y`.

While working on this dissertation, I was supported by the UCLA-Caltech Medical Scientist Training Program; a NIH NLM F31 Predoctoral Individual National Research Service Award 1F31LM012636, the Paul and Daisy Soros Fellowships for New Americans, and a UCLA Dissertation Year Fellowship.

VITA

2009–2013	B.A. (Physics, Molecular and Cell Biology), UC Berkeley.
2016	Paul and Daisy Soros Fellowships for New Americans.
2017	NLM (National Library of Medicine) F31 Predoctoral Individual National Research Service Award.
2017	M.S. (Biomathematics), UCLA.

PUBLICATIONS AND PRESENTATIONS

Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. International journal of epidemiology. 2018 Dec 1;47(6):2005-14.

Tian Y, Baro E, Zhang R. Performance evaluation of regression splines for propensity score adjustment in post-market safety analysis with multiple treatments. Journal of biopharmaceutical statistics. 2019 Sep 3;29(5):810-21.

Schuemie MJ, Cepeda MS, Suchard MA, Yang J, Tian Y, Schuler A, Ryan PB, Madigan D, Hripcsak G. How Confident Are We About Observational Findings in Health Care: A Benchmark Study. Harvard Data Science Review. 2020 Jan 31;2(1).

Tian Y, Reichardt B, Dunkler D, Hronsky M, Winkelmayer WC, Bucsics A, Strohmaier S, Heinze G. Comparative effectiveness of branded vs. generic versions of antihyperten-

sive, lipid-lowering and hypoglycemic substances: a population-wide cohort study. Scientific Reports. 2020 Apr 6;10(1):1-2.

Tian Y, Schuemie MJ, and Suchard MA. Drug Safety and Comparative Effectiveness at Massive Scale. Presented at the Joint Statistical Meetings, Baltimore, Maryland. July 30, 2017

Tian Y, Schuemie MJ, and Suchard MA. Finding Optimal Propensity Score Estimators Using a Modified Plasmode Simulation Framework. Presented at the 33rd International Conference on Pharmacoepidemiology & Therapeutic Risk Management, Montreal, Canada. August 29, 2017

Tian Y. Comparative safety and effectiveness of osteoporosis drugs: a multi-center observational study conducted through the OHDSI network. Invited talk presented to the Vienna Biometrics Section at the Medical University of Vienna, Vienna, Austria. March 15, 2018

Tian Y, Shaddox TR, and Suchard MA. Conquering Massive Clinical Models with GPU Parallelized Logistic Regression. Presented at the Joint Statistical Meetings, Vancouver, Canada. July 30, 2018

Tian Y, Kim Y, Yang J, Huser V, Jin P, Lambert CG, Park H, Park RW, Rijnbeek PR, Van Zandt M, Vashisht R, Wu Y, You SC, Duke J, Hripcsak G, Madigan D, Reich C, Shah N, Ryan PB, Schuemie MJ, and Suchard MA. Comparative Safety and Effectiveness of Alendronate versus Raloxifene in Women with Osteoporosis. Presented at the 34th International Conference on Pharmacoepidemiology & Therapeutic Risk Management, Prague, Czech Republic. August 25, 2018

CHAPTER 1

Introduction

Although randomized controlled trials (RCTs) are considered the gold standard of evidence in medicine, they are not suitable to answer every clinical question of interest. RCTs are prohibitively expensive and time consuming to conduct, underpowered for the detection of rare clinical outcomes, and frequently have inclusion and exclusion criteria so strict as to exclude a substantial proportion of real-world patients. In addition, they are ill equipped to provide answers to patient specific clinical questions in an emerging area of "precision medicine." Observational health data offer an alternative to RCTs as a complementary resource for reliable clinical evidence generation.

The last decade has seen large advances in health data digitization and the development of observational analytics for drug safety surveillance. Catastrophic drug recalls such as the 2004 worldwide withdrawal of the COX-2 inhibitor Vioxx due to increased risks of heart attack and stroke contributed to the passage of the Food and Drug Administration (FDA) Amendments Act of 2007, which mandated the development of an "active postmarket [drug] risk identification system" and validated methods to analyze safety data from at least 100 million patients by 2012. Out of this mandate came the FDA's Sentinel Initiative for drug surveillance and the public-private Observational Medical Outcomes Partnership (OMOP) to inform appropriate methods for observational database use. After its five-year lifespan, OMOP investigators continued their work through Observational Health Data Sciences and Informatics (OHDSI), a multi-stakeholder, interdisciplinary collaborative in which I am a collaborator. OHDSI has over 140 collaborators across the world and nearly 700 million patient records have been converted to the OMOP Common Data Model [2], a standardized representation of health data [3]. Meanwhile, health data digitization has exploded since

the American Recovery and Reinvestment Act of 2009 required all healthcare providers to adopt and demonstrate "meaningful use" of electronic medical records by 2014 under threat of Medicaid and Medicare reimbursements curtailment. As a result, the already growing percentage of office-based physicians using any electronic health records (EHR) system increased from 48% in 2009 to 87% in 2015 [4], and the percentage of hospitals using a comprehensive "basic" EHR increased from 12% to 84% during the same period [5].

The proliferation of digital health data has already transformed medical practice, but we have only begun to realize health data's potential for improving patient outcomes. EHRs have facilitated medical coordination, simplified records access, and in many cases allow providers and administrators to study broad statistics for their patient population — at the cost of substantial time devoted to entering the data into the system [6]. However, what is lacking is quantitative analysis of large-scale observational data to improve individual patient treatment decisions. Non-randomized data are inherently difficult to utilize because they contain systemic biases and missing data that require further methodological research to tame. Furthermore, the sheer scale of observational health data poses computational challenges for clinical researchers, as even simple statistical models may become computationally expensive or infeasible for widespread use. Despite health data digitization, EHR data remain much less available for clinical research than insurance claims data that are less clinically accurate. Because of a lack of research infrastructure and efficient and rigorous observational methods, most modern observational studies are still poorly or slowly executed, and can take weeks to months to complete despite using pre-existing retrospective data [7]. In light of current reality, many tantalizing promises of digital health — rigorous and automated drug safety surveillance, personalized medicine, high-quality evidence at a clinician's or researcher's fingertips — remain years away.

In this dissertation, I present broad advances in observational health data analytics that range from conducting high quality clinical studies that inform medical practice to developing efficient computational solutions for widely used statistical models. I develop statistical and epidemiological methods to control for confounding inherent to non-random health data, with the goal of improving observational study design. I provide statistical computing advances to

accelerate high-dimensional statistical regressions most commonly utilized in observational studies, to allow clinical researchers to conduct quality studies without the need for extravagant computing resources. Using OHDSI tools and methods, I conduct collaborative clinical studies to provide answers to pressing real-world clinical questions.

In Chapter 2, I present an overview of the statistical concepts that underlie our research community's approach to observational studies, including the predominant use of propensity scores (PS) for measured confounding control and negative control outcomes as an emerging tool for quantifying unmeasured confounding. In Chapter 3, I review statistical models commonly used in observational studies and discuss their numerical computation and approaches to model selection. I briefly describe the software environment that I work in, and relevant computing concepts.

The remaining chapters 4-11 can be read as independent articles. The first three involve advances in epidemiological methodology for conducting large-scale observational studies using PS methods. Considering the ubiquity of PS adjustment in observational studies, it is surprising that the choice of PS model is often one based on investigator preference instead of rigorous comparative performance testing. In Chapter 4, I describe a PS evaluation framework that incorporates simulation experiments and negative control experiments. I apply my methods to compare the performance of the popular "high-dimensional propensity score" algorithm to the statistical workhorse L_1-regularization as PS model selection methods.

In Chapter 5, I adapt my PS evaluation framework to study the real-world impact of "instrumental variables" that strongly affect treatment assignment but have no direct causative effect on the study outcomes. These variables have been shown to produce biased treatment effect estimates in small simulations, but their prevalence and consequence in empirical data is unknown. I conduct experiments evaluating instrumental variables in real-world settings and further explore optimum PS model approaches.

In Chapter 6, I move beyond PS model estimation and consider how to best adjust for a PS in outcome models. I compare the familiar PS adjustment method of inverse probability of treatment weighting to spline methods in multiple treatment scenarios. Multiple treatments

require a multidimensional PS, and optimum adjustment methods involve more complex, multivariate strategies.

Remarkable new research infrastructure has dramatically streamlined the study design process for large-scale observational studies. However, study execution time is a frequent research bottleneck, as statistical computation of large models remains frustratingly slow despite advances in statistical software. In Chapter 7, I develop a numerical optimization to generalized linear models widely use in observational studies, including logistic regression, that couples with graphics processing unit (GPU) programming to offer significant improvements in statistical computing time. My implementation allows clinical researchers to conduct large observational studies without complicated and expensive computing resources, and allows those with considerable resources to tackle more ambitious projects such as fitting many thousands of models as a part of drug safety surveillance.

I conduct several collaborative projects to provide high quality evidence for contemporary clinical questions using large-scale longitudinal databases. Chapter 8 describes a large network study across 9 databases to study the comparative effectiveness of two popular osteoporosis medications: alendronate and raloxifene. We focus primarily on fracture prevention outcomes, and also investigate select serious adverse events associated with alendronate.

Chapter 9 describes a clinical study on the highly controversial use of recombinant human bone morphogenetic protein 2 (BMP) in spinal fusion surgeries as a bone graft alternative. BMP use plummeted when misrepresentations of industry-sponsored research came to light, but few studies have evaluated BMP's safety using large-scale PS methods, and none have done so in multiple databases under a single study design.

In Chapter 10, I report on the comparative effectiveness of antidepressant treatments in preventing suicide and suicidal ideation across multiple databases. We embrace a new paradigm of conducting studies within a clinical domain, that of conducting all pairwise comparisons among a large number of available treatments, thus generating hundreds to thousands of hypothesis that are all investigated under a consistent study methodology.

Chapter 11 addresses the comparative effectiveness topic of generic vs branded med-

ications across three clinical domains of hypertension, hyperlipidemia, and diabetes in a comprehensive national database representing nearly all persons in a country of nearly 10 million. Both main results and extensive subgroup analyses favor one medication type over the other, an interesting finding considering the perceived equivalence of generic and branded drugs.

I conclude in Chapter 12 with a discussion of the current state of observational health research, how my contributions fit into an integrated system of generating clinical evidence, and proposals for future research.

CHAPTER 2

Review: The Modern Observational Study

2.1 Introduction

In this chapter I review concepts in causal inference and the statistical analysis of observational data. Medicine comprises informed decisions aimed at improving patient health. Clinical research, whether couched in causal statistical language or not, is used by patients and their physicians to make medical treatment decisions. In the process of determining a treatment's differential effect on an outcome, we have information on a patient's outcome given whether or not she received the treatment, but not the *potential outcome* if – in a parallel universe – she had made the alternative treatment decision. The Rubin causal model is a formal approach to causal inference that operates in the framework of potential outcomes [8]. In this model, randomized treatment assignment readily provides unbiased effect estimates, but nonrandom observational data require additional statistical adjustments in their analysis. The propensity score (PS) is a predominant method to estimate the otherwise unknown treatment assignment probability in observational data, and its adjustment allows for theoretically unbiased observational studies. However, PS methods can only control for measured confounding, and observational data is fraught with unknown sources of bias. Negative and positive control outcomes are emerging tools to control for unmeasured confounding in observational data by estimating systemic biases using a known clinical standard of truth. The combined use of PS and control outcomes offers the most comprehensive available approach to analyzing large-scale observational health data.

It is important to note that observational data analysis in itself is not a novel field, and well developed statistical methods – including propensity score methods – exist for causal in-

ference in econometrics, clinical trial analysis, and epidemiology. However, electronic health data necessitate much larger models than previous methods and computational tools were designed for. Whereas traditional randomized studies include up to thousands of subjects, observational studies can contain hundreds of thousands to millions of patients due to the massive sizes of many health databases. Furthermore, health databases contain tens to hundreds of thousands of unique covariates – drugs, conditions, procedures, measurements – and collinearity among them is commonplace. The high dimension in both patients and covariates challenges popular methods and statistical computing tools.

2.2 Potential Outcomes and the Rubin Causal Model

This section is largely attributable to and motivated by the text *Causal Inference in Statistics, Social, and Biomedical Sciences* [9]. Suppose I have a headache and am considering taking ibuprofen to treat it. I am interested in whether I have less of a headache if I take the drug versus if I do not take the drug, that is, the *causal effect* of ibuprofen on reducing headache. However, I can only decide to take the ibuprofen or not, and observe my headache outcome for that treatment decision. I am unable to observe the *potential outcome* for the treatment decision I did not take, and thus it is impossible to know the actual causal effect of ibuprofen on my headache.

While my *counterfactual* self that took the other treatment decision is unavailable for analysis, we can learn about the causal effect of interest given information on multiple people who took or did not take the treatment. Another person similar to myself could serve as a comparison, and make the opposite treatment decision from me, and the difference in our outcome can be causally attributed to the treatment. In order to do such an analysis, we first rely on the Stable Unit Treatment Value Assumption (SUTVA):

SUTVA: *The potential outcomes for any unit do not vary with the treatments assigned to other units, and for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*

In the above example, SUTVA means that my comparison person's reaction to treatment

has no bearing on my own, and vice versa. Also, the treatment that all treated people in a study receive are identical, so that we are studying a common causal effect across the population. These assumptions can be violated in practice – if I live with the comparison person, their having a headache can affect my own, and we may receive slightly different ibuprofen tablets – but we rely on them to define a consistent causal effect to estimate.

The fundamental problem with comparing people receiving different treatments to infer causal effects on unobserved potential outcomes is that individuals receive treatments for different reasons. A person with a more severe headache may be more likely to take ibuprofen than someone with a minor headache, and thus they would not serve as adequate comparisons for each other. We need to learn about the treatment assignment mechanism in real-world situations to properly compare people who received differing treatments.

Suppose there are n study subjects indexed by i, and each subject has a treatment indicator $w_i = 1$ if they receive a treatment and $w_i = 0$ if they do not. We observe an outcome $Y_i(w_i)$ for each subject, but we do not observe the potential outcome for the treatment they do not receive: $Y_i(1 - w_i)$. For a single person, we are interested in their unit-level causal effect $Y_i(1) - Y_i(0)$. In the study population, each person has their own treatment assignment w_i and pre-treatment variables (a.k.a. covariates) that may affect their treatment assignment, represented as a p-dimensional vector $\boldsymbol{x_i}$. We obtain some *causal estimand* τ of the true treatment effect size, which can be expressed as some function of all of the potential outcomes, pre-treatment variables, and treatment assignment in the study population:

$$\tau = \tau(\boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{X}, \boldsymbol{W}) \tag{2.1}$$

where $\boldsymbol{Y}(0)$ is the vector of all potential outcomes with $w_i = 0$, $\boldsymbol{Y}(1)$ is the vector of all potential outcomes with $w_i = 1$, $\boldsymbol{X}$ is the $n \times p$ matrix of pre-treatment covariates, and $\boldsymbol{W}$ is the vector of all treatment assignments w_i.

Unfortunately, τ cannot be calculated in this form because potential outcomes are unobserved. In order to perform inference, we turn to the treatment assignment mechanism and introduce further assumptions that are sufficient for valid causal inferences.

A subject's assignment probability the sum of probability of all possible treatment assignment vectors $\boldsymbol{W}$ where $w_i = 1$:

$$p_i(\boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)) = \sum_{\boldsymbol{W}: w_i = 1} \Pr(\boldsymbol{W} | \boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)). \tag{2.2}$$

We are interested in the probability of treatment assignment for subpopulations with a specific value of covariates, $\boldsymbol{x}_i = \boldsymbol{x}$, also known as the *propensity score* $e(\boldsymbol{x})$:

$$e(\boldsymbol{x}) = \frac{1}{N(\boldsymbol{x})} \sum_{i: \boldsymbol{x}_i = \boldsymbol{x}} p_i(\boldsymbol{X}, \boldsymbol{Y}(0), \boldsymbol{Y}(1)) \tag{2.3}$$

where $N(\boldsymbol{x})$ is the number of subjects with $\boldsymbol{x}_i = \boldsymbol{x}$.

Under a set of assumptions collectively referred to as *regular assignment mechanism*, we are able to draw valid causal inferences by adjusting for covariates that differ between treated and control units. The unobserved potential outcomes $\boldsymbol{Y}(0)$ and $\boldsymbol{Y}(1)$ are no longer necessary to obtain causal estimands.

Regular treatment assignment *entails the following assumptions:*

1. *the assignment mechanism is individualistic: the unit level assignment probabilities can be written as a common function of that unit's potential outcomes and covariates; that is, all units with the same potential outcomes and covariates share through a common function the same assignment probability*

2. *the assignment mechanism is probabilistic: the unit level assignment probabilities are strictly between zero and one*

3. *the assignment mechanism is unconfounded: all assignment probabilities are free from dependence on the potential outcomes*

Unconfoundedness implies that the propensity score is solely a function of covariates $\boldsymbol{x}_i$:

$$\Pr(w_i = 1 | Y_i(0), Y_i(1), \boldsymbol{x}_i) = \Pr(w_i = 1 | \boldsymbol{x}_i) = e(\boldsymbol{x}_i). \tag{2.4}$$

The propensity score is a balancing score, that is, a function of the covariates $\boldsymbol{x}_i$ sufficient for introducing independence between treatment assignment and covariates:

$$w_i \perp\!\!\!\perp \boldsymbol{x}_i | e(\boldsymbol{x}_i). \tag{2.5}$$

In summary, with a regular treatment assignment mechanism, it is possible to draw valid causal inferences by conditioning on covariates $\boldsymbol{x}_i$. Because the propensity score is a balancing score, it is also sufficient to condition on just the propensity score $e(\boldsymbol{x}_i)$ and not the entire vector of covariates. However, the unconfoundedness assumption is not testable, so we have in essence traded in unobservable potential outcomes for an untestable treatment assignment assumption. Nonetheless, we have arrived at a method (the propensity score) that allows for causal inference within the potential outcomes framework.

2.3 Negative Control Outcomes and Residual Bias

This section is motivated by two papers by Schuemie et al. [10, 11]. Observational studies can be subject to various sources of bias that lead to unreplicable results. Unaccounted-for bias arises from each study's unique study population, study design, and unmeasured confounding. For example, two studies investigating the same clinical question in the same population can reach differing conclusions [12, 13]. We refer to the totality of bias after controlling for measured confounders as residual bias, or systematic error. Even after adjustment through various statistical methods, systematic error exists and biases the results of traditional significance tests such as p-value calculations.

Observational studies lack the stringent study population selection and control of confounding through randomization that randomized trials provide. They therefore attempt to estimate an unknown quantity (the treatment effect size of interest) despite being subject to an unknown amount of residual bias. Negative control outcomes are an emerging tool in observational research that addresses this problem by providing a standard of clinical truth in observational studies [14, 15]. Negative controls are outcomes that the investigator

believes to be differentially unrelated by the compared treatments. As such, the true effect size of the compared treatments on the negative control should be unity, favoring neither treatment. If the estimated effect size on the negative control differs from unity, then the effect can be attributable to residual bias.

With a set of negative controls (perhaps 50 or 100), we can obtain a reliable estimate of the residual bias from the individual negative control estimates. Perhaps one treatment group, for reasons unrelated to the treatments of interest, is more likely to develop recorded outcomes of any type. Alternatively, perhaps one treatment group is differentially affected by an unmeasured confounder that would lead to more outcomes of many types. Whatever the explanation, we are able to fit negative control estimates to an empirical null distribution [10] that approximates the distribution of residual bias. A Gaussian distribution provides a good approximation:

Suppose we are comparing two drug treatments, so that each negative control outcome corresponds to a drug-outcome pair estimate. Let y_i denote the estimated log effect estimate of the i^{th} negative control drug-outcome pair, and τ_i be the associated standard error. There are n total negative controls, $i = 1, \cdots, n$. Let θ_i be the true error associated with pair i, that would be obtained if the population were infinitely large. We assume that y_i is normally distributed around θ_i with standard deviation τ_i. Additionally, we assume that all the θ_i arise from a normal distribution (the null distribution) with mean μ and variance σ^2:

$$\begin{aligned} \theta_i &\sim N(\mu, \sigma^2) \\ y_i &\sim N(\theta_i, \tau_i^2) \end{aligned} \tag{2.6}$$

where $N(a, b)$ denotes a normal distribution with mean a and variance b. The empirical null distribution parameters μ and σ can be estimated through maximum likelihood:

$$L(\mu, \sigma | \theta, \tau) = \prod_{i=1}^{n} \int p(y_i | \theta_i, \tau_i) p(\theta_i | \mu, \sigma) d\theta_i. \tag{2.7}$$

With the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}$, we can calibrate new p-values that

utilize the empirical null distribution. Suppose a new drug-outcome pair (that of our outcome of interest has log effect estimate y_{n+1} and estimated standard error τ_{n+1}. We assume that the true effect size θ_{n+1} arises from the same empirical null distribution:

$$y_{n+1} \sim N(\hat{\mu}, \hat{\sigma}^2 + \tau_{n+1}^2). \tag{2.8}$$

The calibrated one-sided p-value is now

$$\Phi(\frac{y_{n+1} - \hat{\mu}}{\sqrt{(\hat{\sigma}^2 + \tau_{n+1}^2)}}) \tag{2.9}$$

if $y_{n+1} < \hat{\mu}$ and

$$1 - \Phi(\frac{y_{n+1} - \hat{\mu}}{\sqrt{(\hat{\sigma}^2 + \tau_{n+1}^2)}}) \tag{2.10}$$

if $y_{n+1} > \hat{\mu}$, where $\Phi(\bullet)$ is the cumulative distribution function for the standard normal distribution.

We now have a method for estimating the residual bias distribution and using that estimate to calibrate p-values for estimated effect sizes on our outcomes of interest. However, there is no free lunch. We are still burdened by the unverifiable assumption that there is no differential effect of the compared treatments on the negative control outcomes. Sometimes, we have strong confidence in such assumptions. Consider the negative controls used in Chapter 9, in which the compared treatments are spinal fusion surgeries with and without an artificial bone growth factor to promote bone growth. On of the negative controls (and an OHDSI favorite) is "ingrowing nail," as it seems very implausible that having one kind of surgery or another would differentially affect an ingrown toenail. However, another negative control is "alcohol abuse." While it is difficult to imagine surgery having a direct effect on alcohol abuse, it is possible to imagine dramatic differential surgical outcomes that might lead a patient to abuse alcohol. These secondary, plausible (if not somewhat far-fetched) effects are imaginable for some negative controls, and the assumption of no differential effect is not perfect. However, we still have come a long way from acknowledging residual bias and doing nothing about it, as is the case in many published observational studies.

CHAPTER 3

Review: Statistical and Computational Concepts

3.1 Introduction

In this chapter I review statistical methodology for conducting observational studies. I begin by describing logistic regression, which is commonly used for propensity score estimation. I then detail the Cox proportional hazards model, which is commonly used for estimating outcome effect sizes with longitudinal data. To obtain maximum likelihood estimates for our models, we utilize cyclic coordinate descent as our optimization strategy, which is also detailed in Chapter 7. I also review methods for propensity score estimation and adjustment. PS model selection is a major decision, and I describe methods for automated selection of covariates, including regularized regression. There are multiple methods for adjusting for a PS in outcome models, and I describe strategies including matching, stratification, and splines. Finally, I overview relevant programs in the OHDSI software suite that allow us to conduct large-scale observational studies efficiently, from both a design and computational perspective.

3.2 Logistic Regression

Logistic regression is a predominant approach to modeling a binary dependent variable, including the binary treatment variable in a propensity score. Let i index patients $1, \cdots, n$, $\boldsymbol{x}_i$ be a vector of J pretreatment covariates, and y_i be the treatment indicator. We model the treatment assignment process as a Bernoulli distribution in which the assignment probability p_i is a logit transform of the linear predictor $\boldsymbol{x}_i\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of regression coefficients:

$$p_i = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i\boldsymbol{\beta}) + 1}; \tag{3.1}$$

$\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are expanded to include an intercept term.

The log-likelihood function for maximum likelihood estimation over all patients is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \boldsymbol{x}_i \boldsymbol{\beta} - \log[1 + \exp({}_i\boldsymbol{\beta})]. \tag{3.2}$$

3.3 Cox Proportional Hazards Model

This section is derived partly from [16]. The Cox proportional hazards model is a survival model, in which the dependent variable is a time until outcome, and observations can be censored. Censored observations have not had an outcome at the time that the subject is no longer observed. The Cox model is a common model for estimating hazard ratios in time-to-event data such as those from longitudinal databases.

Let T be the variable for time until outcome, and $f(t)$ be its probability distribution function. We deal more with the survival function $S(t)$ that represents the probability of being alive just before time t:

$$S(t) = \Pr\{T \geq t\} = \int_t^{\infty} f(x)dx. \tag{3.3}$$

The hazard function $\lambda(t)$ is the instantaneous rate of occurrence of the event and is given by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + dt | T \geq t)}{dt} = \frac{f(t)}{S(t)}. \tag{3.4}$$

The above expression can be solved for an equation for the survival function $S(t)$:

$$S(t) = \exp(-\int_0^t \lambda(x)dx). \tag{3.5}$$

The Cox proportional hazards model assumes that all subjects share a baseline hazard function $\lambda(t)$ modulated by their linear predictor of covariates $\boldsymbol{x}_i\boldsymbol{\beta}$. Each subject's hazard function is

$$\lambda_i(t|\boldsymbol{x}_i) = \lambda_0(t)\exp(\boldsymbol{x}_i\boldsymbol{\beta}). \tag{3.6}$$

The subject-specific survival function can then be expressed as

$$S_i(t|\boldsymbol{x}_i) = S_0(t)^{\exp(\boldsymbol{x}_i\boldsymbol{\beta})} \tag{3.7}$$

where $S_0(t) = \exp(-\int_0^t \lambda_0(x)dx)$ is the baseline survival function.

Note that the proportional hazards model is a simple additive model for the log of the hazard, $\log\lambda_i(t|\boldsymbol{x}_i) = \alpha_0(t) + \boldsymbol{x}_i\boldsymbol{\beta}$, and the "hazard ratio" of our desired treatment is the coefficient β for the corresponding covariate.

Let be δ_i be the censoring variable, with $\delta_i = 1$ indicating the outcome event and $\delta_i = 0$ indicating censoring. Also let $R_i = \{j : t_j \geq t_i\}$ be the "risk set" of patients j with time to outcome (or censoring) greater than or equal to t_i. The log-likelihood function for maximum likelihood estimation over all patients is:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i(\boldsymbol{x}_i\boldsymbol{\beta} - \log\sum_{j\in R_i}\exp(\boldsymbol{x}_j\boldsymbol{\beta})). \tag{3.8}$$

3.4 Cyclic Coordinate Descent

Numerical optimization refers to the method of finding the estimates $\hat{\boldsymbol{\beta}}$ that maximize the log-likelihood, and are the most likely coefficients given the observed data. We employ cyclic coordinate descent (CCD) that cycles through all J covariates and takes one-dimensional Newton steps in each covariate dimension. A Newton step size is equal to the first derivative of the log-likelihood divided by the second derivative, and one-dimensional Newton steps involve only taking scalar derivatives of the log-likelihood with respect to each covariate [17].

This avoids the inversion of second derivative Hessian matrices present in the multivariate Newton's method and other optimization strategies. With thousands of covariates available in observational studies conducted in large-scale longitudinal databases, these Hessian matrices can become very large and their inversion computationally expensive.

Suppose we are taking a one-dimensional Newton step for covariate k. The first derivative (gradient) of the log-likelihood is $g_k = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_k}$ and the second derivative (hessian) is $h_k = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_k^2}$. The Newton step update is then

$$\beta_k \leftarrow \beta_k - \frac{g_k}{h_k}. \tag{3.9}$$

For logistic regression, these derivatives are:

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_k} &= \sum_{i=1}^{n} y_i x_{i,k} - \frac{x_{i,k} \exp(\boldsymbol{x}_i \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i \beta)} \\ \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_k^2} &= -\sum_{i=1}^{n} \frac{x_{i,k}^2 \exp(\boldsymbol{x}_i \boldsymbol{\beta})}{(1 + \exp(\boldsymbol{x}_i \boldsymbol{\beta}))^2}. \end{aligned} \tag{3.10}$$

For Cox proportional hazards regression, these derivatives are:

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_k} &= \sum_{i=1}^{n} \delta_i x_{i,k} - \delta_i \frac{\sum_{j \in R_i} x_{j,k} \exp(\boldsymbol{x}_j \boldsymbol{\beta})}{\sum_{j \in R_i} \exp(\boldsymbol{x}_j \boldsymbol{\beta})} \\ \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_k^2} &= -\sum_{i=1}^{n} \delta_i \big[\frac{\sum_j \in R_i x_{j,k}^2 \exp(\boldsymbol{x}_j \boldsymbol{\beta})}{\sum_{j \in R_i} \exp(\boldsymbol{x}_j \boldsymbol{\beta})} - \frac{(\sum_{j \in R_i} \exp(\boldsymbol{x}_j \boldsymbol{\beta}))^2}{\sum_{j \in R_i} \exp(\boldsymbol{x}_j \boldsymbol{\beta}))^2} \big]. \end{aligned} \tag{3.11}$$

3.5 Propensity Score Estimation and Regularized Regression

Because the PS models the binary treatment assignment process, it is usually estimated through a logistic regression. The primary topic of concern is how to select covariates to include in the PS model. There are potentially thousands of variables available in observational longitudinal databases, including conditions, procedures, drug exposures, and more. Traditionally, investigators manually select suspected confounders to include in the PS model. However, the reliably of such expert opinion is questionable, and different experts can (and

typically do) arrive at somewhat different sets of covariates. Recently, automated systems for selecting PS model covariates have been developed.

Chapter 4 evaluates two approaches to automated covariate selection for PS models. One approach employs a univariate screen on the covariates to select most likely confounders for the PS model. Covariates are ranked by a metric for their association with the outcome, and an arbitrary number of the top ranked covariates are selected [18] The other approach, often utilized by researchers in the OHDSI community, is to include all pretreatment covariates in the PS model [19]. To avoid model overfitting and perform model selection, statistical regularization is employed. Statistical regularization penalizes the log-likelihood by a function of the covariate coefficients, with the goal of shrinking some coefficient magnitudes. Two common penalties are the L_1 norm of the coefficients (a.k.a. the "lasso" penalty [20]), and the L_2 norm of the coefficients (A.K.A. the "ridge" penalty). The lasso has the attractive property of shrinking some coefficients to exactly zero, thus excluding the covariate from the model. With the lasso penalty, the target for maximum likelihood estimation becomes a penalized log-likelihood $P(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + p(\boldsymbol{\beta})$:

$$p(\boldsymbol{\beta}) = -\lambda \sum_{j=1}^{J} |\beta_j| \tag{3.12}$$

where λ is a hyperparameter controlling the magnitude of penalization.

The corresponding penalized log-likelihood with ridge regression is

$$p(\boldsymbol{\beta}) = -\lambda \sum_{j=1}^{J} \beta_j^2 \tag{3.13}$$

The optimum value of λ is commonly found empirically through a process called cross-validation. In cross-validation, the data are divided into multiple folds, and the folds are left out one at a time. The logistic regression is fit to the remaining folds, and the optimum solution $\hat{\boldsymbol{\beta}}$ is used to calculate an out-of-sample predictive likelihood on the left-out fold. Different values of λ are searched, and the one with the highest average out-of-sample likelihood is selected as the optimum value.

Chapter 5 touches on a related controversy in PS estimation. The PS is defined as an estimate of the treatment assignment probability, and one would think that to estimate it one should build a model to predict the treatment using pretreatment covariates. Performing PS estimation this way allows one to construct a stratified population that mimics a randomized study, allowing for unbiased outcome effect size estimation. Outcome data, that postdates the treatment, should not affect the PS model. This approach to PS estimation is advocated by Rubin, one of the early introducers of the propensity score in observational studies [21], in multiple subsequent papers [22, 23, 24]. We follow this approach in [19] and many OHDSI research projects [25], including only pretreatment covariates and using regularized regression as our model selection strategy.

In contrast, there is a school of thought that believes the outcome data can and should be incorporated into the PS model, because only variables that affect both the treatment and the outcome are true confounders. Other variables would merely introduce bias and variance into the PS-adjusted effect size estimate. This approach, which guides the univariate screen mentioned above, sacrifices the definition of the PS as a treatment prediction model for an attempt to identify and only include true confounders as PS model covariates.

3.6 Propensity Score Adjustment

As many methods there are for selecting covariates to include in a PS model, there are even more methods to adjust for a PS in the outcome model, which is often a Cox proportional hazard model. These methods are partly described in [26], and I have listed the most common methods below.

- Covariate adjustment – the PS is used directly as a covariate in the outcome model, typically as the only covariate other than the treatment indicator. The PS can be included as a single covariate as a linear predictor. This requires the assumption that the treatment effect is linearly associated with the PS. Alternatively, a transformation of the PS, such as spline functions, can be used instead to allow for nonlinear effects. Spline adjustments of the PS are explored in Chapter 6.

- Matching – the PS is used as the metric for matching treated and comparator subjects to create a group of matched sets; the outcome model will be stratified according to the matched sets. The matching is traditionally one-to-one matching, but variable length matching is possible and frequently performed. A caliper is used as a maximum matching distance to prevent subjects with too distant propensity scores from being matched.

- Stratification – the PS is used to stratify subjects into large buckets based on quantiles. Five or ten strata are often used, and the outcome model is again stratified according to the strata.

- Inverse probability of treatment weighting (IPTW) – each subject is weighted by the inverse of their PS. This creates a pseudo-population with an inflated population, in which both treatment cohorts have identical PS distributions. The outcome model now becomes weighted.

3.7 The OHDSI Software Suite

In addition to converting hundreds of millions of patient records across dozens of databases around the world to the OMOP Common Data Model (CDM) [2], the OHDSI community has developed a suite of observational analytics software to facilitate conducting large-scale observational studies. I describe some of these software tools that I utilize the most for my research:

- **ATLAS** – this web tool (`https://atlas.ohdsi.org/`) is a comprehensive portal to explore CDM data and to specify observational studies at the click of a button. One can view dashboard representations of data, explore individual (anonymized) patient records, search the CDM vocabulary, construct concept sets and cohort definitions, generate cohorts on a database, and fully specify an observational study that is automatically constructed as a downloadable R package. For most of the OHDSI collaborations utilized in the chapters of this dissertation, ATLAS was used to construct the

study specifications required for obtaining the clinical data of interest.

- **CohortMethod** – this R package [27] provides a central interface for running estimation studies in which two treatments are compared for one or more outcomes of interest. CohortMethod interfaces with the data server through the package DatabaseConnector using universally translated SQL from the package SqlRender. Covariates are constructed through the package FeatureExtraction. Once the data are obtained, the R package Cyclops does the heavy lifting of regularized regression to obtain propensity scores and to calculate outcome models. P-value and confidence interval calibration with control outcomes is then provided through the package EmpiricalCalibration. CohortMethod also provides the graphical scripts that generate displayable figures for study results.

- **Cyclops** – this R package [28] performs efficient cyclic coordinate descent for generalized linear models, including Poisson regression, logistic regression, Cox proportional hazards regression, linear regression, and the self-controlled case series [29]. Cyclops also performs cross-validation to search for optimum regularization hyperparameters. Cyclops outperforms many existing R software in conducting sparse and regularized regression on high-dimensional data. I extend the capabilities of Cyclops to perform graphics processing unit (GPU) computation in Chapter 7.

CHAPTER 4

Evaluating Large-Scale Propensity Score Performance Through Real and Synthetic Data Experiments

4.1 Introduction

Retrospective observational studies constitute a resource for clinical evidence gathering complementary to randomized controlled trials. Longitudinal databases contain staggering volumes of information available for conducting retrospective studies: all recorded medical conditions, procedures, medications, and clinical measurements for millions of patients in real-world settings [30]. Unfortunately, observational studies suffer deficiences that introduce bias and prevent their more widespread use by the medical community [31, 32]. Chief among these is the unknown and non-random treatment assignment process that precludes the cohort balance inherent in randomized studies.

The propensity score (PS), an estimate of treatment assignment probability, is a predominant tool for confounding control in retrospective studies where the true treatment assignment process is unknown [33, 21]. Propensity scores are frequently estimated using a logistic regression model with pretreatment baseline patient covariates such as demographics and indicators for medical conditions, procedures, and drug exposures [26]. Traditionally, the investigator manually selects suspected confounders to include as PS model covariates [34]. However, the reliability of expert opinion in properly selecting confounders out of all available covariates is suspect [35]. Many aspects of a patient's medical profile could be contributive to a treatment decision, yet escape an expert's contemplative recollection. In contrast, several automated methods for covariate selection better utilize the multitude of covariates available in longitudinal databases. The high-dimensional propensity score (hdPS)

algorithm [18] has gained widespread use in pharmaco-epidemiology [36]. The algorithm screens covariates by marginal associations and includes a predetermined number in the PS logistic regression model. In contrast, other automated PS model selection methods – such as iterative selection procedures [9], the covariate balancing propensity score [37], and statistical regularization [38, 39, 40] – all consider joint effects of covariates instead of individual effects. In particular, L_1-regularization is a workhorse of statistical model selection, and introduces a penalty term into the PS logistical regression likelihood that pushes coefficient values to zero, dropping the respective covariates from the model [20].

Rich literature addresses PS adjustment methods to estimate comparative treatment effects, including stratification, matching, and direct inclusion into the outcome model [26, 9, 41, 42]. However, relatively few studies evaluate PS estimation method performance [40, 43]. In this paper, we address the comparative performance and optimal selection of PS estimators in large-scale observational settings on the order of 100,000 subjects and 100,000 unique covariates. We detail a framework for evaluating PS methods, and conduct a comparison of the hdPS algorithm with L_1-regularization for PS estimation. Our framework includes two aspects: a survival simulation method that extends the "plasmode" framework concept [44, 45], and negative control outcome experiments that utilize outcomes known to be unrelated to the investigated treatments [14, 10]. Synthetic and negative control experiments each have limitations, but their combined use offers value in evaluating PS performance.

4.2 Synthetic Framework

Our synthetic approach realistically simulates survival outcomes while preserving characteristics of real-world clinical cohorts. From a longitudinal database, we construct new user cohorts comparing the effect of two drugs on an outcome of interest [38]. Using the empirical exposure status and baseline covariates, we model the outcome of interest under a Cox proportional hazards model and then simulate new outcomes under a desired true hazard ratio.

4.2.1 Notation

N total study subjects are indexed by i, and have treatment indicator w_i and p-length baseline covariate vector $\boldsymbol{x}_i$. t_i is the event time, and δ_i is the censoring indicator, with $\delta_i = 0$ indicating censoring and $\delta_i = 1$ indicating the outcome of interest. Under the proportional hazards model, η and $\boldsymbol{\beta}$ are the log hazard ratios for the treatment and the baseline covariates, respectively; the subject-specific hazard is then $\theta_i = w_i\eta + \boldsymbol{x_i\beta}$. The baseline survival function $S(t)$ traces the probability of surviving to time t after treatment initiation and $C(t)$ is the analagous baseline censoring function.

4.2.2 Estimate Simulation Components

Outcome simulation requires estimates for $S(t)$, $C(t)$, and $\boldsymbol{\beta}$. We estimate $S(t)$ by fitting a distribution to the observed outcome of interest, and $C(t)$ by fitting a distribution to the censoring times. Critically, the censoring function must be covariate-free to maintain non-informative censoring for the proportional hazards model, meaning that a subject's censoring time and survival time are independent. This point is overlooked in the "plasmode" framework, leading to inaccurate true hazard ratios that are not proportional hazards. Possible forms for $S(t)$ and $C(t)$ include parametric distributions such as exponential, Weibull, Gompertz, gamma, and lognormal; discrete nonparametric estimators such as the Breslow and Kalbfleisch-Prentice estimators [46] (which without covariates are respectively the Nelson-Aalen and Kaplan-Meier estimators); and nonparametric spline functions [47]. The $S(t)$ distribution determines how the covariate coefficients $\boldsymbol{\beta}$ are estimated. For parametric and spline estimators, the parameters that characterize $S(t)$ are jointly estimated with the covariate coefficients often using maximum likelihood estimation on the full survival likelihood function. For the discrete nonparametric estimators, covariate coefficients are first estimated via the partial likelihood function, and then used to produce $S(t)$ [48, 49]. We additionally smooth the discrete nonparametric estimators to avoid excessive simulated outcome time ties that can affect estimation bias (see Web Appendix 1). The subject-specific hazard is then $\theta_i = w_i\hat{\eta} + \boldsymbol{x_i}\hat{\boldsymbol{\beta}}$ and the subject-specific survival function $S(t)^{\exp\{\theta_i\}}$, where $\hat{\eta}$ and $\hat{\boldsymbol{\beta}}$ are

maximum likelihood estimates.

4.2.3 Simulate Outcome and Censoring Times

Under the proportional hazards framework, each subject's survival process is $S(t)^{\exp\{\theta_i\}}$ and censoring process $C(t)$. We use inverse transform sampling and draw for each subject two Unif$(0,1)$ random variables $R_{i,s}$ and $R_{i,c}$. The respective outcome and censoring times are $t_{i,s} = \min\{t : S_i(t) \leq R_{i,s}\}$ and $t_{i,c} = \min\{t : C(t) \leq R_{i,c}\}$. The final simulated event is the miminum time:

$$\begin{aligned} t_i &= \min\{t_{i,s}, t_{i,c}\} \text{ and} \\ \delta_i &= \begin{cases} 1 & t_{i,s} < t_{i,c} \\ 0 & t_{i,s} \geq t_{i,c} \end{cases}. \end{aligned} \tag{4.1}$$

4.2.4 Adjust Simulation for Hazard Ratio and Outcome Prevalence

To simulate under a desired treatment hazard ratio η^*, we replace the empirically estimated $\hat{\eta}$ by η^*: $\theta_i = w_i\eta^* + x_i\hat{\beta}$. The expected resultant simulated outcome prevalence (OP) is

$$p = \frac{1}{N}\sum_i \int_0^\infty \Pr\left(t_{i,s} = t < t_{i,c}\right) \mathrm{d}t = \frac{1}{N}\sum_i \int_0^\infty \left(\frac{\partial}{\partial t} S(t)^{\exp\{\theta_i\}}\right) C(t)\,\mathrm{d}t. \tag{4.2}$$

Let $t_{(k)}$ be the observed outcome times; the corresponding equation for discrete estimators is

$$p = \frac{1}{N}\sum_i \sum_{t_{(k)}} \left[S\left(t_{(k-1)}\right)^{\exp\{\theta_i\}} - S\left(t_{(k)}\right)^{\exp\{\theta_i\}}\right] C(t_{(k)}). \tag{4.3}$$

Similarly to the "plasmode" framework, we simulate under a desired outcome prevalence p by adjusting the baseline survival function by an exponential factor $\gamma \in (0, \infty)$: $S(t) \to S(t)^\gamma$.

Adjustment factor γ is computed numerically to satisfy the outcome prevalence Equation (4.2). This approach is a constant modification to the baseline outcome hazard. In Web Appendix 2, we propose additional approaches to adjusting for outcome prevalence that may suit the investigator.

4.3 Negative Control Outcome Experiments

As an alternative to simulations under known hazard ratios, we perform negative outcome control experiments using sets of outcomes *a priori* believed to be unrelated to the compared treatments, thus having a presumed true hazard ratio of 1 [14, 10]. Negative control outcomes entirely utilize real-world data from the observational database. For the considered cohort, we identify a set of negative control outcomes, and produce a PS-adjusted estimate of treatment effect size for each outcome. Deviations from unity in the estimated hazard ratios may be due to random error, residual systemic biases (possibly arising from inadequate PS adjustment), or incorrect negative control selection. While the precise relative contribution of each of these individual effects cannot be determined or divined, we assume that successfully controlling for one source of bias reduces the absolute bias in the estimated hazard ratio. That is, for a particular set of empirical cohorts and negative control outcomes, we interpret as superior the propensity score method whose adjustment brings the estimated hazard ratios closer to 1.

4.4 Application

Clinical scenarios

We compare PS methods through reproductions of two previously published retrospective cohort studies using the Truven Health MarketScan Medicare Supplemental and Coordination of Benefits Database. Each study compares two drugs: one designated as the active treatment and the other as the reference. See Web Appendix 9-12 for full cohort definitions.

*A*nticoagulants – The first study [50] is a new-user cohort study of dabigatran and warfarin initiators in patients with non-valvular atrial fibrillation. Dabigatran is the active treatment, warfarin is the reference, and intracranial hemorrhage is the outcome of interest.

*N*onsteroidal Anti-inflammatory Drugs – The second study [51] is a new-user cohort study of COX-2 inhibitors and traditional nonsteroidal anti-inflammatory drugs (NSAIDs) initiators. We select celecoxib, a representative COX-2 inhibitor, as the active treatment; diclofenac, a representative traditional NSAID, as the reference; and upper gastrointestinal complications as the outcome of interest.

4.4.1 Covariates

We extract two sets of pretreatment covariates for our studies, termed "CDM Covariates" and "hdPS Algorithm Covariates." The "CDM Covariates" follow the Observational Medical Outcomes Partnership Common Data Model Version 5 format [2], while the "hdPS Algorithm Covariates" are our reproduction of the specific covariates described in the hdPS algorithm paper [18]. Both sets of covariates include demographic information including sex, age, and treatment initiation index year. The "CDM Covariates" used are more expansive than the "hdPS Algorithm Covariates," with the latter including conditions, procedures, and drug covariates, and the former additionally including measurements, observations, aggregate disease scores, and multiple lookback windows. No threshold is used to exclude infrequent covariates. See Web Appendix 13 for full covariate details.

4.4.2 Simulation Methods

We obtain $\hat{\boldsymbol{\beta}}$ through partial likelihood maximum likelihood estimation, and include L_1-regularization on all covariates except treatment to promote model fitting [52]. We manually select the regularization penalty to yield an approximate model size of 500, coinciding with the number of covariates selected by the hdPS algorithm. We use the Breslow estimator for $S(t)$ [49], and the Nelson-Aalen estimator for $C(t)$. We adjust for outcome prevalence by the modification $S(t) \to S(t)^{\gamma}$, with γ obtained numerically through Equation (4.2).

4.4.3 Propensity Score Methods

We compare the hdPS algorithm to L_1-regularization as PS estimation methods. The synthetic model is constructed using both covariate sets combined, but we apply the hdPS algorithm only to the specifically preprocessed "hdPS Algorithm Covariates," and we apply L_1-regularization to "hdPS Algorithms Covariates" alone, "CDM Covariates" alone, and both covariate sets combined. We include two variations of the the hdPS algorithm: "bias-based hdPS" that screens covariates based on their apparent relative risk, a measure of confounding on the outcome [53], and "exposure-based hdPS" that screens based on treatment relative risk [43]. We use default hdPS algorithm settings, including considering the 200 most prevalent covariates in each "data dimension," selecting the top 500 overall ranked covariates, and fitting an unregularized logistic regression model [18]. However, as the unregularized model can lead to "convergence failures" that occur due to the PS estimate nonexistence [36, 54, 55], we evaluate the hdPS algorithm both with and without L_1-regularization. All regularization penalties are selected through 10-fold cross-validation using large-scale regression tools [28]. Table 4.1 lists the 7 compared PS methods.

PS method	Description
L1-Reg-All	L_1-regularization on combined covariates
L1-Reg-CDM	L_1-regularization on "CDM Covariates" only
L1-Reg-HDPS	L_1-regularization on "hdPS Algorithm Covariates" only
bias-hdPS	bias-based hdPS algorithm, without regularization
bias-hdPS-Reg	bias-based hdPS algorithm, with regularization
exp-hdPS	exposure-based hdPS algorithm, without regularization
exp-hdPS-Reg	exposure-based hdPS algorithm, with regularization

Table 4.1: PS methods evaluated across two real-world studies

Using the CohortMethod package [56], we perform PS matching and then estimate the treatment hazard ratio using a stratified Cox survival outcome model with treatment as the only covariate. We avoid one-to-one matching due to inferior covariate balance [57] and bias reduction [58], and instead use variable length matching [59] with a maximum ratio of 10:1 and a propensity score caliper of 0.05, and use a greedy matching algorithm [60]. We use the less prevalent treatment as the "one" in the many-to-one matching to maximize the number

of subjects that are matched.

4.4.4 Negative Controls

For each study, we identify a set of 50 negative control outcomes using the approach described in [10] and the specific method detailed in [61]. Because extremely rare outcomes lead to effect estimates with substantial variance, we exclude outcomes that have less than 0.02% prevalence in the combined treatment groups. After this exclusion, there are 49 negative control outcomes for the Anticoagulants study and 29 for the NSAIDs study. A list of negative outcomes used are given in Web Appendix 6-7.

4.4.5 Metrics

We evaluate PS methods on outcome-dependent and outcome-independent metrics. Outcome-dependent metrics require simulated or real outcome data. In the simulations, we report the estimation bias and 95% confidence interval coverage obtained from the profile likelihood [62]; in the negative control experiments, we report bias from the presumed null true value. Outcome-independent metrics evaluate PS performance absent of any outcome data, and include the c-statistic of the PS model, a.k.a. the area under the receiver operating characteristic curve (AUC), and standardized difference measures of covariate balance [59, 63].

4.5 Results

4.5.1 Cohorts

The Anticoagulants study contains 72,489 subjects: 19,768 new dabigatran users and 52,721 new warfarin users. There are 98,118 unique baseline covariates among all subjects, and the outcome prevalence of intracranial hemorrhage is 0.26%. The NSAIDs study contains 121,317 subjects: 78695 new celecoxib users and 42,622 new diclofenac users. There are 75,425 unique covariates among all subjects, and the outcome prevalence of upper gastrointestinal complications is 1.81%. Table 4.2 reports summary statistics about covariates

in each study. The "CDM Covariates" set is notably larger than the "hdPS Algorithm Covariates" set in both studies.

Study		Covariates		
		All	CDM	hdPS
Anticoagulants	Full cohorts	98,118	82,281	15,854
	Synthetic model	525	446	83
NSAIDs	Full cohorts	75,425	63,004	12,441
	Synthetic model	530	478	60

Table 4.2: Number of covariates in each study, by source covariate set. Both sets share same demographics covariates.

4.5.2 Propensity Score Estimate Existence

To explore the robustness of the default hdPS algorithm without regularization, we conduct tests for hdPS estimate existence under varied simulation parameters (Web Appendix 3). We find that simulations with smaller cohorts and lower outcome prevalences have less likely PS estimate existence. To address this problem, L_1-regularization readily promotes model existence for the hdPS algorithm.

4.5.3 Propensity Score Distributions

Figure 4.1 plots for the NSAIDs study the distribution of preference scores that normalize propensity scores by their prevalence [64]. The exposure-based hdPS algorithm is sharply peaked due to hundreds of subjects with identical PS values, indicating poor treatment group differentiation. These coincident PS values with exposure-based hdPS are also observed in the Anticoagulants study (Web Appendix 5).

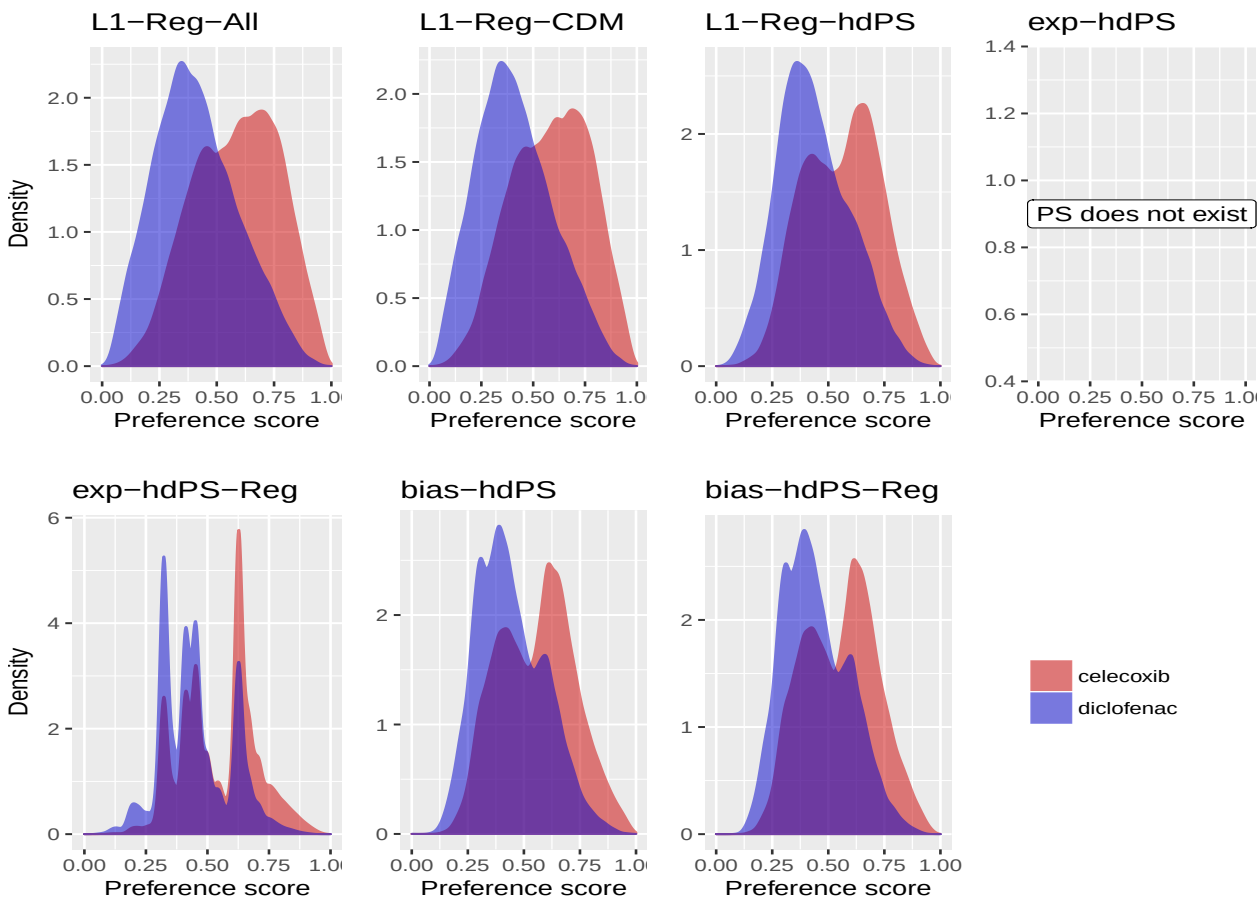


Figure 4.1: NSAIDs study: preference score distributions. Bias-based hdPS algorithm used on the empirical outcome of interest. exp-hdPS fails to construct a PS model.

Figure 4.2 shows the AUC and proportion of matched subjects for compared PS methods. The hdPS algorithm produces similar results with and without regularization. Although the two studies differ in absolute AUC values, they demonstrate a similar ordering of PS methods in order of highest-to-lowest AUC: L1-Reg-All, L1-Reg-CDM, L1-Reg-HDPS, bias-based hdPS, exposure-based hdPS. L1-Reg-All and L1-Reg-CDM have significantly higher AUC than the other methods that use only the "hdPS Algorithm Covariates," suggesting that the larger "CDM Covariates" set allows for improved treatment prediction accuracy. Expectedly, increased AUC and PS distribution differentiation lead to fewer suitable subjects included in the matching process.

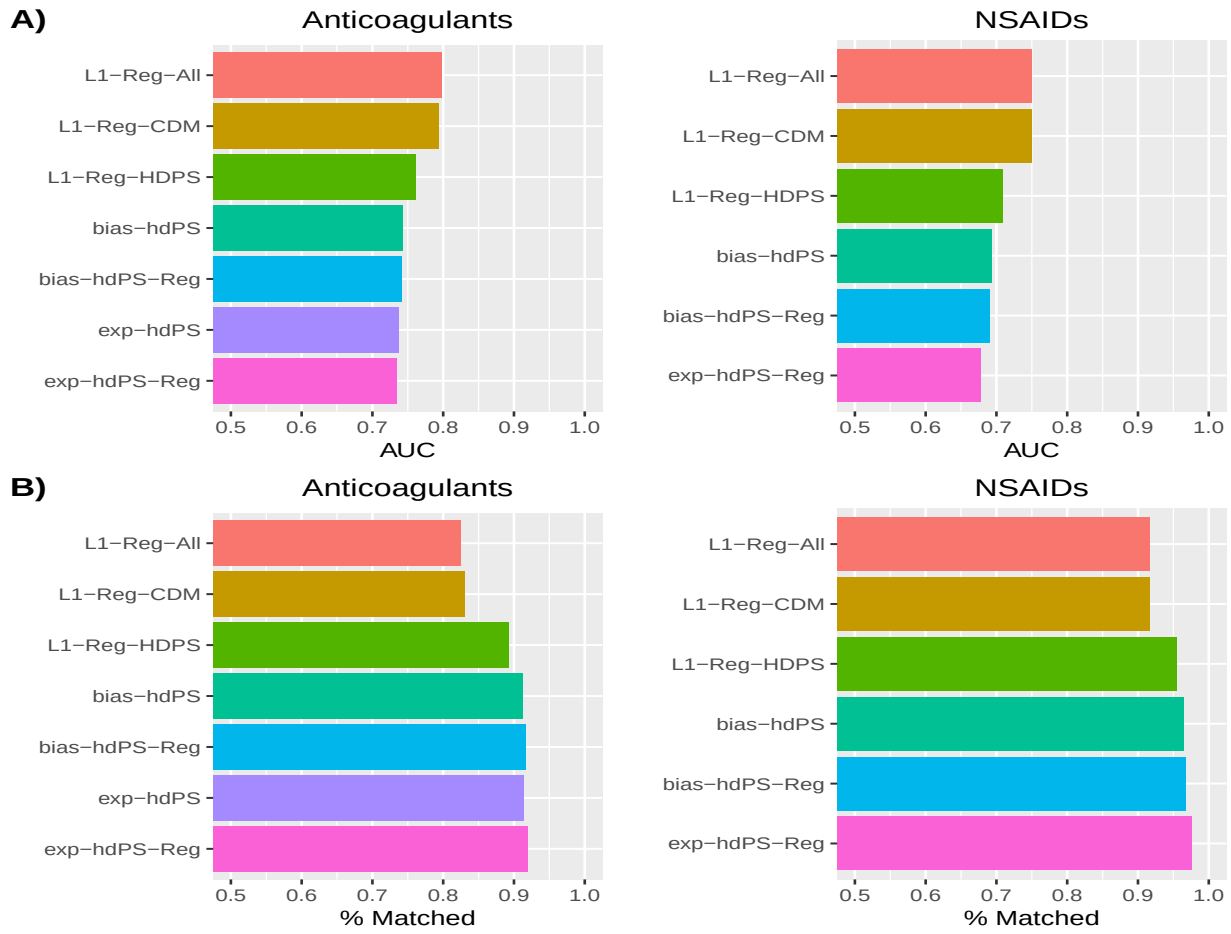


Figure 4.2: A) c-statistic (AUC) of propensity score models. B) percentage of subjects included in matching process. Bias-based hdPS algorithm results from empirical outcome of interest data.

4.5.4 Simulation – Covariate Balance

In the simulation experiments, only the 500 or so synthetic model covariates are true confounders that contribute to estimation bias. Figure 4.3 shows the original and PS adjusted standardized differences for these synthetic model covariates in the Anticoagulants study. A covariate whose standardized difference is improved by PS adjustment will lie below the dotted line. While all PS methods improve covariate balance, L1-Reg-All and L1-Reg-CDM perform best and exposure-based hdPS algorithm worst. Additional analysis reveals L1-Reg-HDPS creates better covariate balance than the bias-based hdPS algorithm (Web Appendix 5). The same relative PS method performance also holds when looking at all covariates instead of just the synthetic model covariates. The NSAIDs study demonstrates similar

results (Web Appendix 5).

The after-matching outlier in Figure 4.3 is the "CDM Covariates" indicator for "Condition Era Overlapping with Cohort Index: Atrial Fibrillation" that is more frequent in the warfarin group. Patients with this covariate have atrial fibrillation records both before and after treatment initiation, and are considered to have chronic atrial fibrillation that may require the stronger anticoagulant control that warfarin is believed to provide. As such, this covariate has high clinical plausibility as a confounder. This derived covariate is absent from the "hdPS Algorithm Covariates" set, and the 5 PS methods that balance only using "hdPS Algorithm Covariates" exacerbate its imbalance.

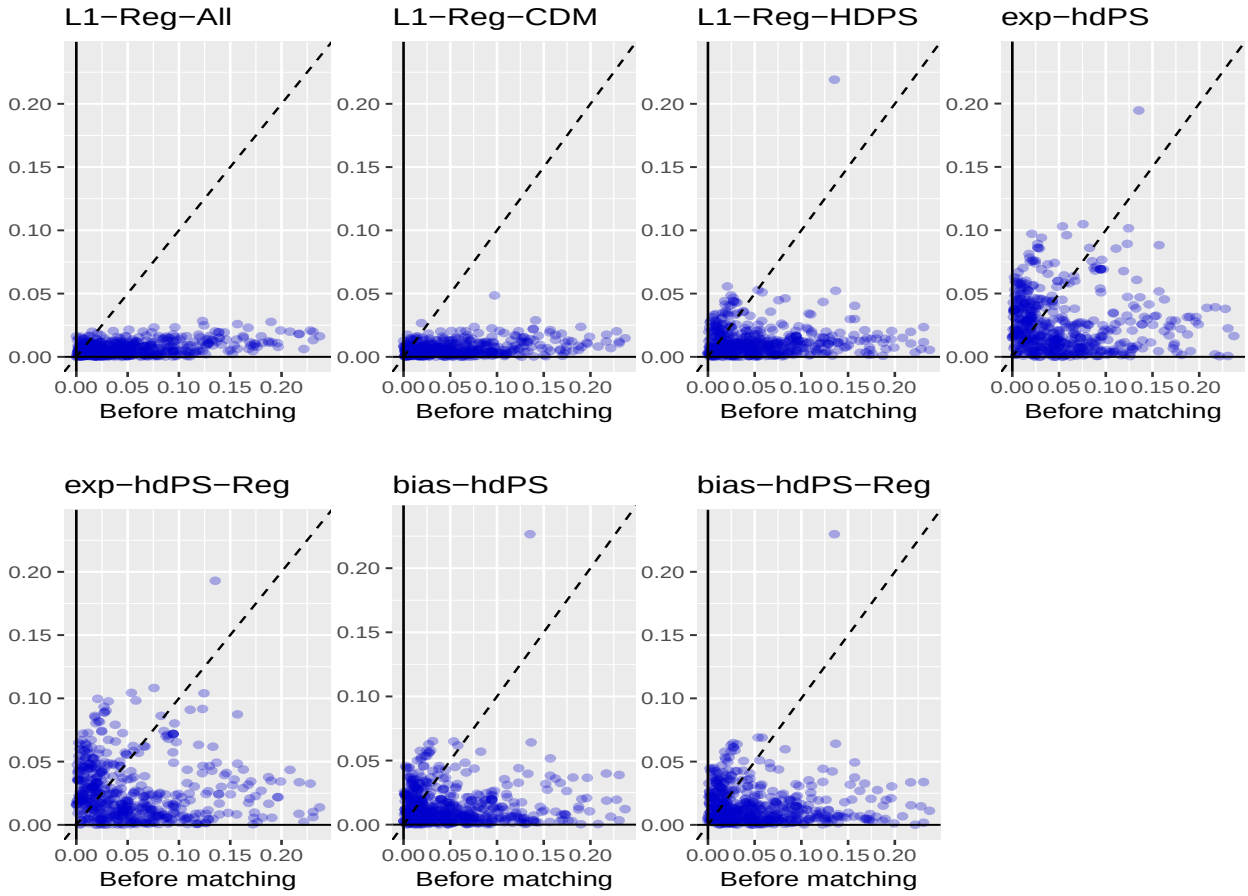


Figure 4.3: Anticoagulants study: before and after PS matching scatterplot of absolute standardized differences for synthetic model covariates. After matching outlier corresponds to higher indicators for "Condition Era Overlapping with Cohort Index: Atrial Fibrillation." in Warfarin group.

4.5.5 Simulation – Hazard Ratio Estimation

Figure 4.4 presents the log hazard ratio estimation bias and confidence interval coverage over 100 simulations for the two studies under varied simulation parameters. In general, all PS methods improve on estimation bias relative to unadjusted, the hdPS algorithm with and without regularization have similar estimates, L1-Reg-All and L1-Reg-CDM are similar, and L1-Reg-HDPS and bias-based hdPS are similar. There is a similar ordering of PS methods relative to the unadjusted estimate in both studies: exposure-based hdPS is closest to unadjusted, followed by L1-Reg-HDPS/bias-based hdPS, with L1-Reg-All/L1-Reg-CDM farthest. Coverage of the true HR expectedly mirrors the estimation bias, and is broadly higher in the NSAIDs study.

Both studies display a strong negative shift in bias with increasing true hazard ratio. This shift dominates the difference between PS methods, and no PS method uniformly has least bias across all simulation parameters. Because positively biased estimates shift past 0, this observation is not explained by our displaying raw instead of proportional bias. Instead, we believe it to be an artifact of the simulation process with its strict proportional hazards assumptions. In Web Appendix 8, we reproduce and explore this trend in the special case of 1-1 matching. Because real-world data do not necessarily conform to a proportional hazard model, this source of bias reveals a limitation of the "plasmode" and, by extension, our simulation framework.

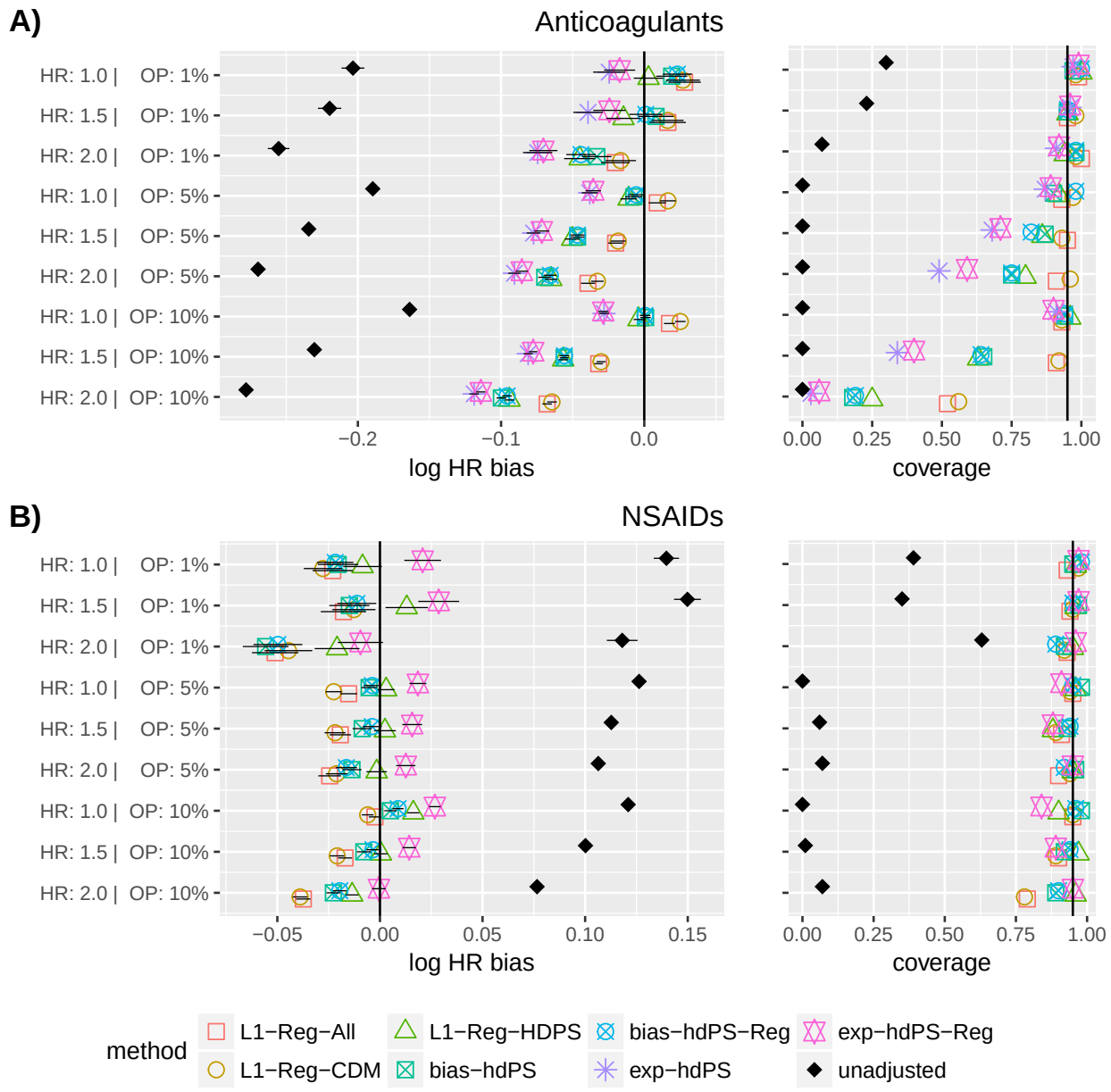


Figure 4.4: Bias in log hazard ratio (HR) with 1 standard deviation intervals, and coverage of true HR by 95% confidence intervals across 100 simulations for A) Anticoagulants and B) NSAIDs study under different simulation parameters of true HR and outcome prevalence (OP). Vertical line drawn at 0 bias and 95% coverage.

4.5.6 Negative Control - Hazard Ratio Estimation

Figure 4.5 shows the hazard ratio estimates and standard errors for the negative control outcomes for the Anticoagulants study. Estimates that lie above the dotted line include

the presumed true hazard ratio of 1 in their 95% confidence interval, and we consider these as "validated" negative control outcomes. In the absence of bias and negative control misspecification, we expect to validate 95% of the negative controls. Adjustment by any PS method substantially increases the number of validated outcomes relative to the unadjusted estimates. The three L_1-regularization methods and bias-hdPS validate between 86% and 90% of the negative control outcomes, while bias-hdPS-Reg, exp-hdPS, and exp-hdPS-Reg validate fewer, between 80% and 82%. The hdPS algorithm methods are able to construct existing PS models for all negative control outcomes. In the NSAIDs study (Web Appendix 5), the unadjusted estimates validate 83% of the negative control outcomes, and most PS methods do not improve significantly on this proportion, except L1-Reg-CDM at 97%. Both bias-hdPS and exp-hdPS demonstrate degrees of PS estimate nonexistence. These relative PS performance results are corroborated by Gaussian empirical null distributions fit to the negative control estimates (Web Appendix 4).

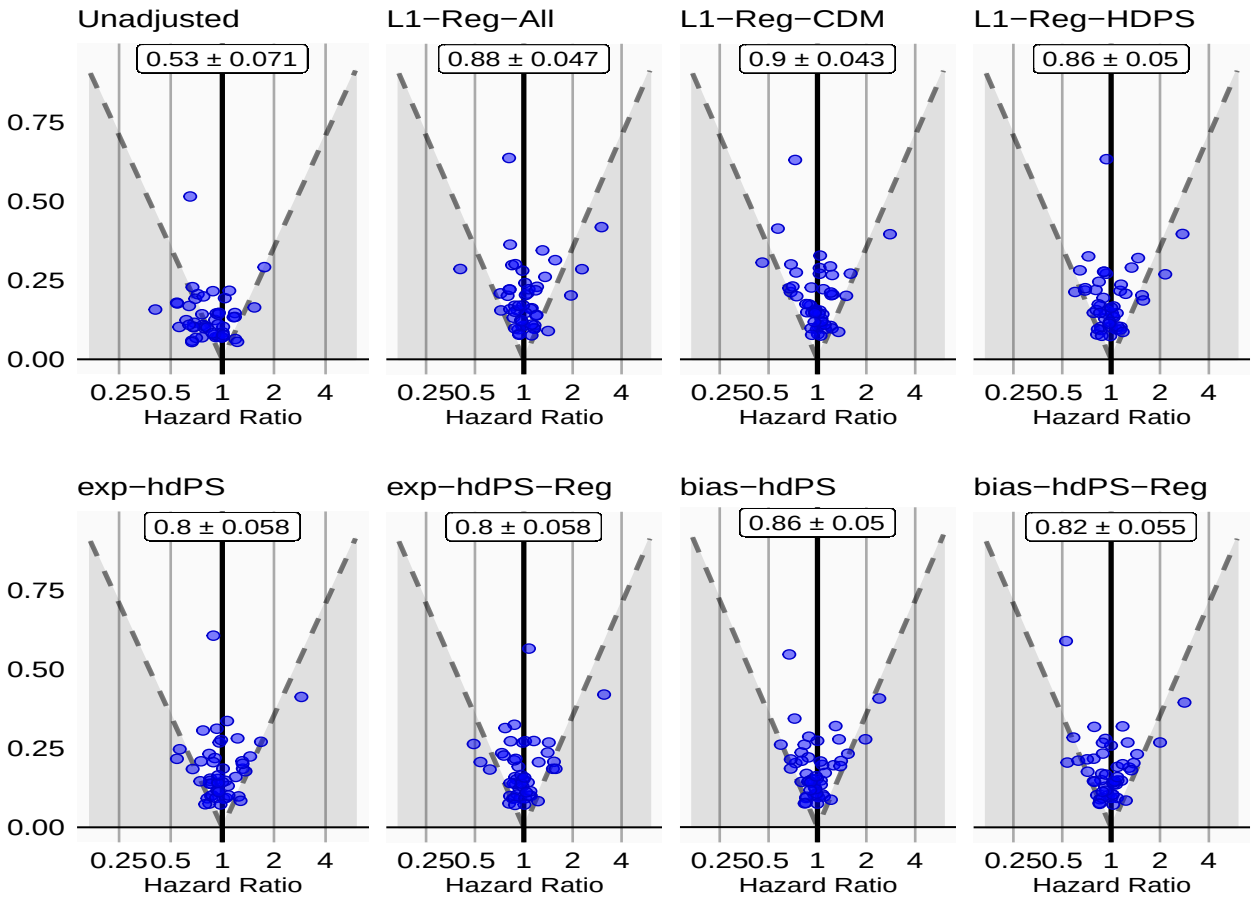


Figure 4.5: Anticoagulants study, estimates and standard errors for 49 negative control outcomes . Dashed line represents the boundary at where the 95% confidence interval does (above) or does not (below) contain the assumed true hazard ratio of 1. Coverage indicates proportion of intervals that contains 1.

4.6 Discussion

In this paper, we detail a combined synthetic and negative control framework for evaluating PS performance, and evaluate PS estimation methods that represent ideological opposites in automated selection: L_1-regularization performs model selection for all covariates simultaneously in a multivariable approach, while the hdPS algorithm relies on a univariate covariate screen. We find that L_1-regularization outperforms both bias-based and exposure-based hdPS algorithm on treatment prediction accuracy and covariate balance, with exposure-based hdPS algorithm having the worst performance. We also find that the use of a larger,

more comprehensive covariate set substantially improves treatment prediction accuracy and covariate balance. In the simulations, either L_1-regularization or the bias-based hdPS algorithm generally offers the least estimation bias, but results are strongly influenced by simulation parameters, revealing a limitation of the simulation framework. In the negative control experiments, PS adjustment offers significant improvement over unadjusted in one of two studies, in which L_1-regularization produces the closest to nominal number of validated negative controls and least biased empirical null distributions.

While defined as the probability of treatment assignment, propensity scores are used to reduce confounding bias by constructing covariate-balanced cohorts such as those inherent in randomized studies. So should PS estimation methods be judged on their success on outcome-independent metrics of treatment prediction and covariate balance, or outcome-dependent metrics of bias reduction? Bias reduction is the most immediate metric; after all, what good are treatment prediction and cohort balance if the PS cannot deliver unbiased estimates? However, the incorporation of outcome data can introduce arbitrary adjustment decisions and biases that complicate a clear comparison of PS methods, and methods that work well for one outcome may for another. In contrast, outcome-independent metrics are influenced by fewer study design decisions, and are more generalizable than outcome-dependent metrics determined on single outcomes.

Our survival simulation method extends the "plasmode" framework [44] by detailing additional distributional forms for the survival and censoring processes; proposing additional outcome prevalence adjustment methods; using non-informative, covariate-free censoring to avoid violating the proportional hazards model; and using the more accurate outcome prevalence Equation (4.2). While simulations benefit in having known effect sizes, even the most realistic simulations cannot capture the full complexity of real-world data. Our negative control outcome experiments entirely utilize real-world data for method evaluation, and avoid numerous simulation design choices that can introduce investigator bias. Granted, the use of control outcomes comes with the obvious concern of their misspecification, and uncertainty of their true effect sizes. The proper specification of negative and positive outcome controls will require continuous effort based in expert medical opinion, randomized trial results, and

testing across multiple databases.

Under the potential outcomes framework, the propensity score is a balancing score, such that its adjustment preserves unconfoundedness – the independence between treatment assignment and potential outcomes [9]. Unconfoundedness is violated by propensity score estimators that utilize outcome information such as the bias-based hdPS algorithm and the outcome-adaptive lasso [65], and by outcome-sensitive adjustment techniques such as disease risk scores [66]. To be clear, the data may well be unconfounded, and in the case of simulations they can be known to be; it is the PS estimator that discards unconfoundedness and that should be avoided if operating under the potential outcomes framework. Furthermore, outcome-dependent PS estimator performance should be evaluated using negative and/or positive outcome controls instead of simulations. Using simulated outcomes generated through a known process can favorably bias outcome-dependent PS estimators in an unrealistic and prophetic fashion. As an extreme example, one could construct the PS model with the exact covariates present in the synthetic model, and thus produce artificially unbiased effect estimates.

An argument in favor of outcome-dependent propensity scores, and more broadly investigator selected propensity score models, is the concern over pre-treatment variables that are uncorrelated with true confounders, strongly predict the treatment, and contribute no confounding on the outcome. These variables, sometimes known as "instrumental variables," promote treatment prediction without balancing confounders, and can inflate estimation bias or variance. The potential harmful effects of instrumental variables have been shown in theoretical examples and simulation experiments [34, 67]. However, the prevalence of instrumental variables in real-world data is debatable and their identification difficult. In our experiments, the bias-based hdPS algorithm that should avoid instrumental variables is not superior to L_1-regularized methods that include all available covariates, suggesting of a lack of instrumental variables. Comprehensive methods for instrumental variable identification and characterization in real-world observational data, and knowledge of the consequences on propensity score estimator selection, are still lacking and require further investigation.

The hdPS algorithm performs two functions: it presents tactics to address observational

data quality, and it utilizes a univariate screen for PS model selection. On the issue of data quality, the hdPS algorithm's separation of data sources, attention to coding granularity, and data augmentation by covariate frequency are clear-eyed approaches to the unique challenges of observational health data. On the issue of model selection, our results show the hdPS algorithm's univariate screen suffers from covariate interdependence in large-scale data. We show that hdPS estimate nonexistence, or "nonconvergence," is a problem in smaller sample sizes and with lower outcome prevalences, corroborating published observations [54, 55]. And, if there is enough covariate interdependence and collinearity to render the hdPS algorithm inoperable in smaller studies, there is no reason to believe covariate interdependence is not a serious problem despite algorithm convergence in larger studies. For example, in our studies L_1-regularization outperforms the hdPS algorithm in treatment prediction (AUC), despite the exposure-based hdPS explicitly selecting for marginal treatment associations. An undeniable benefit of a univariate screen is its computational efficiency; in our problem sizes, on the order of 100,000 subjects and 100,000 covariates, the hdPS algorithm can screen covariates in mere minutes. However, modern computational machinery increasingly handles large-scale regressions in observational health research. The Cyclops package [28] can run similarly sized, cross-validated logistic and Cox survival regressions in reasonable hours of compute time on ubiquitous personal computers. Computer parallelization and future statistical computing advances can further improve large-scale observational analyses, reducing computational burden as a barrier to utilizing appropriate methods.

CHAPTER 5

Evaluating Instrumental Variables in Propensity Score Models Using Synthetic and Negative Control Experiments

5.1 Introduction

Propensity scores (PS), an estimate of treatment assignment probability, are widely used for confounding control in observational studies [33, 21]. PS adjustment allows for the comparison of only similar treated and control persons in a cohort, thus approximating randomized experiments that are the gold standard in clinical evidence [22]. There remains controversy over the issue of variable selection for the PS model in high-dimensional datasets where the number of covariates can range from the hundreds to the many tens of thousands. Traditionally, clinical investigators construct a PS model using expert domain knowledge, including only covariates known or suspected to the investigators as confounders. However, this human-dependent process can be substantially and inexorably biased [57].

Various automated propensity score model selections exist to eliminate human bias from the task of selecting a parsimonious PS model out of thousands of available covariates. Still, concern remains over whether to include all pretreatment covariates in the automated selection process or whether to first curate them to only include "real" confounders that will ultimately reduce estimation bias in the outcome of interest. In particular, there are concerns over instrumental variables (IVs) that causally affect the outcomes only through their effect on the treatment [68]. Instrumental variables are covariates that are associated with treatment, independent of all confounders, and independent of the outcome conditional on

treatment and confounders [68]. When used to estimate average treatment effects, instrumental variables analysis provides unbiased bounds on the treatment effect size [69, 70, 71]. "Instrumental variables" is also more broadly to refer to variables that meet the mentioned criteria, and can be conditioned on, as in a propensity score, irrespective of conducting an IV analysis.

When used as conditioning variables, IVs are the source of "Z-bias," whereby they may increase bias from unmeasured confounders in observational data [31, 72]. As such, IVs are also known as "bias amplifiers" for amplifying existing residual bias after conditioning on other measured confounders [73, 67]. The potential deleterious effects of both IV have been shown in both theoretical frameworks [67, 74] and small simulation studies [75, 76, 77]. In addition to amplifying bias, conditioning on IVs may also reduce precision [34, 78].

While IVs can be easily simulated, it is controversial how prevalent IVs are and how to identify them in real-world data. IVs are sensitive to deviations from their unverifiable definitional assumptions [71, 79], and perfect IVs are difficult to identify for IV analyses [80]. In real-world observational health data, researchers often use provider characteristics as IVs, such as distance to health care facility or physician variation [81], but these IVs can be flawed and also unavailable in large-scale insurance claims databases that are used for many observational studies. In the absence of tools for identifying quality IVs, there is a movement to only include covariates associated with the outcome in propensity score models [82]. The popular high-dimensional propensity score (HDPS) [18] selects only covariates that have a high apparent relative risk with the outcome [83].

In this study, we conduct simulations and negative control experiments to explore the effect of IVs in real-world data and optimal PS models for reducing bias in the presence of IVs. By basing our simulations on real-world data, we utilize much larger models than those used in existing simulation studies. We also explore calendar year as a potential IV that would readily available in longitudinal data [84, 85]. Our contrasting PS models pit the approach of selecting covariates to purely predict the treatment to solely considering association with the outcome. In addition to reporting the simulation bias and precision, we measure effects on the residual bias using negative control outcomes [14, 15].

5.2 Methods

5.2.1 Clinical Study

We base our experiments on a real-world anticoagulants study of first time dabigatran to warfarin users among patients with atrial fibrillation from 2010 - 2018 using the Truven Health MarketScan Medicare Supplemental and Coordination of Benefits Database. The primary outcome is gastrointestinal bleeding. This is a reproduction of a published observational study [50] and follows a new user cohort study design [86, 38]. Extracted baseline patient covariates are encoded in the Observational Medical Outcomes Partnership (OMOP) Common Data Model Version 5 format [2], and indicator variables for demographics, conditions, procedures, drugs, observations, and measurements. See the Supplementary Material for more pretreatment covariate details. We use the CohortMethod R package [87] to construct the study cohort.

5.2.2 PS Models

We experiment with a large number of PS models, all of which are fit using large-scale regression models [19] of the real-world anticoagulants study through the Cyclops R package [28]. Six of these models do not include simulated IVs. Firstly, we conduct unadjusted analyses without a propensity score. Secondly, we use all measured covariates (All Covariates) in the PS model to maximize treatment prediction. Thirdly, we examine the fitted All Covariates PS model and select the calendar year covariate with the largest absolute coefficient, which is the indicator for 2010. This indicator for 2010 is a strong predictor for warfarin because dabigatran was only just coming onto the market at that time, and physician awareness and preference could have been a factor in dabigatran use. We then fit a large-scale Cox proportional hazards outcome model for GI bleed, and analyzed the model coefficients to confirm that 2010 has zero coefficient, thus no conditional association with the outcome. Fourthly, we exclude all calendar year indicators from the All Covariates model. Fifthly, we screen the most prevalent 500 covariates according to the HDPS apparent relative risk

criterion and only include in the PS model the 200 top ranked covariates. These covariates have the highest univariate relative risk with the outcome, and we call this covariate set the "HDPS Set." Sixthly, we use the fitted outcome model and include in the PS model only the covariates that have non-zero coefficients. These covariates have the highest conditional association with the outcome in a multivariate model, and we call this covariate set the "Cox Set."

The remaining PS models include a simulated IV. We use three baseline covariate sets corresponding to the second, fifth, and sixth PS models described above. To these PS models, we add a single simulated IV with one of three prevalences ($p = 0.025\%, 0.05\%, 0.1\%$), and one of three relative risks with the treatment variable ($r = 1.5, 2, 4$). For each of the three baseline covariate sets, there are 9 additional PS models, one for each prevalence-relative risk combination. There are a net total of 33 PS models, listed in Table 5.1. DAGs representing simulations using simulated outcomes are shown in Figures 5.1A and 5.1C.

1. Unadjusted	7. All + 0.025/1.5	16. HDPS + 0.025/1.5	25. Cox + 0.025/1.5
2. All Covariates	8. All + 0.025/2	17. HDPS + 0.025/2	26. Cox + 0.025/2
3. No 2010	9. All + 0.025/4	18. HDPS + 0.025/4	27. Cox + 0.025/4
4. No Years	10. All + 0.05/1.5	19. HDPS + 0.05/1.5	28. Cox + 0.05/1.5
5. HDPS Set	11. All + 0.05/2	20. HDPS + 0.05/2	29. Cox + 0.05/2
6. Cox Set	12. All + 0.05/4	21. HDPS + 0.05/4	30. Cox + 0.05/4
	13. All + 0.1/1.5	22. HDPS + 0.1/1.5	31. Cox + 0.1/1.5
	14. All + 0.1/2	23. HDPS + 0.1/2	32. Cox + 0.1/2
	15. All + 0.1/4	24. HDPS + 0.1/4	33. Cox + 0.1/4

Table 5.1: Evaluated PS models. Simulated IVs have prevalence p and relative risk with treatment r represented as p/r. Models 7-15 add a simulated IV to the Model 2. Models 16-24 add a simulated IV to Model 5. Models 25-33 add a simulated IV to Model 6.

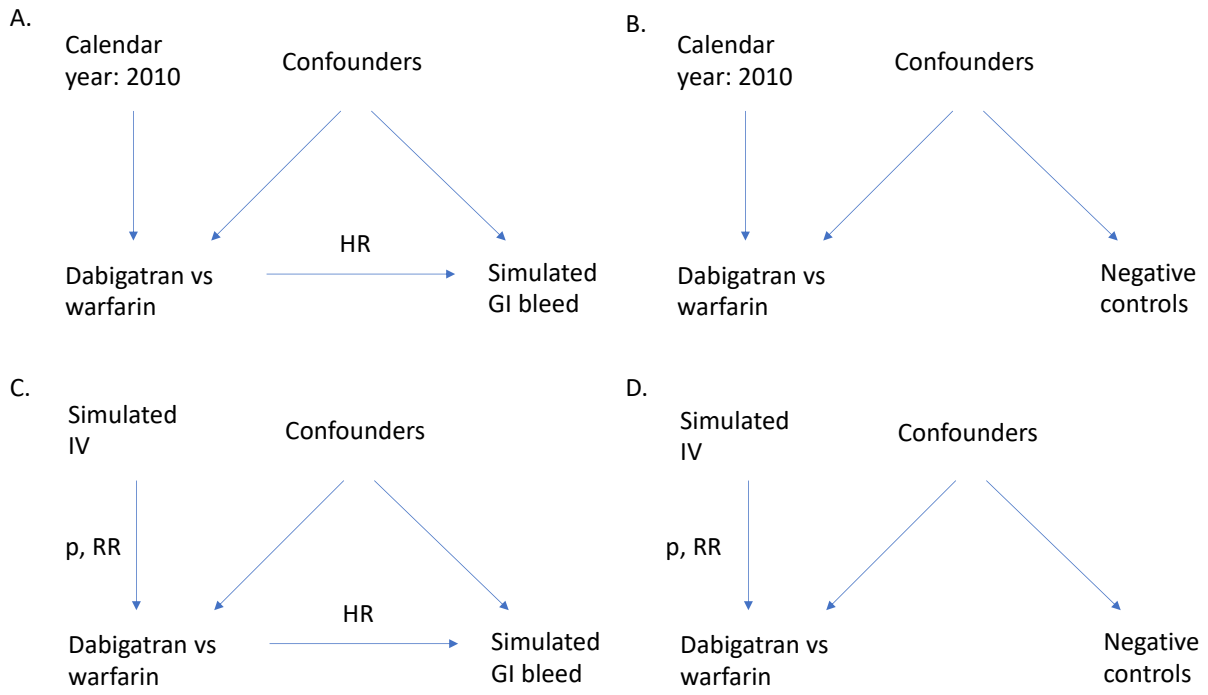


Figure 5.1: A) directed acyclic graph (DAG) showing calendar year 2010 as an IV affecting estimation of hazard ratio (HR) of simulated outcomes of GI bleed. B) DAG showing negative control outcomes with no presumed effect from compared treatments. C) DAG showing effects of simulated IV under specified prevalence and relative ratio on the HR estimate of simulated outcome. D) DAG showing effects of simulated IV on negative control estimates

5.2.3 Outcome Simulations

Using the real-world anticoagulants study as a simulation framework, we simulate new outcomes according to a "plasmode" design [44, 45]. The specific plasmode design we employ is detailed in [19]. We first fit the data to a Cox proportional hazards model to obtain a realistic survival model for outcome simulation, including covariate coefficients and survival functions for the outcome and censoring events. Keeping the original covariates, we calculate each subject's linear predictor, and simulate outcome times and censoring times under a desired true hazard ratio. We simulate under four true hazard ratios (1, 1.5, 2, 4). With each fitted PS model, we perform variable length PS matching [59] with a maximum ratio of 10:1 and a caliper of 0.2, and use a greedy matching algorithm [60]. We then fit a PS-stratified Cox proportional hazard model with the simulated outcomes, to obtain point estimates and 95% confidence intervals of the treatment effect size.

5.2.4 Negative Controls

Negative control outcomes are an emerging tool in observational research that allow for experiments using real data by providing a standard of clinical truth – that of no effect between exposure and outcome [14, 15]. In a clinical observational study setting, a negative control is an outcome that investigators can determine, with some confidence, is not differentially affected by the active treatment or reference treatment. Effect estimation on a large set of negative control outcomes provides a distribution whose deviation from the expected null effect approximates the systemic study bias, or residual bias after controlling for measured confounding [10].

We identify 49 negative control outcomes through a data-rich algorithm [61] combined with manual curation. Similar to the simulated outcomes, we perform variable ratio PS matching and fit PS-stratified Cox proportional hazards models for the negative control outcomes. We fit the set of negative control estimates – each presumed to have a true hazard ratio of 1 – to an empirical null distribution [10]. This distribution characterizes the study residual bias after PS adjustment [11] and arises from both unmeasured confounding and inappropriate control of measured confounding, such as inclusion of IVs. DAGs representing simulations using negative control outcomes are shown in Figures 5.1B and 5.1D.

5.2.5 Metrics

For both simulated and negative control outcomes, we compare the bias and standard deviation (SD) of the estimated hazard ratios to the true hazard ratios (known for the simulated outcomes and presumed to be 1 for the negative control outcomes). We fit the negative control estimates to empirical null distributions, and report the distribution means and SDs. To assess how the instrumental variables affect the covariate balance through the PS, we compare before and after matching standardized mean differences (SMDs) for all covariates. We plot the distribution of SMDs and also note the number of after-matching SMDs that cross a threshold of 0.1. We also plot the distributions of the fitted PS models.

5.3 Results

In the anticoagulants study, there are 20474 first-time dabigatran users and 56648 first-time warfarin users. There are 52729 total unique covariates in the All Covariates set, of which 900 have nonzero coefficients in the fitted PS model. The HDPS Set of covariates contains the 200 covariates with the highest apparent relative risk out of the 500 covariates with the highest prevalence. Of these, 170 have nonzero coefficients in the fitted PS model. The Cox Set of covariates obtained from a large-scale outcome model contains 74 covariates, and 73 of them have nonzero coefficients in the fitted PS model. There are 31 covariates that overlap between the PS model for All Covariates and the Cox PS model, 26 that overlap between the HDPS PS model and the Cox PS model, and 93 that overlap between the All Covariates PS model and the HDPS PS model.

Figure 5.2 plots for the PS models the preference score distributions that normalize propensity scores by their prevalence. There are few discernible differences among the PS plots for All Covariates and All Covariates with 2010 removed and with all calendar years removed. These three PS distributions show moderately strong differentiation between the dabigatran and warfarin populations. In contrast, the PS distributions built on the HDPS Set and Cox Set of covariates show more overlap between distributions. All three distributions with a simulated instrumental variable have large spikes in the dabigatran distribution close to 1, showing that the simulated IV has a strong effect on the PS distributions. On inspection of the PS models over 100 simulations, every single simulated PS model includes the simulated IV as a covariate with nonzero coefficient.

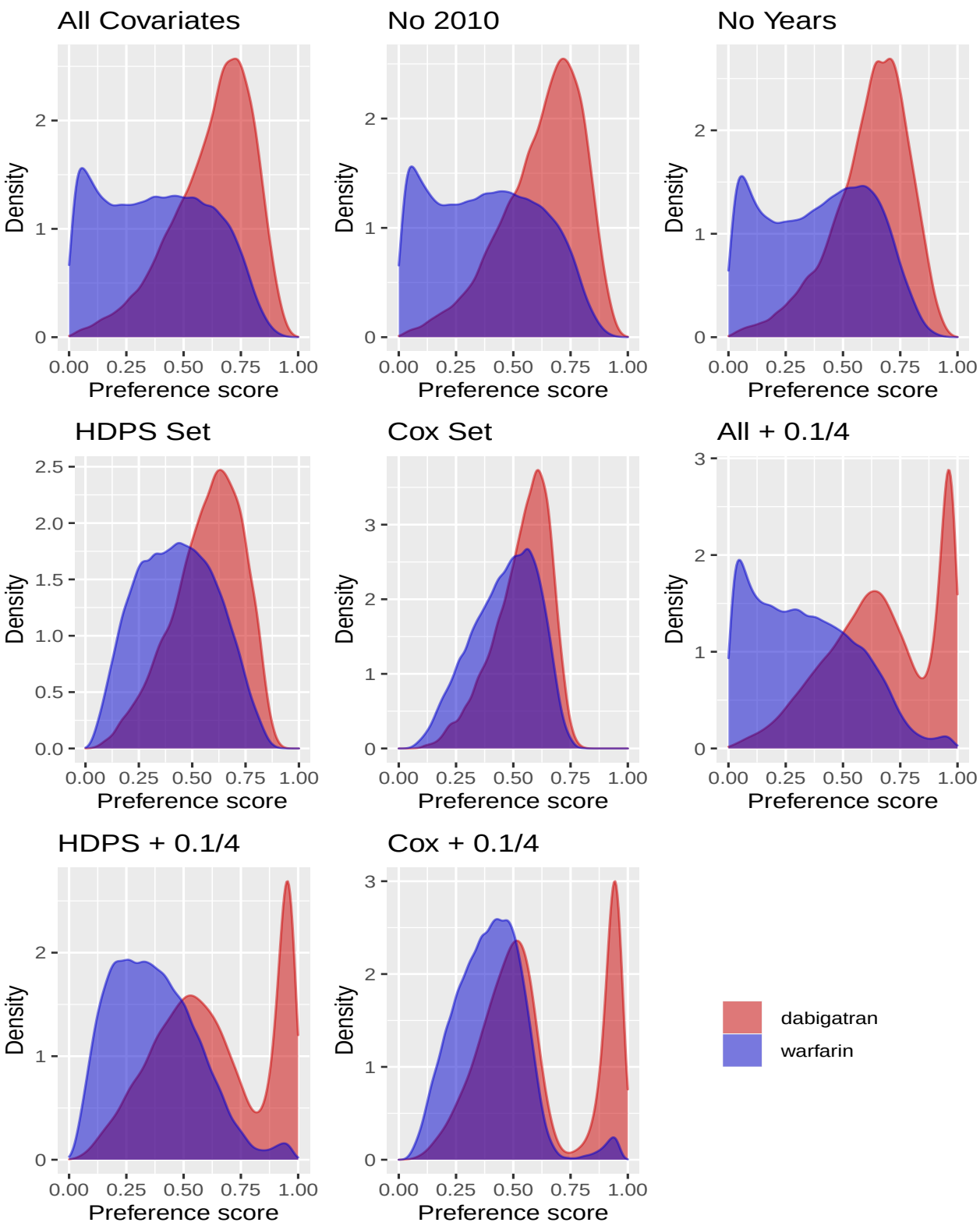


Figure 5.2: Propensity score distributions represented as preference scores that normalize propensity scores by prevalence. The three plots from simulated IV PS models are taken from simulations with 10% IV prevalence and relative risk of 4.

Simulation results under a true hazard ratio of 4 are shown in Figure 5.3, with detailed mean and SD provided in Table 5.2. The unadjusted estimate has by far the largest bias and lowest coverage of the true effect size of all compared methods. Relative to All Covariates, removing calendar year 2010 very slightly increases the bias, and removing all calendar years increases the bias even more. The HDPS Set has smaller bias and variance than All Covariates, and the Cox Set has almost no bias and even smaller variance. For simulations with simulated IV based on All Covariates and HDPS Set, increasing the simulated IV prevalence and relative risk actually decreases the study bias, while having mixed effects on the variance. The simulations with simulated IV added to Cox Set has the smallest bias, followed by those added to HDPS Set, then those added to All Covariates. All methods other than unadjusted have high coverage of the true effect size. Simulation results under the other three simulated true hazard ratios display similar patterns and are shown in the Supplementary Material.

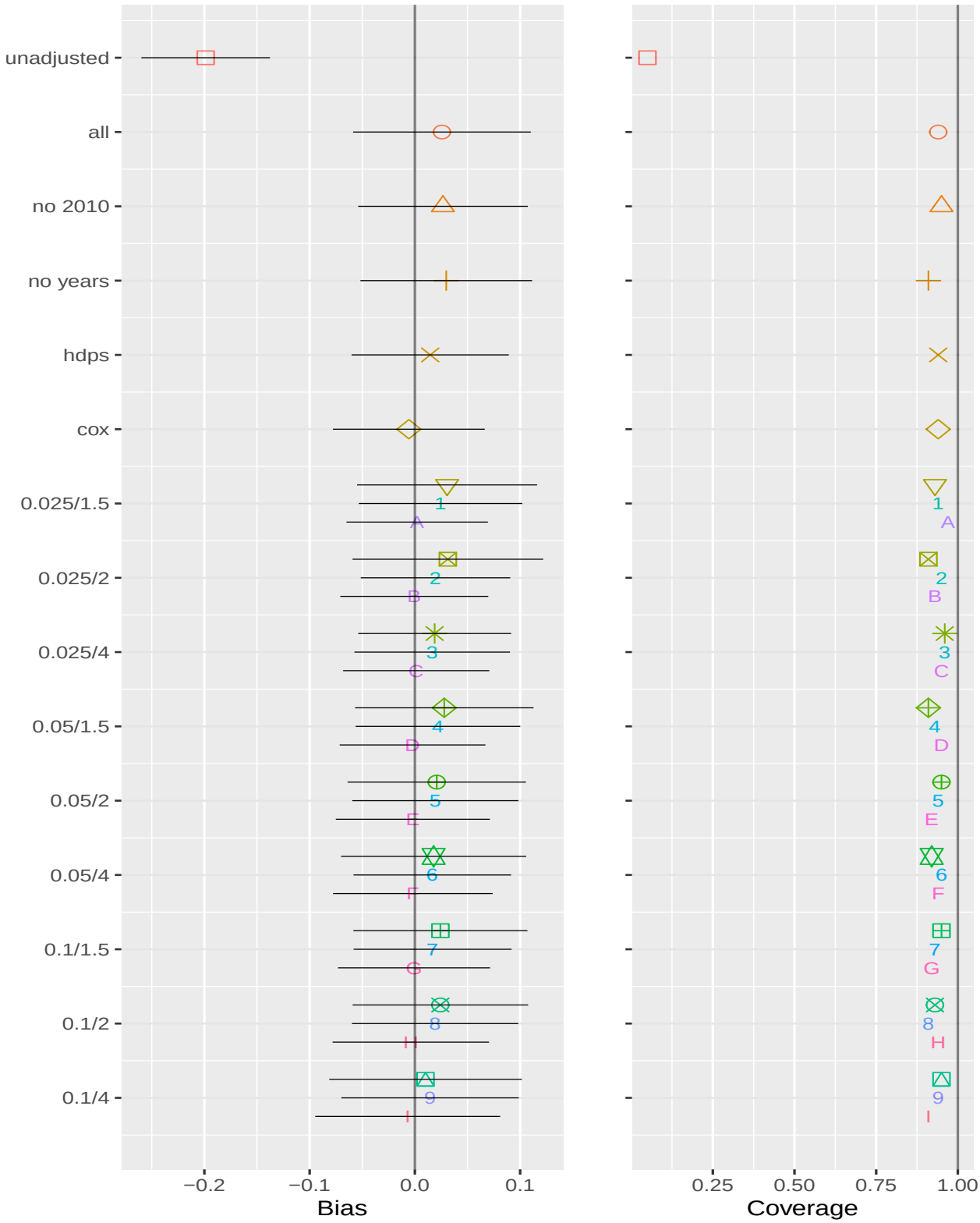


Figure 5.3: Left: bias and SD of simulation experiments with true hazard ratio of 4. Right: coverage of true effect size of $HR = 4$ across 100 simulations. For the 9 simulated IV settings, the shapes represent All Covariates, the numbers represent HDPS set, and the letters represent Cox set.

1. Unadjusted	-0.199 (0.061)	16. HDPS + 0.025/1.5	0.024 (0.078)
2. All Covariates	0.026 (0.084)	17. HDPS + 0.025/2	0.020 (0.071)
3. No 2010	0.027 (0.081)	18. HDPS + 0.025/4	0.016 (0.074)
4. No Years	0.030 (0.082)	19. HDPS + 0.05/1.5	0.022 (0.078)
5. HDPS Set	0.015 (0.075)	20. HDPS + 0.05/2	0.019 (0.079)
6. Cox Set	-0.006 (0.072)	21. HDPS + 0.05/4	0.017 (0.075)
		22. HDPS + 0.1/1.5	0.017 (0.075)
		23. HDPS + 0.1/2	0.019 (0.079)
		24. HDPS + 0.1/4	0.014 (0.084)
7. All + 0.025/1.5	0.031 (0.086)	25. Cox + 0.025/1.5	0.002 (0.067)
8. All + 0.025/2	0.031 (0.091)	26. Cox + 0.025/2	-0.001 (0.070)
9. All + 0.025/4	0.019 (0.073)	27. Cox + 0.025/4	0.001 (0.069)
10. All + 0.05/1.5	0.028 (0.085)	28. Cox + 0.05/1.5	-0.002 (0.069)
11. All + 0.05/2	0.021 (0.085)	29. Cox + 0.05/2	-0.002 (0.073)
12. All + 0.05/4	0.018 (0.088)	30. Cox + 0.05/4	0.002 (0.076)
13. All + 0.1/1.5	0.024 (0.083)	31. Cox + 0.1/1.5	-0.001 (0.072)
14. All + 0.1/2	0.024 (0.083)	32. Cox + 0.1/2	-0.004 (0.074)
15. All + 0.1/4	0.010 (0.091)	33. Cox + 0.1/4	-0.007 (0.088)

Table 5.2: Simulation bias for true $HR = 4$, as Mean (SD), for all PS models

While the above results represent performance under a known outcome model, the negative control distributions approximate the residual study bias. The null distribution means and SD are shown in Figure 5.4 and Table 5.3. The unadjusted estimate has by far the largest bias (deviation from 0 mean) and variance, and the lowest coverage of unity HR by the individual negative control estimates. Among PS models without simulated IVs, All Covariates has the smallest bias, while removing calendar year 2010 and all calendar years creates increasingly larger bias. HDPS Set has larger bias and variance than All Covariates, though higher coverage. Meanwhile, Cox Set has even larger bias and variance than HDPS Set and lower coverage than All Covariates. Among PS models with a simulated IV, for each

IV prevalence and relative risk the All Covariates estimate has smaller bias and variance and lower coverage than the HDPS Set estimate, which in turn has smaller bias and variance and lower coverage than the Cox Set estimate. Adding an instrumental variable to All Covariates increases both the bias and variance but also the coverage, though the magnitude of increase is not clearly associated with the strength of the IV. However, adding an IV to the HDPS Set slightly decreases the bias overall, and noticeably increases the variance and lowers the coverage. Finally, adding an IV to the Cox Set decreases the bias and variance, and manages to increase the coverage under some settings. Increasing the relative risk of the simulated IV slightly increases the coverage throughout.

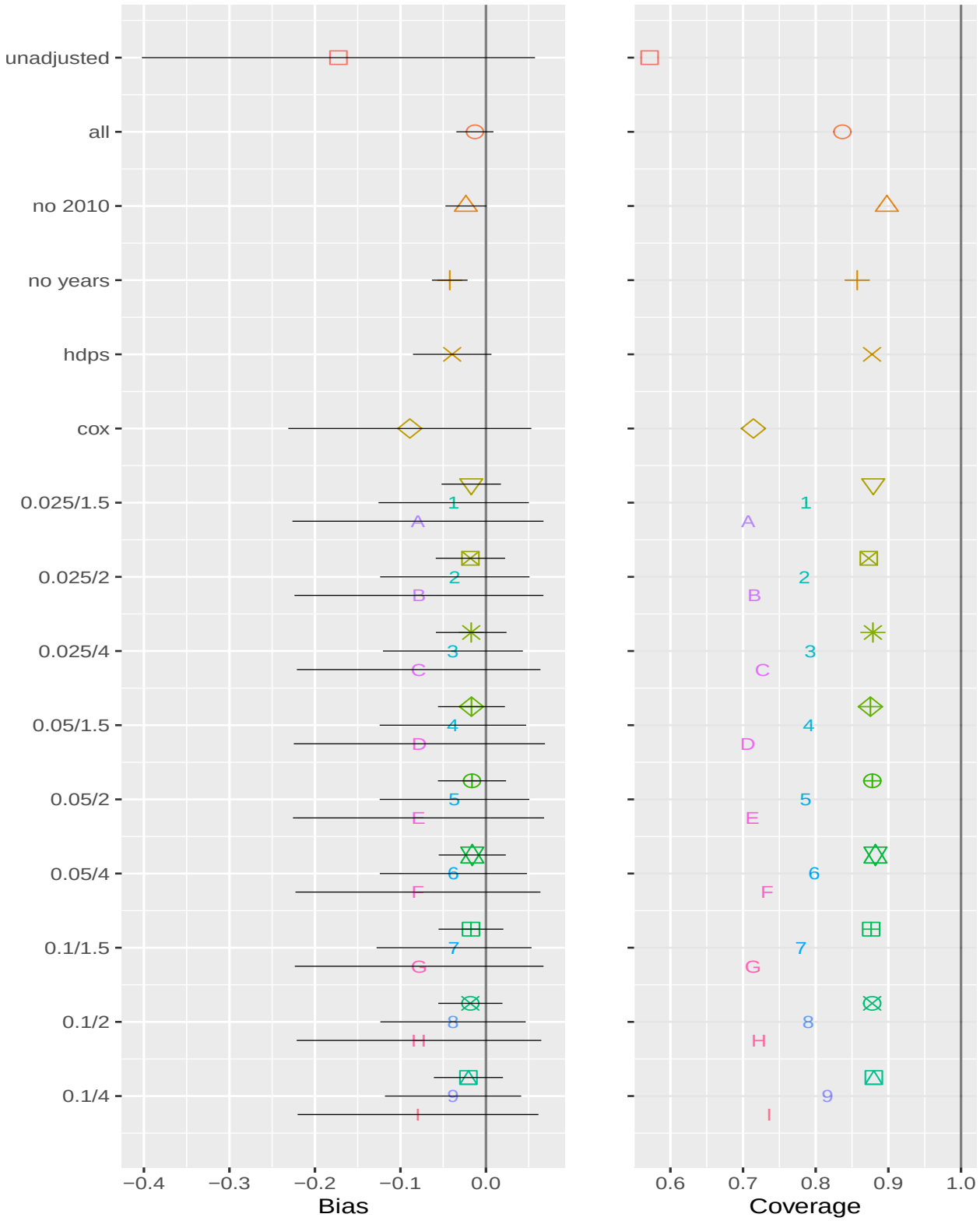


Figure 5.4: Left: mean and SD of fitted negative control distributions, characterizing residual study bias. Right: coverage of presumed true effect size of 1 HR by negative control estimates. For the 9 simulated IV settings, the shapes represent All Covariates, the numbers represent HDPS set, and the letters represent Cox set.

1. Unadjusted	-0.173 (0.229)	16. HDPS + 0.025/1.5	-0.038 (0.088)
2. All Covariates	-0.013 (0.028)	17. HDPS + 0.025/2	-0.036 (0.087)
3. No 2010	-0.023 (0.024)	18. HDPS + 0.025/4	-0.039 (0.082)
4. No Years	-0.042 (0.017)	19. HDPS + 0.05/1.5	-0.039 (0.086)
5. HDPS Set	-0.039 (0.051)	20. HDPS + 0.05/2	-0.037 (0.088)
6. Cox Set	-0.089 (0.147)	21. HDPS + 0.05/4	-0.038 (0.086)
		22. HDPS + 0.1/1.5	-0.037 (0.091)
		23. HDPS + 0.1/2	-0.038 (0.085)
		24. HDPS + 0.1/4	-0.038 (0.080)
7. All + 0.025/1.5	-0.017 (0.035)	25. Cox + 0.025/1.5	-0.079 (0.147)
8. All + 0.025/2	-0.018 (0.041)	26. Cox + 0.025/2	-0.078 (0.146)
9. All + 0.025/4	-0.017 (0.041)	27. Cox + 0.025/4	-0.079 (0.142)
10. All + 0.05/1.5	-0.017 (0.039)	28. Cox + 0.05/1.5	-0.078 (0.147)
11. All + 0.05/2	-0.016 (0.040)	29. Cox + 0.05/2	-0.079 (0.147)
12. All + 0.05/4	-0.016 (0.039)	30. Cox + 0.05/4	-0.080 (0.143)
13. All + 0.1/1.5	-0.018 (0.038)	31. Cox + 0.1/1.5	-0.078 (0.145)
14. All + 0.1/2	-0.018 (0.038)	32. Cox + 0.1/2	-0.078 (0.143)
15. All + 0.1/4	-0.020 (0.040)	33. Cox + 0.1/4	-0.079 (0.141)

Table 5.3: Negative control distributions, as Mean (SD), for all PS models

Propensity scores reduce confounding by creating comparable cohorts that are balanced with respect to pretreatment covariates. Figure 5.5 shows the covariate balance for the All Covariates set of covariates. The All Covariates PS model does the best in balancing the covariates, and removing calendar year 2010 or all calendar years from the PS model results in the respective calendar years becoming unbalanced. The HDPS Set PS model does a worse job with covariate balance, and the Cox Set PS model does an even worse job. Adding a strong IV to the All Covariates, HDPS Set, and Cox Set PS models has very little effect on the covariate balance distribution, even though we have seen it has a strong effect on the PS distribution (Figure 5.2). Figure 5.6 shows the covariate balance of just the HDPS

Set covariates. The HDPS Set PS model does the best at balancing covariates, and the All Covariates PS model also keeps all after-matching standardized differences below 0.05. However, the Cox Set PS model fares poorly on these covariates' balance, and fails to balance numerous covariates.

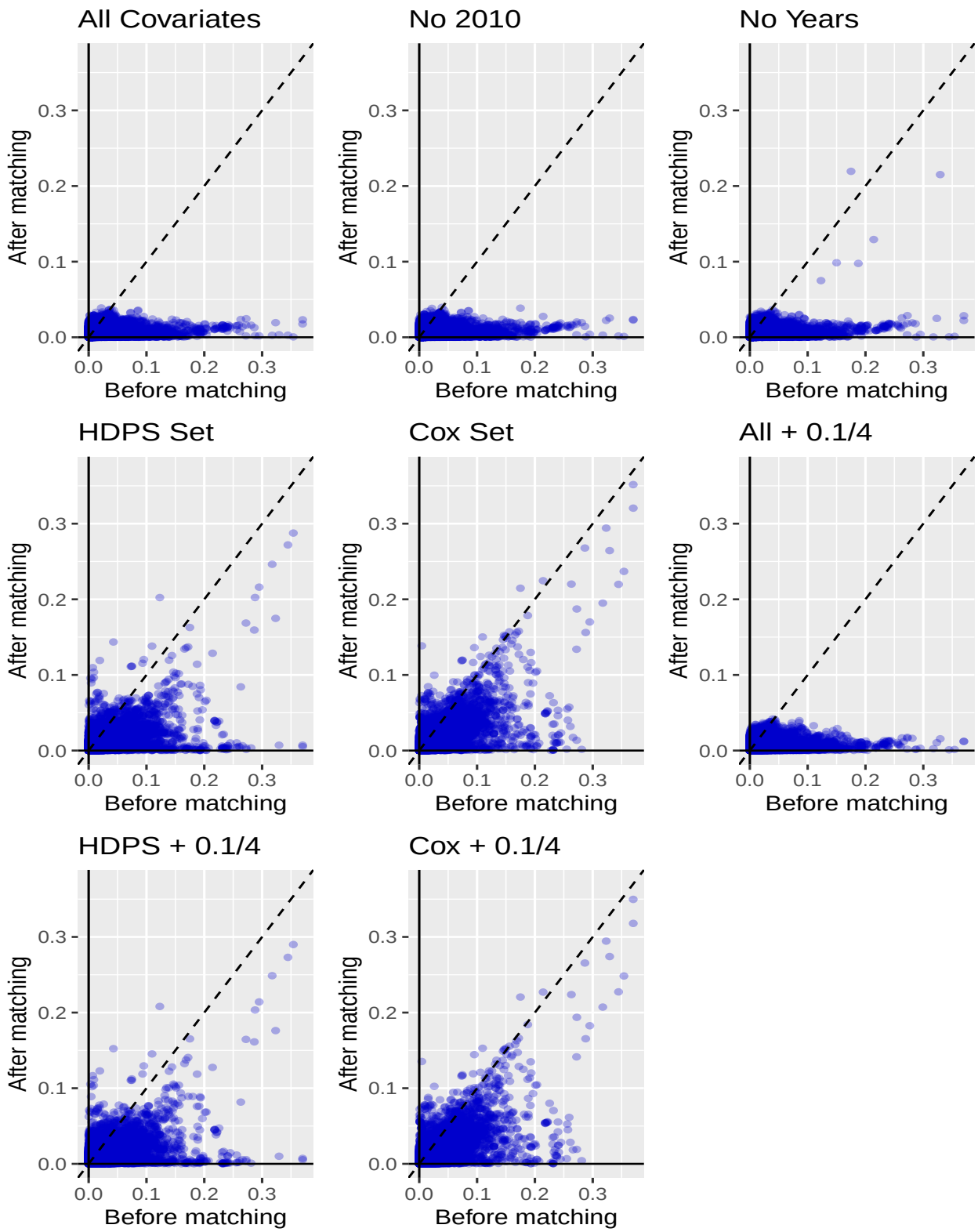


Figure 5.5: Pre-matching vs post-matching covariate absolute standardized differences for All Covariates. Each point represents one covariate. The three plots from simulated IV PS models are taken from simulations with 10% IV prevalence and relative risk of 4.

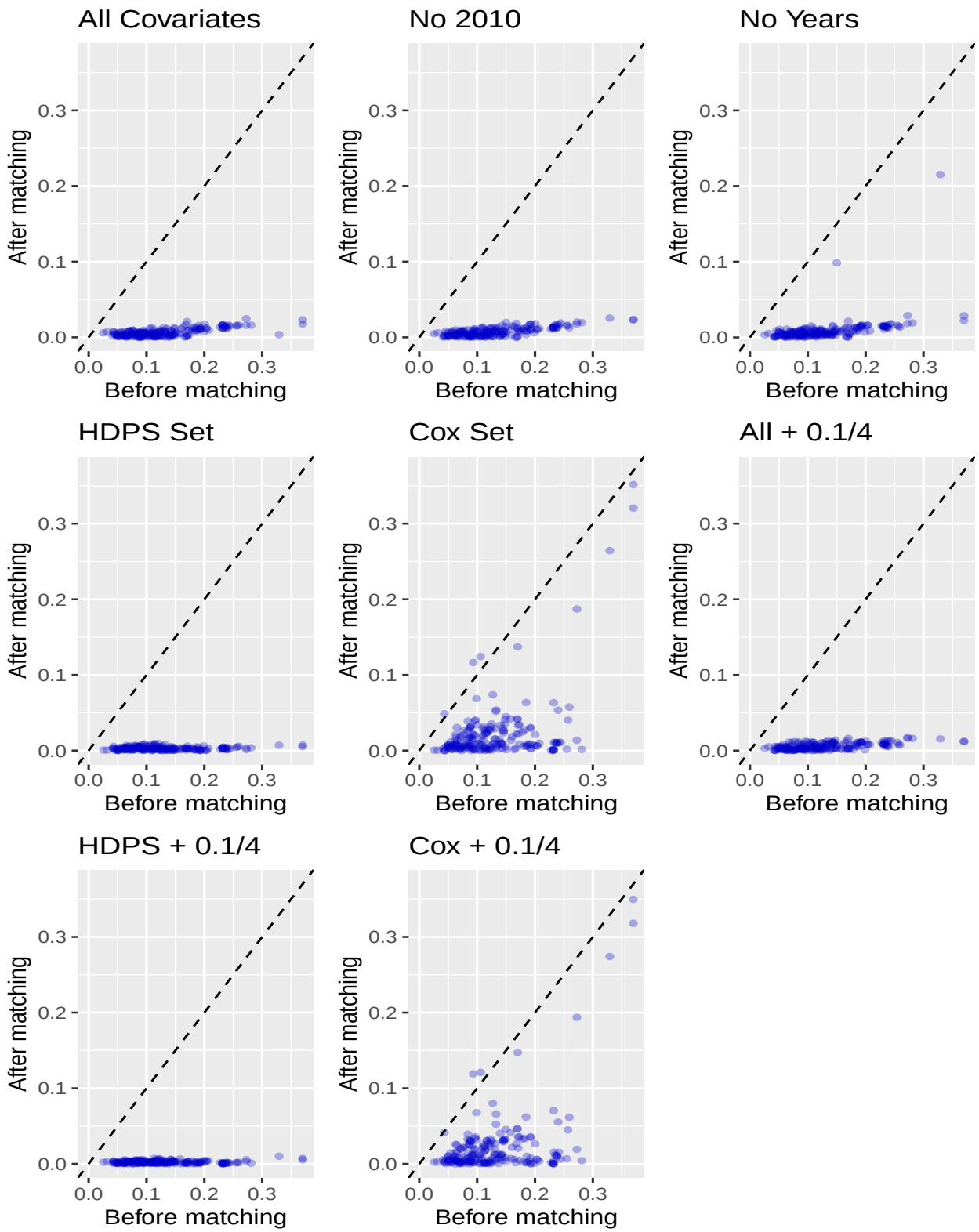


Figure 5.6: Pre-matching vs post-matching covariate absolute standardized differences for HDPS Set covariates. Each point represents one covariate. The three plots from simulated IV PS models are taken from simulations with 10% IV prevalence and relative risk of 4.

Figure 5.7 shows the negative control outcome estimates generated by the PS models, along with the coverage by the estimates of the presumed true hazard ratio of 1. Nominally, 95% of the estimates' 95% confidence intervals should include 1. The unadjusted estimates have the smallest coverage, and have a mean estimate that is noticeably negative. At 84%, the All Covariates PS model has higher coverage than the HDPS Set at 61% and the Cox Set at 71%. Removing calendar year 2010 and all calendar years both increase the coverage from the All Covariates PS model. Adding a simulated IV to the All Covariates, HDPS Set, and Cox Set PS models increases the number of negative control estimates that produce nonsignificant confidence intervals.

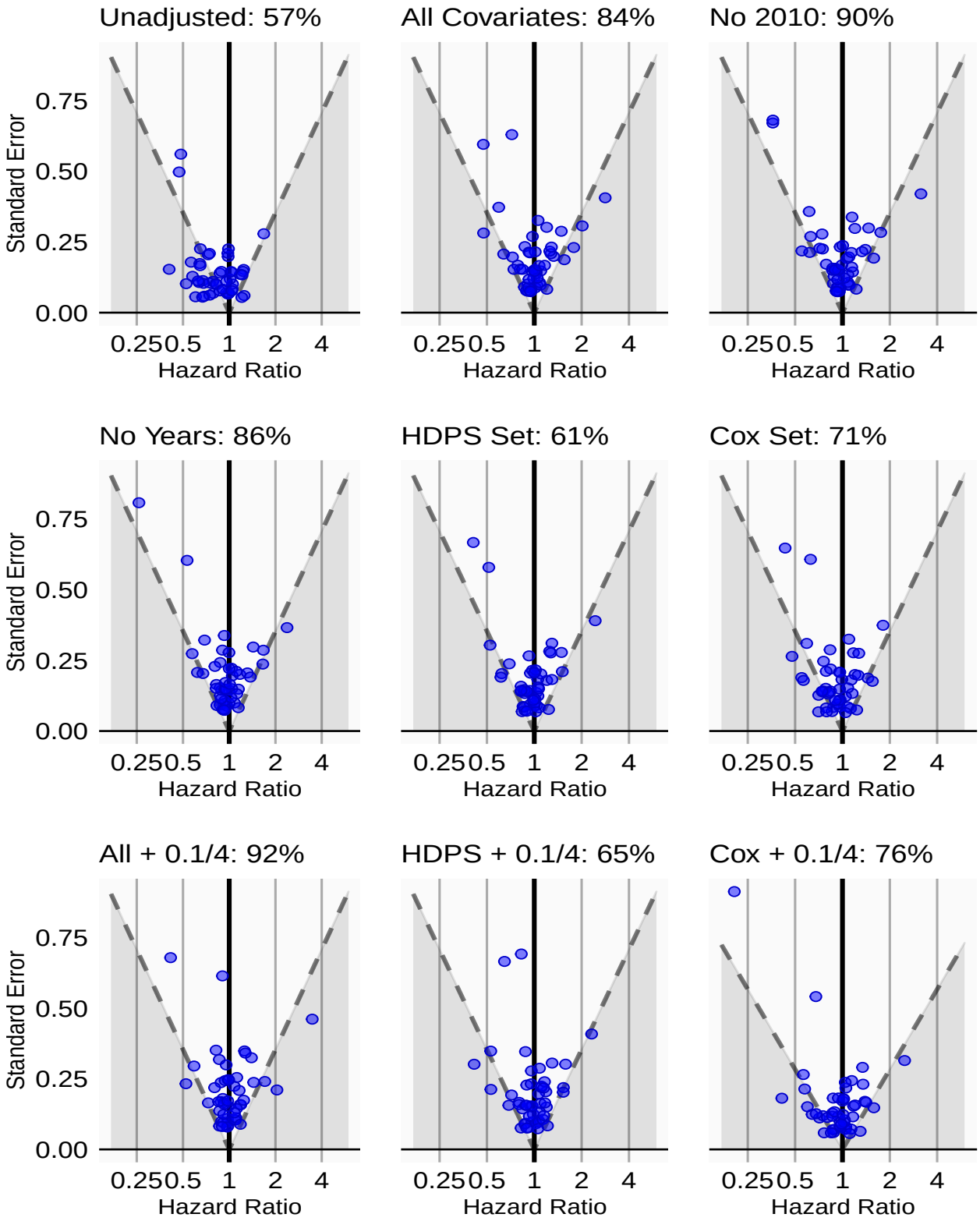


Figure 5.7: Negative control outcome estimates with associated coverage of presumed true hazard ratio of 1. Each point represents one negative control estimate. Estimates above the dotted lines include 1 in their 95% confidence intervals and are not statistically significant, while estimates below the dotted lines no not include 1 and are statistically significant.

5.4 Discussion

The propensity score is defined as an estimate of treatment assignment probability using pretreatment covariates, and its use is sufficient in removing bias from all observed covariates [21, 9]. In other words, knowing the true treatment assignment probability would allow for perfectly unbiased outcome effect estimates in observational studies, through the design of stratified studies that approximate randomized trials [22]. Knowing this, it makes definitional sense to build the PS model with the goal of treatment prediction, and include only pretreatment covariates; the outcome, which postdates the treatment, would have no role in the PS model [23, 24]. Large scale regularized regressions are a natural approach to estimating the PS in the presence of thousands of covariates, as are available in longitudinal observational databases [19].

Instrumental variables are established to be bias amplifiers in numerous theoretical and simulation studies [67, 74, 75, 76, 77]. There seems to be agreement that known IVs should be removed from the set of conditioning variables [88]. The question becomes, how can we identify IVs in real-world observational data that adhere to multiple unverifiable causal criteria? Faced with this dilemma, some authors have abandoned the purpose of the PS as a treatment prediction model, and advocate for including covariates based on association with outcome [82]. This approach embraces a novel ideology for PS estimation: fit an outcome model, use the results to build a PS, and use the PS in another outcome model. The high-dimensional propensity score takes this to an extreme in utilizing a univariate screen to identify the most outcome-associated covariates for PS model inclusion [18]. The HDPS has become a common tool for automated PS model construction [36, 89].

We observe that sacrificing treatment prediction to create PS model covariate sets based on outcome association has expected consequences in PS distribution quality. Larger PS models do increasingly better jobs at separating PS distributions of the target and comparator treatment (Figure 5.2). The All Covariates PS model, built on thousands of covariates, achieves the most separation between dabigatran and warfarin populations, while leaving enough overlap to allow for meaningful study comparison. In contrast, the HDPS Set PS

model, built on 200 covariates identified through a univariate screen for outcome association, does not identify a large set of warfarin patients with low preference score. The Cox Set PS model, built on 74 covariates identified in a multivariate regularized outcome model, has very little preference score separation between the two groups. For all three PS models, inclusion of a simulated IV causes a dramatic spike in the preference score distribution at the high end of the scale near 1, showing that indeed the simulated IV strongly affects treatment assignment probability.

One might presume that a simulated IV dominating the PS distribution would diminish the relative contribution of other covariates in the PS model and have a detrimental effect on other covariates' balance. Surprisingly, we observe that inclusion of a simulated IV has almost no perceptible effect on covariate balance (Figures 5.5 and 5.6). We do observe a similar relationship between PS model size and covariate balancing performance: the larger the PS model, the more covariates are successfully balanced (Figure 5.5). The All Covariates PS model is built on all covariates, and even though the resultant PS model only has 900 nonzero coefficients, all covariates are satisfactorily balanced. The HDPS Set and Cox Set PS models fail to restrict all after-matching standardized differences to below 0.1. When we observe the covariate balance only for the 200 HDPS Set covariates, the HDPS Set PS model unsurprisingly performs the best covariate balancing (Figure 5.6). However, the All Covariates PS model also performs excellently, showing that including [even vastly] more PS model covariates does not compromise the balance of a smaller subset of covariates that may be of interest to the investigator.

In our plasmode simulations under a known true hazard ratio, the Cox Set PS model unsurprisingly demonstrates the least bias, as it builds the PS using the exact covariates used to build the parametric outcome generating model (Figure 5.3 and Table 5.2). Surprisingly, the HDPS Set PS model has slightly smaller bias than the All Covariates PS model, even though the HDPS Set PS model does a poorer job balancing the 74 covariates of the Cox Set that are used for simulated outcome generation. The HDPS Set PS model also has fewer covariates, 26, that overlap with the Cox Set than the All Covariates PS model, at 31. These results suggest that there is merit to selecting PS model covariates by outcome association

when it comes to study bias. Interestingly, removal of our suspected IV calendar year 2010 from the All Covariates PS model increases – rather than decreases – study bias, and removal of all calendar years further increases the bias. Additionally, inclusion of a simulated IV to the All Covariates and Cox Set PS models often decreases the study bias. Most strikingly, the All Covariates PS models with simulated IV have smaller bias with stronger and more prevalent simulated IV. Inclusion of simulated IV does seem to generally increase variance across observed PS models.

Instrumental variables are widely known as bias amplifiers [67], yet our simulation results show them having the opposite effect: removing suspected calendar year IVs slightly increases bias, while adding a simulated IV sometimes decreases bias. We notice that published simulation studies utilize small simulation models in which the IV is one of a few – if not the only – simulated covariates [75, 76, 77]. Meanwhile, we are adding simulated IVs (or removing suspected IVs) from much larger models with at least 74 covariates and up to tens of thousands of covariates. Our large PS models more accurately reflect real-world scenarios in which longitudinal observational databases provide many thousands of potential confounding covariates. While we cannot explain the observed paradoxical IV effects, we believe that IVs have much weaker effect in real-world data than in small simulations. A similar view, that adjusting for a suspected (and possibly imperfect) IV likely reduces net bias, is shared by one of the aforementioned simulation studies [77].

Our plasmode simulations reveal somewhat of a circular result: using the exact covariates that affect the outcome in the PS model produces almost no study bias. Unfortunately, it is impossible to know the exact outcome generating process of real-world outcomes of interest. Whether through univariate screens or multivariate regressions, whatever outcome models we utilize to select covariates are inherently parametric and likely fail to capture the "true" generative model. Negative controls are able to provide what simulation experiments cannot – a standard of clinical truth (that of no effect) in real-world data. By using real data as negative control outcomes, our negative control experiments are able to estimate the distribution of residual study bias. Our negative control experiments show a clear result: a PS model based on modeling treatment assignment results in less residual study bias as

measured by negative control distributions (Figure 5.4 and Table 5.3). The All Covariates PS model and associated simulated IV PS models have smaller residual bias, smaller variance, and higher coverage than the respective HDPS Set PS models, which in turn perform better than the Cox Set PS models. Inclusion of simulated IV does seem to generally increase negative control distribution variance across observed PS models.

We offer a word of a caution in selecting covariates based on outcome association through the apparent relative risk of Bross [83]. When we select for the "HDPS Set" the top apparent relative risk covariates out of all covariates, the resultant PS model completely fails to separate the treatment and comparator cohorts. Furthermore, both the plasmode simulation bias and the residual study bias are substantially larger than that of other methods. This is due to low prevalence covariates completely dominating the ranked list of covariates by apparent relative risk. The HDPS algorithm [18] only selects among highly prevalent covariates for the PS model, thus avoiding this phenomenon, and we are sure to select our HDPS Set from among the 500 most prevalent covariates. Our use of regularized regression [19] to fit PS models avoids this issue in the All Covariates PS model despite inclusion of all covariates, as lowly prevalent covariates are shrunk by the lasso penalty [20] to have zero coefficients.

In conclusion, we find that IVs have at most a weak effect on bias in simulations and negative control experiments based on large-scale real-world data, though they do reduce precision. IVs also have very little effect on covariate balance despite strongly affecting PS distributions. We agree with the conviction that PS models should be based on estimating treatment assignment probability as per the definition of the propensity score, and that outcome data not be used in selecting covariates for the PS model [23, 24]. Because real-world data cannot be omnisciently modeled, simulation experiment results may not offer practical comparisons of PS methods. Instead, we prefer to conduct negative control experiments that approximate residual bias, and those results confirm that large PS models curated through regularized regression [19] offer least bias with or without instrumental variables.

CHAPTER 6

Performance Evaluation of Regression Splines for Propensity Score Adjustment in Post-Market Safety Analysis with Multiple Treatments

6.1 Introduction

Observational health data provide a key resource for monitoring post-market drug safety and effectiveness. However, while many medical situations present with multiple (more than two) treatment options, comparative effectiveness research has been largely limited to comparing only two treatments. As a result, conclusions about multiple treatments rely on comparing results from disparate studies, with possibly differing study design elements.

The propensity score (PS), an estimate of treatment assignment probability conditional on observed baseline characteristics, is a widely used tool for confounding control in observational studies [21, 33]. There is substantial debate as to how to best select variables to include in the PS, how to estimate the PS, and how to use the PS to adjust for confounding in the outcome model [90, 34, 40]. Simulation studies assuming correct specification of the PS model have shown that inverse probability treatment weighting (IPTW) provides better mean square error than matching or stratifying on the PS, with reduced precision for PS matching and increased bias for PS stratification [91, 92]. PS adjustment through direct inclusion of the PS in the outcome model is an alternative to weighting, and its application through spline functions has been frequently utilized [90, 34, 93] in two-treatment settings. Recently, an extensive simulation study [42] found that PS splines can provide better performance than other PS methods, including IPTW.

Estimation of causal effects using the PS has straightforward extensions in multiple treatment settings [94], but there is little research on the relative performance of PS methods for multiple treatments [95, 96]. A recent study found IPTW to be substantially more biased than PS matching and matching weights in a three-treatment setting [96]. However, PS matching and matching weights improve bias at the cost of restricting the population of interest. To our knowledge, the relative performance of IPTW in comparison to splines has not been investigated in multiple (more than two) treatment settings. In this paper, we conduct simulation experiments to compare IPTW and spline methods for estimating the average treatment effects for three treatments. We compare performance between IPTW and splines with simulations under a range of PS distributions, outcome prevalences, constant and heterogenous treatment effect sizes.

6.2 Background

6.2.1 Notation

For the $i = 1, \cdots, n$ individuals in the observed data, let $T_i \in \{0, 1, 2\}$ denote the treatment variable with observed value t_i, and Y_i denote the binary outcome variable with observed value y_i. We observe a vector $\mathbf{x}_i = (x_{i,1} \cdots x_{i,p})$ of p pretreatment baseline covariates. The propensity score (PS) has three components indicating probability of assignment to each treatment: $\mathbf{e}_i = (e_{i,0}, e_{i,1}, e_{i,2})$, where $e_{i,t} = \Pr(T_i = t|\mathbf{x}_i), t = 0, 1, 2$. Let $Y_i(t)$ be the potential outcome if individual i had received treatment t, as defined in the Rubin causal model.

6.2.2 Average Treatment Effect Definition

We are interested in the average treatment effect (ATE) of treatment $t, (t \in \{0, 1, 2\})$ relative to treatment $t', (t' \neq t)$, and the risk ratio (RR) as the measure of effect size. We define our estimands of interest as: $RR_1 = \frac{\sum_{i=1}^{n} Y_i(1)}{\sum_{i=1}^{n} Y_i(0)}$, $RR_2 = \frac{\sum_{i=1}^{n} Y_i(2)}{\sum_{i=1}^{n} Y_i(0)}$, and $RR_3 = \frac{RR_2}{RR_1}$.

6.2.3 IPTW Estimation

Inverse probability treatment weighting (IPTW) uses weights based on the PS to construct a pseudo-population that is balanced on observed covariates [90, 97]. For ATE estimation, the weight for patient i is the inverse of the propensity score of his/her received treatment: $w_i = 1/e_{t_i,i}$. The IPTW weights are used to conduct a weighted logistic regression of the binary outcome on the treatments:

$$\text{logit}(\Pr(Y_i = 1|T_i = t_i)) = \beta_0 + \beta_1 * I(t_i = 1) + \beta_2 * I(t_i = 2) \tag{6.1}$$

where I is an indicator function.

Because all IPTW weights are greater than one, the weighted outcome model has a larger effective sample size than the original population, which can lead to variance underestimation [98]. We therefore use stabilized IPTW weights that adjust each weight by the marginal probability of the received treatment: $w_i = p_{t_i}/e_{i,t_i}$, where $p_t = \frac{1}{n}\sum_{i=1}^{n} I(T_i = t)$ for $t = 0, 1, 2$ [98].

We use the maximum likelihood estimates of the outcome model parameters $\hat{\beta}$ to calculate the marginal relative risk estimators:

$$RR_k = \frac{\sum_{i=1}^{n} \Pr(Y_i = 1|\hat{\beta}, T_i = k)}{\sum_{i=1}^{n} \Pr(Y_i = 1|\hat{\beta}, T_i = 0)}, k = 1, 2. \tag{6.2}$$

We use a sandwich variance estimator for the covariance matrix of $\hat{\beta}$ [99]. We then estimate the marginal relative risk variance through the Delta method [100]. See Section 2 of Supplementary Material for details.

6.2.4 Estimation of Regression Splines

Direct regression methods utilize the PS directly in the outcome model as a predictor. Because the three PS components are linearly constrained to sum to 1, when we use the PS directly in the outcome model regression we drop the first component $e_{i,0}$. The simplest

regression approach would be to include the separate PS components in the outcome model as linear predictors, but this requires the restrictive assumption of linear effects. Instead, we allow a nonlinear effect by modeling a spline function on the logit of the propensity score, that has θ as model parameters.

$$\text{logit}(\Pr(Y_i = 1|T_i, \mathbf{e}_i)) = \beta_0 + \beta_1 * I(t_i = 1) + \beta_2 * I(t_i = 2) + s(\mathbf{e}^*; \theta) \tag{6.3}$$

where $\mathbf{e}_i^* = (e_{i,1}^*, e_{i,2}^*) = (\text{logit}(e_{i,1}), \text{logit}(e_{i,2}))$ denotes the logit of the PS, and s denotes the spline function.

In our experiments, we use natural cubic splines and thin plate regression splines of the PS. Cubic splines fit data to piecewise cubic polynomials in between "knots" and require continuous first and second derivatives at the knots [101]. Natural cubic splines (a.k.a. restricted cubic splines) additionally impose linearity outside the boundary knots, and are represented by B-spline basis functions [101]. These cubic splines are functions of a single variable, so for a PS with two components, we assume an additive effect from two separate natural cubic splines:

$$s(\mathbf{e}_i^*; \theta) = \sum_{j=1}^{m} B_{1,j}(e_{i,1}^*) + \sum_{j=1}^{m} B_{2,j}(e_{i,2}^*) \tag{6.4}$$

where $B_{t,j}$ represents the j^{th} B-spline function for PS of treatment $e_{i,t}^*$, and m is the number of total basis functions determined by the number of knots. Commonly, the number of knots is predefined, knots are placed at equally spaced percentiles of the data, and five or fewer knots are generally sufficient [102].

Unlike natural cubic splines, thin plate splines utilize a basis function that can extend to multiple variables. A "thin plate regression spline" (TPRS) uses thin plate splines with a knot placed at every data point, with additional penalization to avoid overfitting [103] and fit to a desired low rank approximation. We use TPRS in two ways: an "additive" approach with a separate TPRS for each PS component, and a "joint" approach with a TPRS on the 2-component PS.

$$\begin{aligned} \text{Additive: } s(\mathbf{e}_i^*;\theta) &= \sum_{j=1}^{m} U_{1,j}(e_{i,1}^*) + \sum_{j=1}^{m} U_{2,j}(e_{i,2}^*) \\ \text{Joint: } s(\mathbf{e}_i^*;\theta) &= \sum_{j=1}^{m} V_j(e_{i,1}^*, e_{i,2}^*) \end{aligned} \tag{6.5}$$

where $U_{t,j}$ represents the j$^{\text{th}}$ thin plate function for $e_{i,t}^*$, and V_j represents the j$^{\text{th}}$ thin plate basis function for the 2-component PS $\mathbf{e}_i^*$, and m is the total number of basis functions determined by the desired low rank approximation.

For all spline models, we use the maximum likelihood estimates of the spline outcome model parameters $\hat{\beta}$ to calculate the marginal relative risk estimators. We then use a model-based variance estimator for the covariance matrix of $\hat{\beta}$ and estimate the marginal relative risk variance through the Delta method.

6.3 Simulation Experiments

6.3.1 Data Generation

Following the data generation process in two existing studies [96, 63], for each individual i we simulate ten covariates $\mathbf{x}_i = (x_{i,1}, \cdots, x_{i,10})$ (3 binary, 1 categorical, 6 continuous) according to prespecified distributions. We simulate treatment T_i from a multinomial logistic distribution with probabilities $(e_{i,1}, e_{i,2}, e_{i,3})$. We simulate binary outcome Y_i from a logistic distribution with probability $\Pr(Y_i = 1|T_i, \mathbf{x}_i)$. See Sections 3.1-3.3 of Supplementary Material for precise details.

We simulate 24 "scenarios" that represent all combinations of different simulation parameters for treatment prevalence (equal, 33:33:33; unequal, 10:45:45), PS overlap (good, fair, poor), outcome prevalence (rare, approximately 2%; common, approximately 10%), and true treatment effect (null, non-null with $RR_1 = 0.8$ and $RR_2 = 0.6$). We simulate 1000 times with a sample size of $n = 5000$. Figure 6.1 shows the PS distribution for simulations under unequal treatment prevalence and fair PS overlap, and the other 5 treatment generating distributions are provided in Sections 3.2.1-3.2.6 of Supplementary Material.

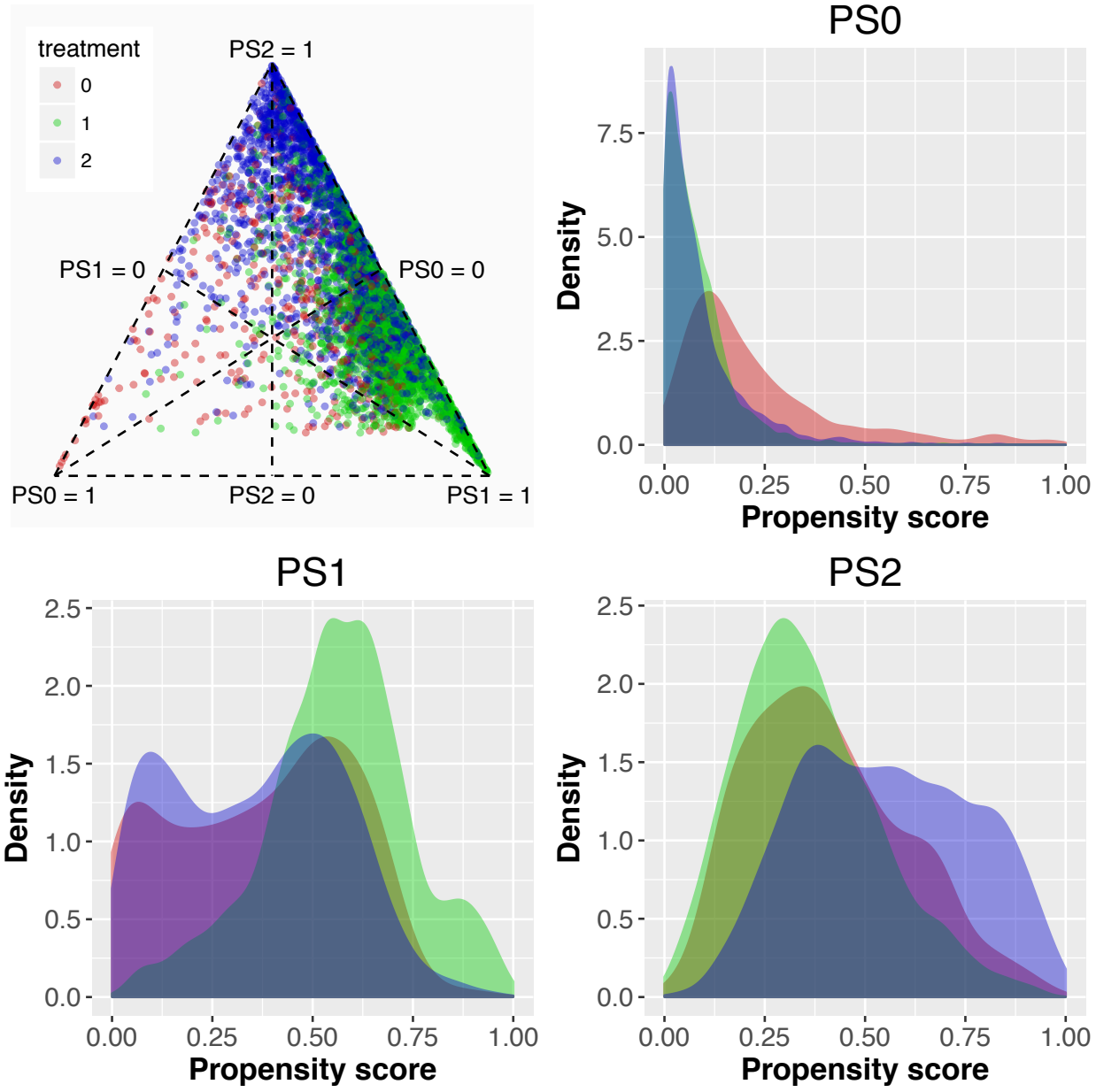


Figure 6.1: Propensity score distribution for simulations with unequal (10:45:45) treatment prevalence and fair PS overlap, drawn for 5000 sample size. PS0, PS1, PS2 represent three components of PS for treatments 0, 1, 2

Across simulations, we vary the PS distribution from good to poor by changing the coefficients of the covariates in the treatment generating model. We also simulate outcomes under different prevalences by changing the intercept term in the outcome generating model. We vary the true marginal relative risk by altering the coefficients of the treatment variables in the outcome generating model. We simulate outcomes with and without treatment effect

heterogeneity, which we model as an interaction between treatment and one covariate (X_4) in the outcome generating model.

6.3.2 Model Fitting

We estimate the PS using a multinomial logistic regression of treatment on covariates. We then use the ten adjustment methods in Table 6.1 to obtain estimates of the marginal relative risk. Truncated IPTW was used to reduce the effects of extreme weights that may inflate variance [104, 105]. Two of the adjustment methods allow treatment effects to vary with propensity score splines through interaction terms, which may be helpful when the underlying data exhibits treatment effect heterogeneity.

We fit each of the 10 adjustments methods under four "settings":

1. Heterogeneity: treatment-covariate X_4 interaction added in outcome generating model 2. Trimming: exclusion of subjects based on PS percentiles [106], and applied to study population before PS adjustment 3. PS misspecification: intentional removal of covariate X_9 from PS estimation process 4. Standard: no heterogeneity, trimming, or PS misspecification

Adjustment method	Description
Outcome Model	Direct outcome regression on covariates, no PS adjustment
IPTW	Stabilized IPTW
IPTW Fixed Trunc.	Stabilized IPTW truncated to [0.10, 10]
IPTW % Trunc.	Stabilized IPTW truncated to 99^{th} percentile by treatment
Cubic 1	Additive natural cubic splines with 1 interior knot at median
Cubic 4	Additive natural cubic splines with 4 interior knots at quantiles
TPRS 1D	Additive thin plate regression splines
TPRS 2D	Joint thin plate regression spline
Cubic + interaction	Cubic 4 + interaction between treatment and splines
TPRS + interaction	TPRS 2D + interaction between treatment and spline

Table 6.1: Compared PS adjustment methods

See Section 4 of Supplementary Material for greater detail.

6.3.3 Performance Evaluation

We assess the methods' performance based on their bias, variance, root mean square error (rmse), and coverage of the true marginal relative risk across simulations. As a supplemental consideration, we also evaluate IPTW weight distributions and covariate balance after weighting, which are pre-outcome diagnostic metrics for assessing validity of IPTW estimates. Because splines utilize the PS directly in the outcome model, no similar diagnostics exist prior to looking at outcome data. The diagnostics for the scenario with unequal treatment prevalence and fair PS overlap are shown in Figure 6.2 and those for all treatment generating distributions are provided in Section 5 of Supplementary Material.

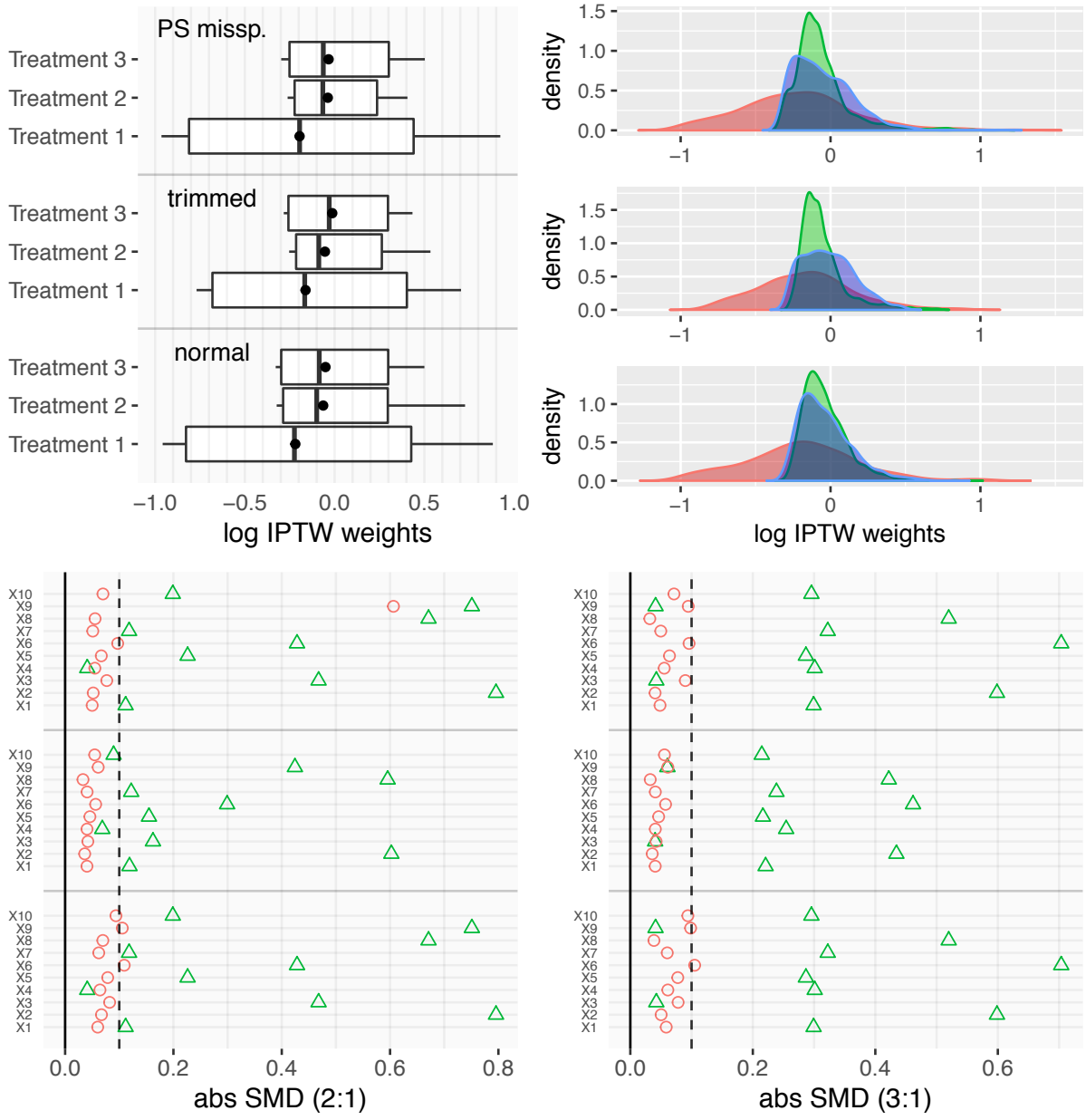


Figure 6.2: IPTW diagnostics for simulations under unequal treatment prevalence and fair covariate balance, with normal/heterogeneous analyses in bottom row, trimmed analysis in middle row, misspecified PS analysis in top row. Top left: Percentiles of IPTW weights. Tips are 1^{st} and 99^{th} percentile, box spans 5^{th} to 95^{th} percentile, middle line is median, dot is mean. Top right: density of IPTW weights. Bottom: Before (green triangle) and after (red circle) IPTW weighting absolute standardized mean differences between treatment groups 2 to 1 (left) and treatment groups 3 to 1 (right).

6.4 Results

Full results for all simulations are provided in Section 6 of Supplementary Material. Across all simulations, results (bias, variance, rmse, and coverage) for null and non-null treatment effects are extremely similar, when other simulation parameters are fixed. Figure 6.3 shows the rmse in estimating the marginal relative risk for all six combinations of treatment prevalence and PS overlap, under 10% outcome prevalence, null true treatment effect size, and standard setting (no treatment effect heterogeneity, trimming, or PS misspecification). Under good PS overlap, there are few differences among PS methods, and all demonstrate small rmse. Under fair and poor PS overlap, direct outcome regression and the four spline methods without interaction have the smallest rmse and perform similarly to each other. Under this setting, all IPTW methods have high rmse, though truncation reduces the rmse. The spline models with interaction perform worse than the spline methods without interaction, which could be due to overfitting in the absence of treatment effect heterogeneity.

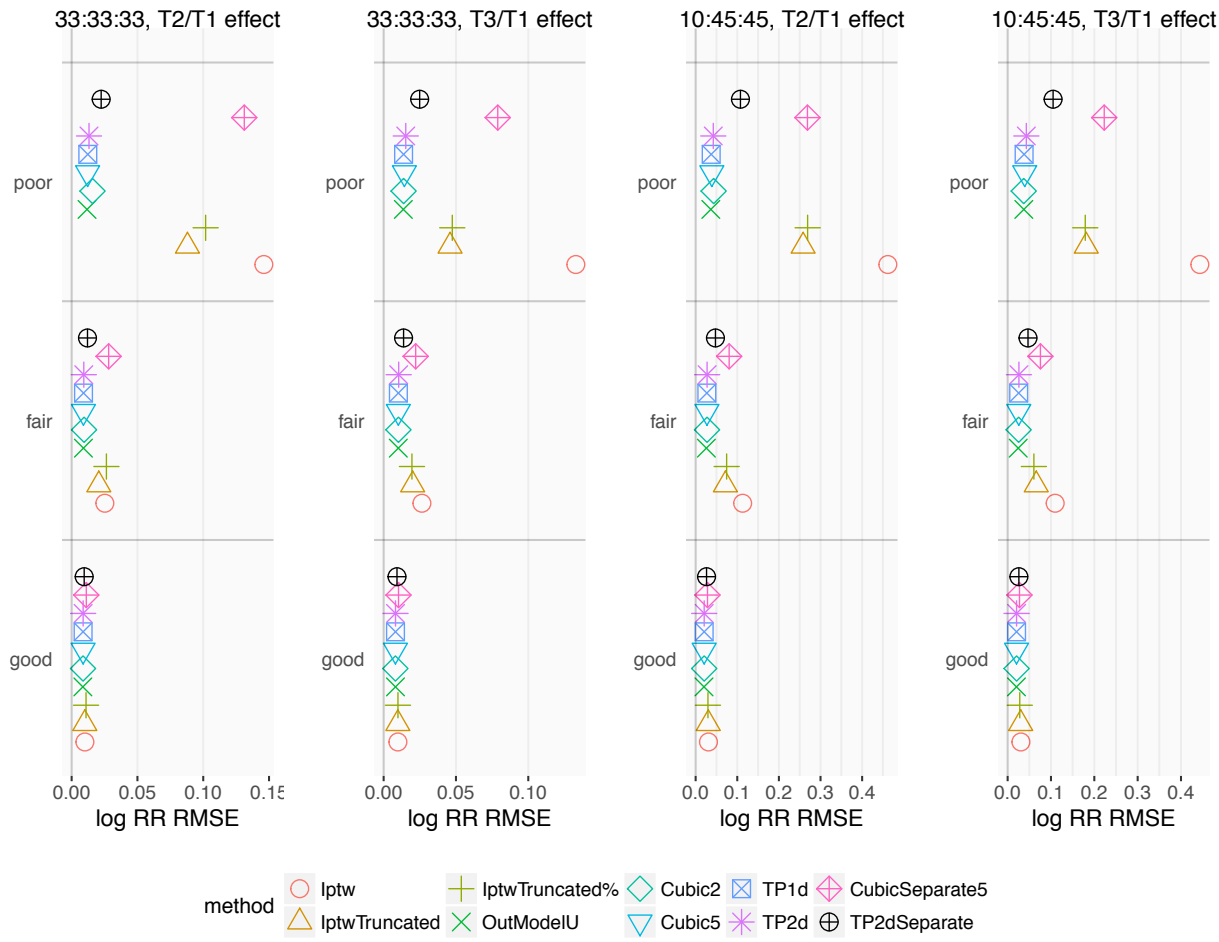


Figure 6.3: RMSE in scenarios with 10% outcome prevalence, null effect sizes, by degree of PS overlap (good, fair, poor)

Figure 6.4 shows the bias, variance, rmse, and coverage of the true marginal relative risk for the simulations under unequal treatment prevalence, fair PS overlap, 10% outcome prevalence, null true effect size, for each of the four settings (heterogeneity, trimming, PS misspecification, and standard). In all four settings, direct outcome regression and the four spline methods without interaction produce similar results and have the smallest rmse. Although additive natural cubic splines have a slightly lower coverage with 1 interior knot at the median compared to 4 interior knots at quantiles in this and several other scenarios, the splines' performance with 1 knots and 4 knots are generally comparable (see full results in Section 6 of Supplementary Material). In all but the misspecified PS setting that has large bias, the rmse is dominated by the variance over the bias. Compared to IPTW, truncated

IPTW adds bias when PS overlap is good to fair but reduces bias when there is substantial lack of overlap (poor overlap and unequal treatment prevalence scenarios). Additionally, across all scenarios, IPTW truncation substantially reduces variance for smaller rmse. PS trimming reduces the bias of IPTW methods, suggesting that extreme weights pose challenges to these methods. Trimming also improves coverage and greatly reduces variance of the IPTW estimate, leading to similar rmse as the truncated IPTW estimates. However, with trimming, the variance of direct outcome regression and spline methods without interaction increase slightly due to the smaller sample size, but these methods still have the smallest rmse. Interestingly, PS trimming generally reduces the bias of spline models with interaction.

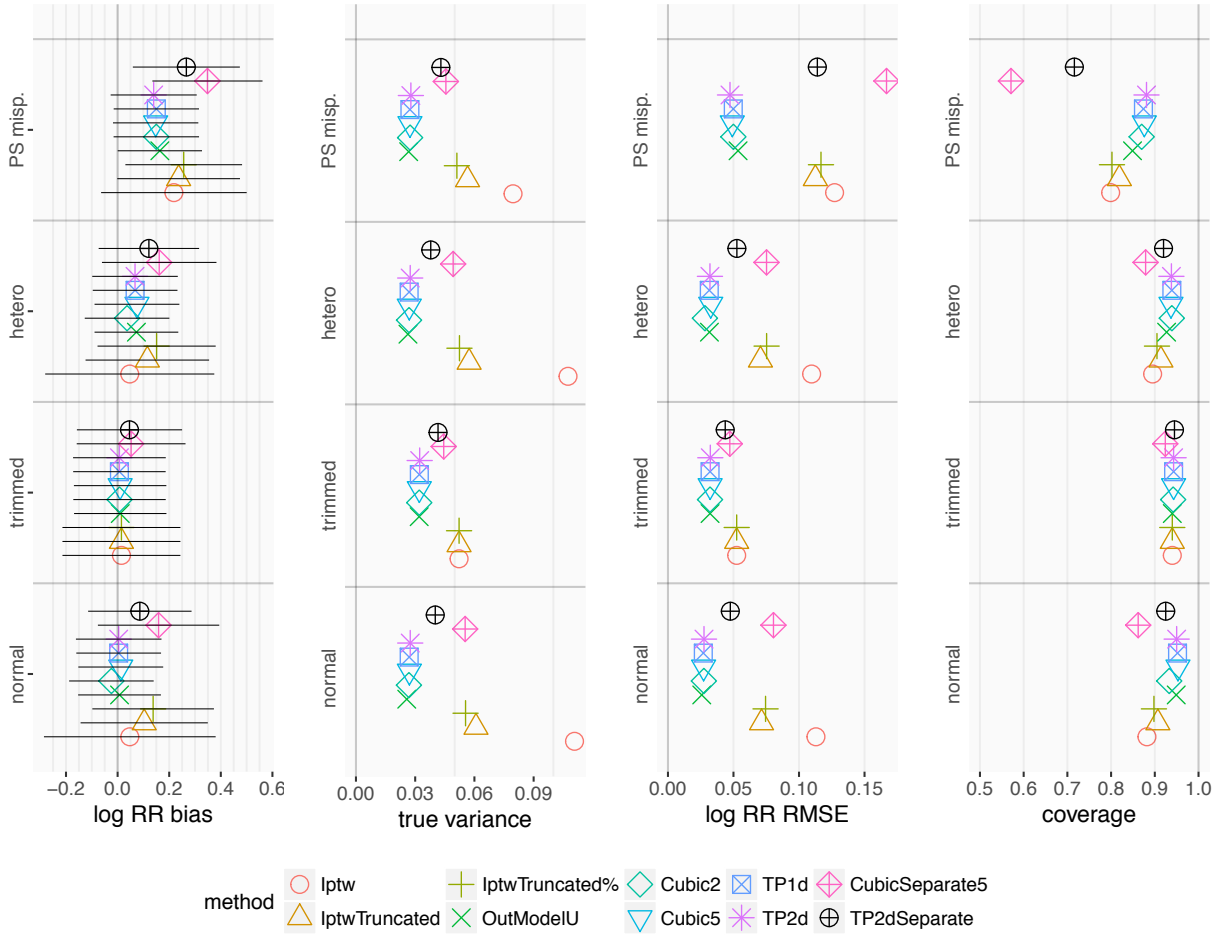


Figure 6.4: Results for 10% outcome prevalence, unequal treatment prevalence, fair PS overlap, null effect sizes, T_1/T_0 effect

Treatment effect heterogeneity in the true outcome generating model causes direct outcome regression and spline methods without interactions to become misspecified. This leads to higher bias for these methods in the heterogeneity setting compared to the standard setting, unlike IPTW where the bias is similar in both settings. However, the methods' variance are generally similar in the heterogeneity setting and the standard setting and dominate over the bias. Interestingly, the splines with interaction that attempt to model treatment effect heterogeneity continue to have higher bias and higher variance than the spline methods without interaction. PS misspecification increases the bias for all methods to levels that meaningfully affect the rmse, and there is a noticeable decline in coverage. However, the relative performance of methods does not change from the standard setting.

Figure 6.5 shows the effect of simulating under different outcome prevalences (rare, approximately 2%; common, approximately 10%) and standard setting, for otherwise the same simulation parameters as in Figure 6.4. Direct outcome regression and four spline methods without interaction maintain their low bias and good coverage in both outcome prevalence settings, while IPTW methods have a large increase in bias and a large notable decline in coverage in the rare prevalence setting. Interestingly, cubic spline with interaction has smaller bias under the rare prevalence setting. Overall, all methods have increased variance and rmse in the rare outcome prevalence simulations, but their relative performance does not change from the common outcome prevalence setting.

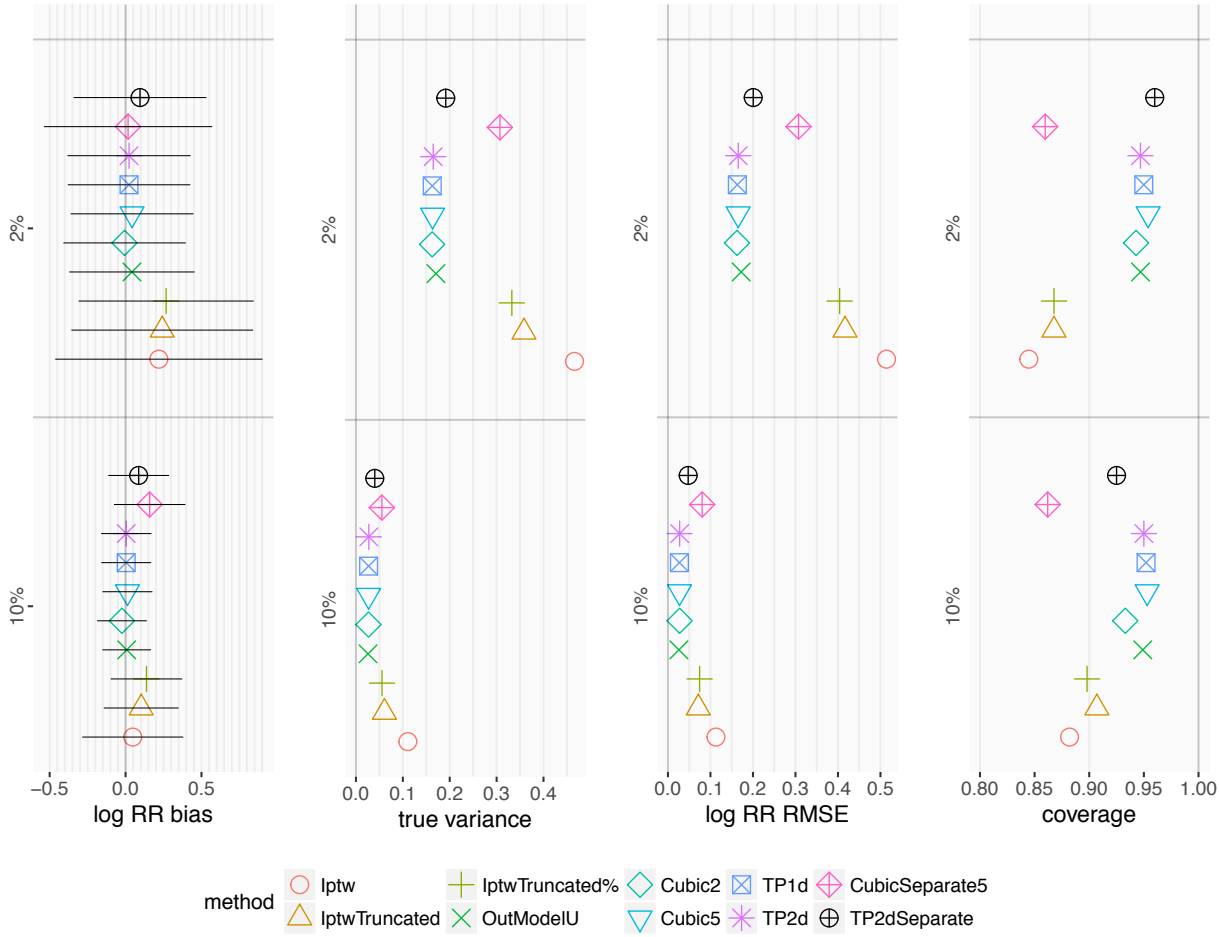


Figure 6.5: Results comparing common and rare outcome prevalence in simulations with unequal treatment prevalence, fair PS overlap, null effect sizes, normal analysis, T_1/T_0 effect

6.5 Discussion

Across a range of simulations, we find that PS adjustment using splines generally provide smaller rmse, bias, and variance than IPTW in a three-treatment setting. Direct outcome regression on covariates also provide comparably favorable performance as splines. However, our simulations are based on a model with only 10 covariates, and direct outcome modeling is known to struggle in realistic studies that have more covariates and relatively few outcomes.

When there is treatment effect heterogeneity, we observe that PS splines without treatment interaction provide biased estimates. We suspect that this is because they misspecify

the outcome model by assuming a constant treatment effect. We indeed see that bias increases for the spline methods, while not changing much for IPTW. However, using splines with interactions between treatment and PS to model the treatment effect heterogeneity did not improve bias or rmse. These heterogeneity results stand in contrast to the simulations in [42] that found that fitting separate spline models by treatment group provided less bias compared to a single spline model in simulations with treatment effect heterogeneity. In that study, however, the treatment effect is directly proportional to the PS, resulting in much more pronounced heterogeneity than in our simulations. Similarly, the authors of [107] find that their approach of separate spline models with multiple imputation provides superior performance under simulations with treatment effect heterogeneity modeled through generating outcomes from different distributions for the different treatment groups. Such an outcome generating model also leads to rather extreme treatment effect heterogeneity. Perhaps under a more extreme model of treatment effect heterogeneity, our simulations may also find that the spline methods become biased, and that the splines models with interaction are indeed preferable.

Parsimony and interpretability are important model considerations in clinical and regulatory settings, where methods and results need to be understood by a diverse group of stakeholders including patients and physicians. In using PS splines, there are several implementation decisions: what spline function to use, how many knots to employ, whether and how to incorporate smoothing. Because the PS is multidimensional in the multiple treatment setting, separate spline functions on the individual PS components may not provide as much model complexity as a multidimensional spline function. However, in our simulations, the "joint" thin plate regression spline on the multidimensional PS provide similar results as the separate "additive" thin plate splines and the cubic spline methods. This favors the easy-to-understand cubic spline, "piecewise cubic polynomials," over the more advanced thin plate regression spline, which we would be harder pressed to explain to a clinical researcher as a "low rank approximation to the full smoothing spline using thin plate basis functions." We do find that in some simulations, the cubic splines with only a single interior knot provided noticeably different estimates and lower coverage than the cubic splines with 4 interior knots.

When analyzing real data, we suggest using data driven methods to select the number of knots. Other studies have argued that a relatively small number of knots is sufficient for fitting cubic splines [102].

Our results for multiple treatments extend findings established in studies with two treatments. For example, other studies [108, 96] find that IPTW performs poorly with low PS overlap (a.k.a. positivity) and low outcome prevalence. Our generally favorable spline results are also consistent with other studies [109, 42]. However, real-world multiple treatments studies are more likely to deviate from ideal conditions than two-treatment studies. The multidimensional PS space for three treatments is more difficult to evenly populate than that of a linear PS for two treatments; higher dimension PS spaces with more than three treatments would be even more challenging. As a result, multiple treatments studies may inherently demonstrate poorer positivity with increasing number of treatments. Additionally, trimming based on the PS is a more difficult task in multiple dimensions, as it is more likely to eliminate an unacceptably large proportion of patients. For example, trimming based on the fixed boundaries 0.10-0.90 would leave 80% of the PS space in a two treatment setting, while the analagous boundaries in a three treatment setting would leave the square of that proportion, 64%. Our trimming approach based on percentiles in each treatment group, as is popularly done in the two-treatment setting [106], would also trim away progressively more patients with more treatments.

Regulators often rely on companies to provide evidence of product safety and must do their due diligence to prevent investigator biases such as selecting analysis parameters that happen to produce desired results. For this reason, PS methods such as matching and weighting are attractive for separating PS adjustment from treatment effect estimation, allowing for pre-outcome diagnostics to assess whether the method is likely to provide a valid estimate [99]. Our IPTW diagnostics (Figure 6.2 and Section 5 of Supplementary Material) show a much wider weight distribution for a treatment group with low prevalence, and an inability to control weighted covariate standardized mean differences to below 0.10 in the poor PS overlap simulations. These conditions both presage poor IPTW performance. Spline methods have the disadvantage that they have no similar pre-outcome diagnostics.

However, in our simulations they increasingly outperform IPTW as the IPTW diagnostics become less favorable. In real-world studies, we suggest conducting sensitivity analyses to verify that conclusions from IPTW analyses still hold with the splines.

While we have mainly focused on splines and IPTW as PS adjustment methods, other methods may also be used in multiple treatment settings. Matching for multiple treatments is possible but becomes computationally challenging with more than three treatments [110]. Instead, the method of matching weights [111] has shown favorable performance compared to IPTW and matching for three treatments [96]. However, matching weights creates a matched pseudo-population and does not estimate the ATE in the entire sample; in our study, we prioritized comparing methods that estimate the ATE. Splines are also far from the only outcome regression PS adjustment method. One could include the PS as a continuous or categorical variable, for example. However, we don't expect these approaches that assume linear effects to provide superior results to the flexible modeling of splines.

CHAPTER 7

GPU Parallelization of Cyclic Coordinate Descent for Large Scale Cross Validated Logistic Regression

7.1 Introduction

Contemporary longitudinal observational health databases contain time-stamped patient-level data on up to hundreds of millions of patients, and offer staggering amounts of data for clinical research [30]. These data are particularly useful for post-market safety surveillance of drug adverse events [112]. Observational health data can provide thousands of unique patient characteristics – demographics, drugs, conditions, procedures – as potentially confounding covariates in statistical analyses. Generalized linear models with unknown parameter regularization offer a rich tool for estimating the association between drugs and outcomes of interest while accounting for these many covariates [113, 29]. Statistical regularization methods such as the lasso [20] provide Bayesian priors on the covariate model parameters while providing shrinkage on the number of model covariates [114]. However, the large scales of the resultant regression models can pose a taxing computational burden on typical computing resources available to the clinical researcher.

One solution to challenging computations is to distribute work across multiple CPU cores or a CPU cluster. This approach is well suited to fitting GLMs while simultaneously searching for optimal regularization parameters because cross-validation – a popular method for the regularization parameter search – involves fitting a number of independent, separate models in an "embarrassingly parallel" fashion [115]. However, the small number of cores on a typical desktop computer limit the maximum speed up of GLM fitting, while CPU clusters can be expensive or inaccessible for many interested clinical researchers. Further-

more, CPUs have begun to hit a hardware limit on their clock speeds and thus their future improvement in computing power [116]. In contrast, graphics processing units (GPUs) are relatively inexpensive, easy-to-use hardware that offers impressive potential for speeding up computations [117]. A GPU can be connected directly to a personal computer and, with compatible software, require no additional expertise to use.

GPUs for general computing have seen considerable use in the field of computer science and machine learning, such as for Support Vector Machine computation [116, 117, 118, 119, 120]. However, their use in statistical computing has been more limited. Suchard et al. [121] provides an introduction to using GPUs for statistical model fitting. The impressive benefits of GPUs achieving one to two orders of magnitude improvement over CPUs is demonstrated for a Bayesian self-controlled case series model in Suchard et al. [28]. A similar magnitude of improvement is seen in Zhou et al. [122] that applies GPUs to high-dimensional optimization. In this paper, we extend the work of Genkin et al. [123] by developing a GPU implementation for fitting logistic regression, a widely used model for binary classification, through cyclic coordinate descent (CCD). Instead of fitting cross-validation folds independently in an embarrassingly parallel fashion, we exploit fine-grain parallelism and fit them synchronously to achieve maximal GPU efficiency. We implement our GPU program in the R package Cyclops [28], and provide numerical results comparing GPU to single-threaded and multi-threaded CPU.

7.2 Methods: Background

7.2.1 GPU Architecture

GPUs are many-core architectures that consist of a number of multiprocessors that each contain numerous cores. A modern GPU can consist of thousands of total cores, allowing for massive parallelization of computational tasks. In this paper we utilize a NVIDIA Titan V GPU, which is based on the NVIDIA Volta architecture comprising 64 cores per multiprocessor and 80 total multiprocessors, for 5,120 total cores. This GPU can deliver 14.90

TFLOPS of single-precision performance, or 7.45 TFLOPS of double-precision performance.

OpenCL is a computing platform for GPUs that provides compatibility across heterogeneous GPU hardware, and is what we use in this paper [124]. Cano [117] provides an introduction to GPU architecture. GPU cores execute the same GPU program ("kernel") on different elements of large data arrays. A single element, or index, of the overall GPU task is known as a thread. Threads are organized into thread blocks that are mapped onto the many multiprocessors. Each GPU core has 32-bit memory registers to store local variables for a given thread. All threads in the same thread block have access to a small shared memory on the multiprocessor, and accessing shared memory is as fast as accessing registers [125]. All threads have access to a large high-bandwidth global memory on the GPU. GPU kernels are executed across multiple thread blocks, the size of which is specified by the user and the number of which depend on the total amount of threads. These thread blocks are delegated to the multiprocessors, and (on NVIDIA hardware) broken down into 32-thread "warps" that run on individual cores in a pipelined fashion. The GPU schedules warps to maximize efficiency, by swapping out idle warps waiting on data access or function results with ones ready for computation [117].

7.2.2 GPU Programming

GPUs contain thousands of processor cores that can apply the same numerical operations simultaneously to elements of large data arrays under a "Single Instruction, Multiple Threads" (SIMT) programming paradigm. While GPUs offer great potential for parallelism through its many processor cores, there are several main limitations to their performance:

- Kernel overhead - the overhead to launch a kernel can be on the order of microseconds [126]. Moreover, launching multiple kernels can become costly if kernels are sequentially dependent on each other, such as with each step of CCD. Increasing the amount of work per kernel and reducing the number of kernels can improve performance.

- Global memory access - accessing global memory is hundreds of times slower than accessing local memory [127], so reads and writes to global memory should be kept at

a minimum. Because data are retrieved in 128 byte segments, 16 sequential double-precision global memory read/writes can be serviced by a single global memory transaction, known as a "coalesced" memory access. Noncontiguous read/writes are, in contrast, "noncoalesced." Data should be organized as much as possible to maximize coalesced memory accesses.

- Memory transfer - transferring data between the CPU (host) and the GPU (device) is extremely slow, and has overhead on the order of tens of microseconds [128]. The amount of data transfer operations should be kept at a minimum. Separate data transfers should be grouped together into a single transfer to reduce overhead, and as much computation should be done on the GPU as possible. While the latency in global memory access can be hidden by switching among active warps, memory transfer latency cannot be hidden.

Optimizing GPU performance around these principles is our main goal in achieving maximum performance in our logistic regression implementation.

7.2.3 Logistic Regression

Logistic regression is a common regression model for problems with a binary dependent variable. In large-scale observational clinical studies, logistic regression is widely used to estimate the propensity score – an estimate of treatment assignment probability conditional on pretreatment patient covariates – for confounding control [33]. The number of covariates can be hundreds or thousands in more expansive propensity score model [38, 19]. Let there be N patients, and the observed treatment be $y_i = 1$ for the treatment of interest and $y_i = 0$ for the reference treatment. We model the treatment assignment process as a Bernoulli distribution where the assignment probability p_i is a logit transform of a linear combination of J pretreatment covariates $\boldsymbol{x}_i$:

$$p_i = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i\boldsymbol{\beta}) + 1} \tag{7.1}$$

where $\boldsymbol{\beta}$ is the vector of J regression coefficients. $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ have been expanded to include an intercept term.

The log-likelihood for maximum likelihood estimation over all patients is:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \boldsymbol{x}_i \boldsymbol{\beta} - \log[1 + \exp(\boldsymbol{x}_i \boldsymbol{\beta})]. \tag{7.2}$$

7.2.4 Statistical Regularization

Observational health databases offer thousands of conditions, procedures, drugs, and other recorded medical characteristics that might be included as study covariates. Statistical regularization is often necessary in these high dimensional settings to avoid model overfitting [113, 114, 123]. We focus on "lasso" regularization that penalizes the likelihood by a penalty $p(\boldsymbol{\beta})$ equal to the L_1 norm of the covariate vector $\boldsymbol{\beta}$ [20]. The lasso penalty is also equivalent to imposing independent Laplace priors on the β_j. The degree of penalization is controlled by a single hyperparameter λ. The target for maximum likelihood estimation becomes the penalized log likelihood $P(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + p(\boldsymbol{\beta})$:

$$p(\boldsymbol{\beta}) = -\lambda \sum_{j=1}^{p} |\beta_j| \quad (L_1 \text{ regularization}). \tag{7.3}$$

The size of λ strongly affects the model fitting process through maximum likelihood estimation [129]. If λ is too small, the model may be overfitted or have no unique solution, whereas if it is too large we have the opposite problem of underfitting and excessively shrinking the coefficients of important covariates. Therefore, the optimum value of λ is often found through cross-validation that divides the data into multiple folds, fits the data with one fold left out at a time, and uses the fitted model to compute the out-of-sample likelihood in the excluded fold to avoid overfitting [114]. Different values of λ are searched, and the one with the highest average out-of-sample likelihood is selected as the optimum value.

Additionally, we also implement ridge regression, which penalizes the L_2 norm of the parameters $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}) = -\lambda \sum_{j=1}^{p} \beta_j^2 \quad (L_2 \text{ regularization}). \tag{7.4}$$

7.2.5 Maximum Likelihood Estimation Using Cyclic Coordinate Descent

Cyclic coordinate descent (CCD) optimizes the penalized log likelihood by cycling through all J covariates and taking one-dimensional Newton steps in each, a process that only involves taking scalar first and second partial derivatives of the log likelihood [17]. This process avoids inversion of a large second derivative Hessian matrix present in the multivariate Newton's method and other multivariate optimization strategies [130]. In addition, we follow the optimization approach in [123] that employs an adaptable trust-region bound on $\Delta\beta_j$, the unbounded one-dimensional Newton step. Algorithm 1 details the steps to implement CCD by alternatingly calculating coordinate updates $\Delta\beta_j$ and updating a vector of linear predictors $\boldsymbol{\theta} = \{\theta_i\}, \theta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ for the N subjects.

initialize initial search vector $\boldsymbol{\beta} = \mathbf{0}$;
while $\boldsymbol{\beta}$ *not yet converged* **do**
  for $j \leftarrow 1$ **to** J **do**
    compute univariate gradient $\frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta})$ and hessian $\frac{\partial^2}{\partial {\beta_j}^2} L(\boldsymbol{\beta})$;
    compute $\Delta\beta_j$ from $\frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta})$, $\frac{\partial^2}{\partial {\beta_j}^2} L(\boldsymbol{\beta})$, and derivatives of $p(\boldsymbol{\beta})$;
    if $\Delta\beta_j \neq 0$ **then**
      $\beta_j \leftarrow \beta_j + \Delta\beta_j$;
      update $\theta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ for subjects with $x_{i,j} \neq 0$;
    end
  end
end

Algorithm 1: Cyclic coordinate descent

7.2.6 Computational Work

In logistic regression, the univariate derivatives of the log likelihood function $L(\boldsymbol{\beta})$ are

$$\begin{aligned}\frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta}) &= \sum_{i=1}^{N} y_i x_{i,j} - \frac{x_{i,j}\exp(\theta_i)}{1+\exp(\theta_i)} \\ \frac{\partial^2}{\partial {\beta_j}^2} L(\boldsymbol{\beta}) &= -\sum_{i=1}^{N} \frac{x_{i,j}^2 \exp(\theta_i)}{(1+\exp(\theta_i))^2}\end{aligned} \tag{7.5}$$

where $\theta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ are values of the linear predictors. We divide the work in Algorithm 1 into the following steps:

A.1 Compute $\sum_{i=1}^{N} y_i x_{i,j}$ for all j (One-time computation)

A.2 Loop over all j:

(a) Compute gradient component $-\sum_{i=1}^{N} \frac{x_{i,j}\exp(\theta_i)}{1+\exp(\theta_i)}$ and hessian term $-\sum_{i=1}^{N} \frac{x_{i,k}^2\exp(\theta_i)}{(1+\exp(\theta_i))^2}$. When the data matrix $\boldsymbol{X}$ is dense, these operations have serial complexity $O(N)$. When $\boldsymbol{X}$ is sparse, these operations have serial complexity $O(X_{max})$ where X_{max} is the number of nonzero $x_{i,j}$

(b) Compute derivatives of penalty term, $\frac{\partial}{\partial \beta_j} p(\boldsymbol{\beta})$ and $\frac{\partial^2}{\partial {\beta_j}^2} p(\boldsymbol{\beta})$. This is done in constant time $O(1)$

(c) Compute $\Delta\beta_j$ in constant time $O(1)$

(d) If $\Delta\beta_j \neq 0$, update $\beta_j \leftarrow \beta_j + \Delta\beta_j$. This is done in constant time $O(1)$

(e) If $\Delta\beta_j \neq 0$, update $\theta_i \leftarrow \theta_i + x_{i,j}\Delta\beta_j$. This operation has the same serial complexity as in Step A.2a

A.3 Calculate and check convergence criterion for $\boldsymbol{\beta}$. Criteria that depend on the linear predictors $\boldsymbol{\theta}$ have serial complexity $O(N)$

The penalty $p(\boldsymbol{\beta})$ for lasso L_1 regression requires directional derivatives [17].

We will refer to the entirety of Algorithm 1 as a "CCD algorithm" that proceeds until $\boldsymbol{\beta}$ converges. Step A.2 is a "CCD cycle" that cycles through each of the J covariates, and each covariate step is a "CCD step."

7.2.7 Data Sparsity and Memory Access

When the data $\boldsymbol{X}$ are dense, Steps A.2a and A.2e are dense operations with serial complexity $O(N)$. When $\boldsymbol{X}$ are sparse, only nonzero subjects $\{i : x_{i,j} \neq 0\}$ and their corresponding covariate values $x_{i,j}$ are stored and need to be accessed or updated. The serial complexity reduces to $O(X_{max})$, where X_{max} is the number of nonzero $x_{i,j}$ that can be substantially smaller than N for very sparse data. However, even with sparse $\boldsymbol{X}$ the linear predictors $\boldsymbol{\theta}$ are still dense, and must be accessed non-contiguously, resulting in noncoalesced and expensive memory access. When covariate prevalences are high, such as in the double digit percentages, the extra memory overhead of sparse data representation and irregular memory access can outweigh the benefits of sparsity. However, in large observational medical studies with hundreds or thousands of covariates, most covariates usually have sufficiently low prevalence to warrant sparse representation.

7.3 Methods: GPU Implementation for Logistic Regression

Logistic regression is relatively resistant to GPU optimization because the amount of local operations is small relative to the amount of expensive global memory transactions. The gradient and hessian calculation of Step A.2a involves a simple transformation of the global vectors $\boldsymbol{X}$ and $\boldsymbol{\theta}$ and reduction over all nonzero indices. Updating $\boldsymbol{\theta}$ in Step A.2e requires writing to the global vector $\boldsymbol{\theta}$. There is relatively little arithmetic in between these two steps to calculate $\Delta\beta_j$. In addition, CCD is an inherently serial algorithm, as each coordinate cannot proceed until the last one is updated. However, despite these limitations we find fruitful areas for GPU optimization, especially for cross-validation.

7.3.1 Updating Gradient and Hessian

Listing 7.1 provides a GPU kernel for Step A.2a that sums together the components of the gradient and hessian for index (covariate) j. The offsets for index j in the index and data vectors are offK and offX, respectively. The kernel is called as TBS thread blocks each

of size TPB (threads per block), and these TBS * TPB threads collectively work over the N (a.k.a. X_{max}) non-zero entries of index j. Within each thread block, the kernel features a transformation of respective nonzero parts of two vectors $\boldsymbol{X}_{-,j}$ and $\boldsymbol{\theta}$ and a subsequent reduction over all TPB threads. This "fused" transformation-reduction has been described in [28]. Finally, the TBS partial gradient and hessian sums are written to locations in global memory, where they will be accessed by the next kernel. We call these locations "buffers" because they are written to and used across different kernels.

7.3.2 Updating Delta

Listing 7.2 provides a GPU kernel that completes summing the gradient and hessian components, and calculates the covariate step size $\Delta\beta_j$ using the gradient, hessian, regularization hyperparameter, previous coefficient β, and the adaptable trust-region bound [123]. The kernel is launched as a single work-group of size TPB. The reduction over the TBS gradient and hessian components is performed across multiple threads, while computing $\Delta\beta_j$ is performed on a single thread. Although the work done in this kernel is minimally parallel, we still opt to perform it on the device instead of the host, to avoid expensive host-device memory transfers.

7.3.3 Updating Linear Predictors

Listing 7.3 provides a GPU kernel that uses $\Delta\beta_j$ to update the nonzero entries of covariate j in the vector of linear predictors $\boldsymbol{\theta}$ (a.k.a. XBeta). Notice the kernel uses the same patterns of data access as in updating the gradient and hessian. Whereas Listing 7.1 read from the nonzero entries of $\boldsymbol{\theta}$, now we write to them.

Together, the three above kernels are run sequentially for each covariate index to constitute one CCD loop. We then run another kernel to compute the convergence criterion for the CCD algorithm. We use the absolute change in $\boldsymbol{x}_i\boldsymbol{\beta}$ as our difference and $\epsilon = 1\text{e}{-7}$ as our threshold. This convergence criterion, and others such as the likelihood, involve independent calculations and a reduction across all subjects, which is a parallel task well suited for the

GPU. We do not detail the GPU kernel to calculate the convergence criterion, except to say that we use the same patterns of data access, transformation and reduction as in Listing 7.1.

7.3.4 CCD Loop ("Single") Kernel

Instead of separate kernels to update the gradient and hessian, delta, and the linear predictors $\boldsymbol{\theta}$, we can implement the entirety of a CCD loop, including iteration over all covariates, into a single GPU kernel. Thus, if there are J covariates, instead of launching $3J$ kernels, we launch only one kernel per data representation type (dense, sparse, indicator, intercept). The advantages of such a combined kernel include saving on kernel overhead, and avoiding having to read from and write to the global buffers BUFFER (in between Listing 7.1 and Listing 7.2) and DELTAVECTOR (in between Listing 7.2 and Listing 7.3). The disadvantage of this approach is that instead of distributing the work of computing the gradient and hessian, and updating $\boldsymbol{\theta}$, across TBS thread blocks, we use only a single thread block. We do this because we require synchronization across all threads, and synchronization commands are available only within a thread block, not across different thread blocks. It may seem counterintuitive to reduce the number of parallel tasks on a GPU that is designed to handle massive parallelization. However, we will show that the benefits of this single kernel method in some cases outweigh its inefficiencies compared to the separate kernels.

7.3.5 Synchronized Cross Validation

To search for the optimum regularization hyperparameter λ, we use k-fold cross-validation, which at each value of λ searched has the following steps:

B.1 Set λ

B.2 Divide data into k folds

B.3 For each fold $0 \leq l < k$:

(a) Leave fold l out of data

(b) Run CCD algorithm to completion on remaining data

(c) Using fitted model $\hat{\boldsymbol{\beta}}^l$, calculate out-of-sample likelihood in fold l

B.4 Average out-of-sample likelihoods across all k folds

B.5 Repeat Steps B.2 - B.4 r times with different data partitions

By repeating k-fold cross-validation r times, we reduce spurious effects from random data partitioning [131]. There are now $k * r = R$ total repetition-fold combinations, which we call "replicates," and R total CCD algorithms to fit.

On the CPU, the R CCD algorithms are fit serially, leading to approximately R times the runtime of a single CCD algorithm at each λ value searched. Additionally, we can run multithreaded CPU to fit the replicates in parallel. We could then utilize the same GPU kernels in Listings 7.1 - 7.3 or the single combined kernel, by serially calling the kernels for each replicate. However, this fails to utilize the full resources of the GPU, that can handle many parallel tasks. We increase the parallel tasks involved by "synchronizing" replicates, so that they all proceed through the steps of the CCD algorithm in lockstep. That is, all R replicates take the first coordinate step together, then the second, and so on. Each replicate uses the same underlying data $\boldsymbol{X}$, but has a different set of binary weights $\boldsymbol{w}^l = \{w_i^l\}$, where l indexes the replicate. In addition, each replicate has its own model parameters $\boldsymbol{\beta}^l$ and linear predictors $\boldsymbol{\theta}^l = \{\theta_i^l\}, \theta_i^l = \boldsymbol{x}_i\boldsymbol{\beta}^l$. This synchronized approach is detailed in Algorithm 2, where the loop over R replicates is performed in parallel.

7.3.6 Memory Representation

As described in Section 7.2.7, for sparse and indicator data we access the nonzero $x_{i,j}$ contiguously but they refer to non-contiguous indices of the dense linear predictors $\boldsymbol{\theta}$. This leads to expensive noncoalesced reads of $\boldsymbol{\theta}$ in calculating the gradient and hessian, and non-coalesced writes to $\boldsymbol{\theta}$ when updating $\boldsymbol{\theta}$. In cross-validation, every CV replicate accesses the same data matrix $\boldsymbol{X}$ but their own $\boldsymbol{\beta}^l, \boldsymbol{\theta}^l, \boldsymbol{w}^l$. These replicate specific vectors can be interleaved as in $(\boldsymbol{\theta}_0^0, \boldsymbol{\theta}_0^1 \cdots, \boldsymbol{\theta}_0^{R-1}, \cdots, \boldsymbol{\theta}_{n-1}^0, \boldsymbol{\theta}_{n-1}^1, \cdots, \boldsymbol{\theta}_{n-1}^{R-1})$. This interleaved memory layout

initialize all initial search vectors $\boldsymbol{\beta}^l = \mathbf{0}$;
while $\boldsymbol{\beta}^l$ *not yet all converged* **do**
  for $j \leftarrow 1$ **to** J **do**
    for $l \leftarrow 1$ **to** R **do**
      compute univariate gradient $\frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta}^l)$ and hessian $\frac{\partial^2}{\partial {\beta_j}^2} L(\boldsymbol{\beta}^l)$;
      compute $\Delta\beta_j^l$ from $\frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta}^l)$, $\frac{\partial^2}{\partial {\beta_j}^2} L(\boldsymbol{\beta}^l)$, and derivatives of $p(\boldsymbol{\beta}^l)$;
      if $\Delta\beta_j^l \neq 0$ **then**
        $\beta_j^l \leftarrow \beta_j^l + \Delta\beta_j^l$;
        update θ_i^l for subjects with $x_{i,j} \neq 0$;
      end
    end
  end
end

Algorithm 2: Cyclic coordinate descent for cross-validation

promotes coalesced reads and writes to $\boldsymbol{\theta}$, as opposed to the noninterleaved layout that puts each replicate's vectors end-to-end: $(\boldsymbol{\theta}_0^0, \boldsymbol{\theta}_1^0, \cdots, \boldsymbol{\theta}_{n-1}^0, \cdots, \boldsymbol{\theta}_0^{R-1}, \boldsymbol{\theta}_1^{R-1}, \cdots, \boldsymbol{\theta}_{n-1}^{R-1})$. We will compare both of these memory representation layouts in the Demonstration section.

7.3.7 2-Dimensional Kernels for Cross Validation

For cross-validation, we extend Listings 7.1 through 7.3 to work on R replicates in Listings 7.4 through 7.6. Our work grid is now two dimensional: one dimension indexing the replicates, the other to index the work that each replicate needs to do. For updating the gradient/hessian and $\boldsymbol{\theta}$, the thread blocks have size (TPB0, TPB1), where TPB0 represents the number of replicates each thread block handles, and TPB0 * TPB1 = TPB, the same size as the thread blocks for non cross-validated kernels. When using the interleaved memory layout, TPB0 = 16, and when using the noninterleaved memory layout, TPB0 = 1. The associated global work size is (R, TBS*TPB1). For computing $\Delta\beta_j^l$, instead of a single thread block of size TPB we now utilize R thread blocks of size TPB, one for each replicate. We additionally extend the single kernel for cross-validation, so that each thread block completes the CCD loop for TPB0 replicates.

7.4 Demonstration

We examine the performance of GPU vs CPU in simulated data of varying sizes and also in a real dataset of anticoagulants patients. As described, we utilize a NVIDIA Titan V GPU. For CPU computations we utilize a Intel(R) Xeon(R) W-2155 CPU that runs at 3.3GHz and has 10 cores.

7.4.1 Non Cross-Validated Experiments

In our simulations, we simulate $\boldsymbol{X}$ with 2% sparsity, and draw $\boldsymbol{\beta}$ from a normal distribution with mean 0 and standard deviation 2. We then set 80% of the $\boldsymbol{\beta}$ values to 0 to simulate sparsity in the coefficient effect sizes. We then calculate the logit of the linear predictors and draw outcomes under a Bernoulli distribution. In the non cross-validated simulations, we fit the data under a fixed lasso penalty with $\lambda = \sqrt{2}$. In the cross-validated simulations, we fit the data under several (typically 3-6) values of λ, using an automated search strategy [123]. We use 10-fold cross-validation with between 1-100 repetitions, resulting in between 10 to 1,000 cross-validation replicates.

Our first comparison is in fitting non cross-validated simulated data under a fixed number of covariates ($p = 1,000$) and a variable sample size n. In Figure 7.1 we compare the CPU (CPU), the GPU with separate kernels for each step (GPU), and the GPU with a single kernel for the CCD loop (GPU single). The GPU work group size is TPB = 512, and the number of work groups is TBS = 16. Although successive sample sizes differ by approximately 3×, the ratios of CPU runtimes increase from 3× between the first two sample sizes to 8× between the last two sample sizes (the number of iterations are roughly equal). This reflects increasingly slower memory access as the total data are increased. In comparison, the GPU single method increases in runtime approximately 3× between successive sample sizes at larger n. Surprisingly, the GPU method, which launches $3J$ kernels per CCD loop, has approximately constant runtime throughout, with a slight increase at the highest sample size. This suggests that the kernel overhead may be dominating the relatively little work performed by each thread block. Across TBS = 16 thread blocks of size TPB = 512 each,

we can handle approximately $8,000$ nonzero indices before having to loop within a thread block, a size achieved only at the largest sample size (sample size * sparsity $= 20,000$). Overall, the GPU single method offers increasing speedup compared to CPU with increasing sample size, and the GPU method begins to shine when the data are substantially large.

We see less decisive GPU gains when fixing the sample size at $n = 100,000$ and increasing the number of covariates, as shown in Figure 7.2. Referring back to Figure 7.1, the sample size $n = 100,000$ places us in a regime where the GPU method is only slightly faster than CPU, and GPU single is approximately $3\times$ faster than CPU. We see in Figure 7.2 that these ratios roughly hold across different covariate counts p, except at the lowest levels $p = 100$ and $p = 300$. Between successive covariate values p, the runtimes increase by more than $3\times$ because of increasing iteration counts. CCD demands serial processing of covariates, so it is not surprising that GPU does not offer dramatically increasing gains as the covariate count is increased. The advantage of the GPU is more in being able to process the entire sample size n in parallel within a CCD step.

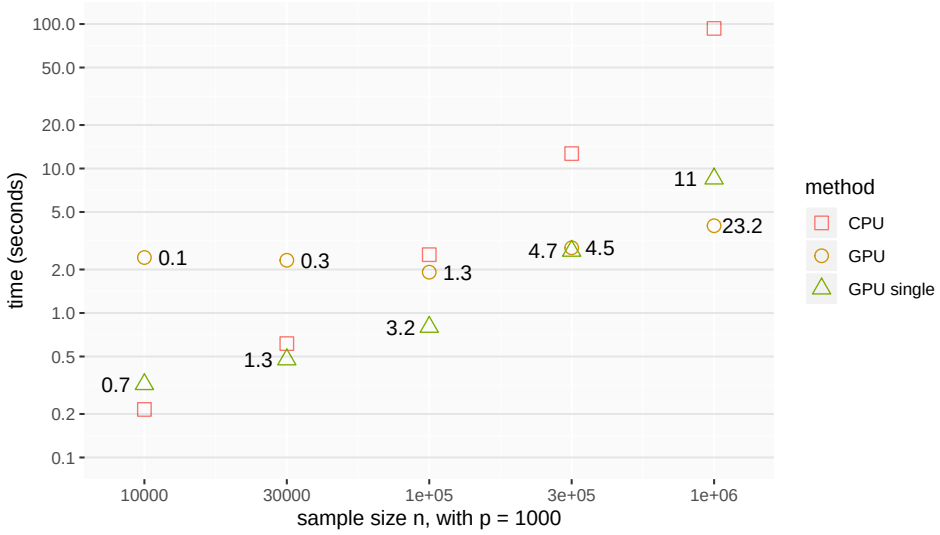


Figure 7.1: GPU vs CPU runtimes for variable n, p = 1,000. For the two GPU methods, GPU to CPU speedup displayed as ratio of CPU runtime to GPU runtime. GPU single refers to using the single combined kernel.

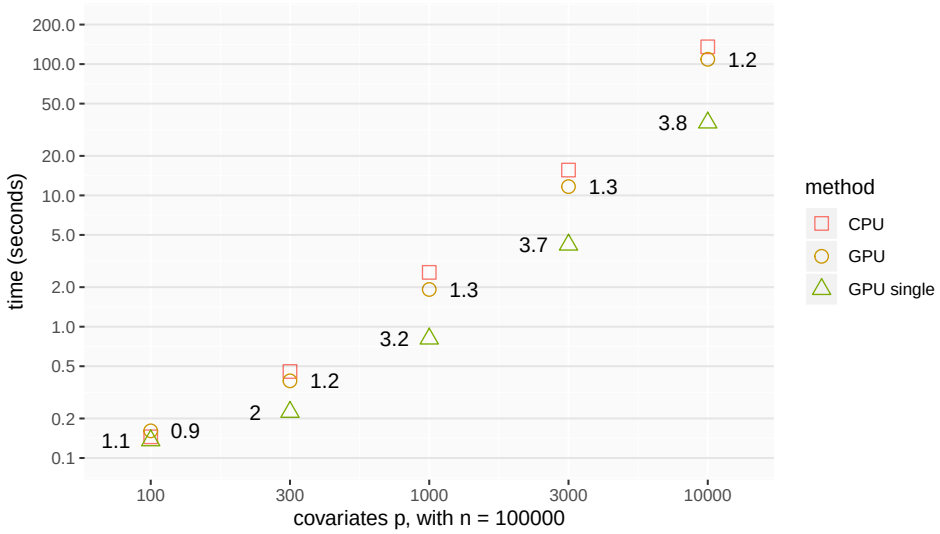


Figure 7.2: GPU vs CPU runtimes for variable p, n = 100,000. GPU to CPU speedup shown as ratio next to CPU points. Vector memory limit reached for CPU dense with p = 10,000

7.4.2 Cross-Validated Experiments

In testing cross-validated logistic regression, we have more GPU methods to compare. We can synchronize replicates or fit them serially (no-sync). Among synchronized approaches, we can either use the interleaved (inter) or non-interleaved (non-inter) memory layout. Additionally, we can utilize separate kernels for the individual CCD step components, or a single kernel (single) to fit the entire CCD loop. We begin by comparing CPU (CPU) and multithreaded CPU with 4 threads (CPU multithreaded) to six GPU methods for simulated data of size $n = 100,000$ and $p = 1,000$ with different numbers of cross-validation replicates, shown in Figure 7.3. Although our CPU can support more than 4 threads, we actually find diminishing returns with greater threads; for example, 4 threads performs faster than 10. When not synchronizing the GPU replicates, we achieve the smallest gains compared to CPU, approximately 1.5× speedup for GPU no-sync and 4.6× for GPU no-sync single. Compared to CPU multithreaded, GPU no-sync is only 0.56× as fast at 1,000 replicates, and GPU no-sync single is 1.7× faster. At the larger replicate counts, the two interleaved memory methods shine, with both delivering approximately 100× speedup compared to CPU at $R = 1,000$, and 38× compared to CPU multithreaded. At a more reasonable $R = 100$

replicates, the 4 synchronized GPU methods have close runtimes and deliver between 27.2× and 71.1× speedup compared to CPU and between 10.1× and 37.4× compared to CPU multithreaded. At lower replicate counts of $R = 10$ and $R = 30$, it is not more advantageous to use interleaved memory, and the non-interleaved, single kernel method has the best performance. Using a single kernel is advantageous for the non-interleaved memory layout across simulations, while it is disadvantageous for the interleaved memory layout.

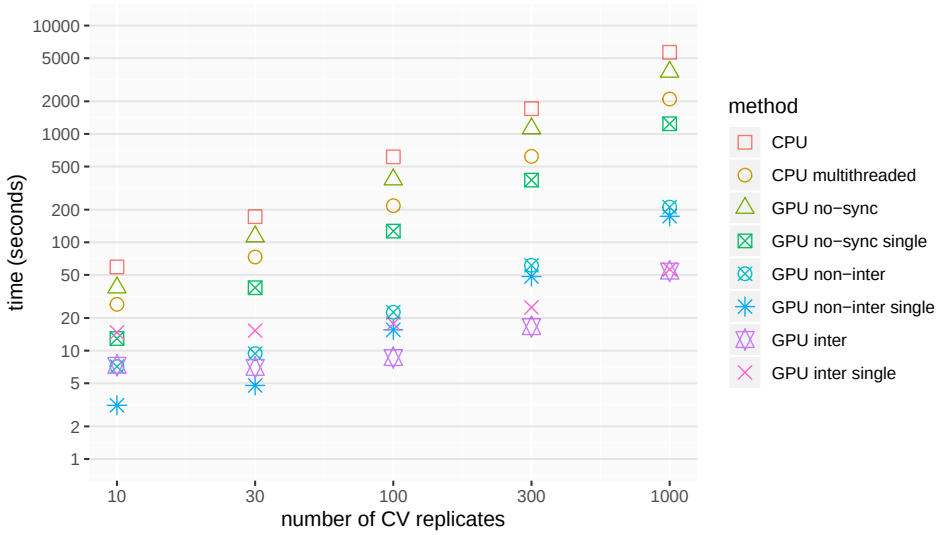


Figure 7.3: Cross validated GPU vs sparse CPU runtimes for n = 100,000, p = 1,000. GPU to CPU speedup shown as ratio

We perform logistic regression on a real dataset by estimating a propensity score of two anticoagulants, dabigatran vs warfarin in patients with non-valvular atrial fibrillation, in the Truven Health MarketScan Medicare Supplemental and Coordination of Benefits Database. Dabigatran is the active treatment, and warfarin is the reference. The study size is 77,122 patients, with 20,474 dabigatran users and 56,648 warfarin users. We use 12,392 total pre-treatment covariates that are indicator variables for demographics, patient conditions, procedures, and drugs.

Because the previous demonstration clearly shows the inefficiency of non-synchronized GPU methods, we omit them from this next comparison. Figure 7.4 shows the results from fitting cross-validation to this real dataset at different replicate counts. We see some similar patterns as before. The non-interleaved, single kernel method is superior at the lower

replicate counts, offering 5.7× speedup at $R = 10$ and 9.8× at $R = 30$ compared to CPU, and respectively 2.0× and 3.8× compared to CPU multithreaded. CPU multithreaded is the second fastest method at the lower replicate count $R = 10$, but its relative performance diminishes with more replicates. At $R = 100$, non-interleaved, single kernel method again has the best performance, with 13.0× speedup compared to CPU and 4.0× compared CPU multithreaded, though the interleaved method is not far behind, at 11.4× and 3.5×, respectively. At the highest replicate count tested, $R = 300$, the interleaved method has the best performance with 24.3× speedup compared to CPU and 8.3× compared to CPU multithreaded, reducing the runtime from 14.4 hours on CPU and 4.9 hours on CPU multithreaded to 35 minutes on GPU. Using a single kernel is again advantageous for the non-interleaved method but disadvantageous for the interleaved method. Both interleaved methods have a fewer than 2× in runtime from the smallest replicate count $R = 10$ to the largest replicate count $R = 300$, suggesting excellent scaling with increased replicates.

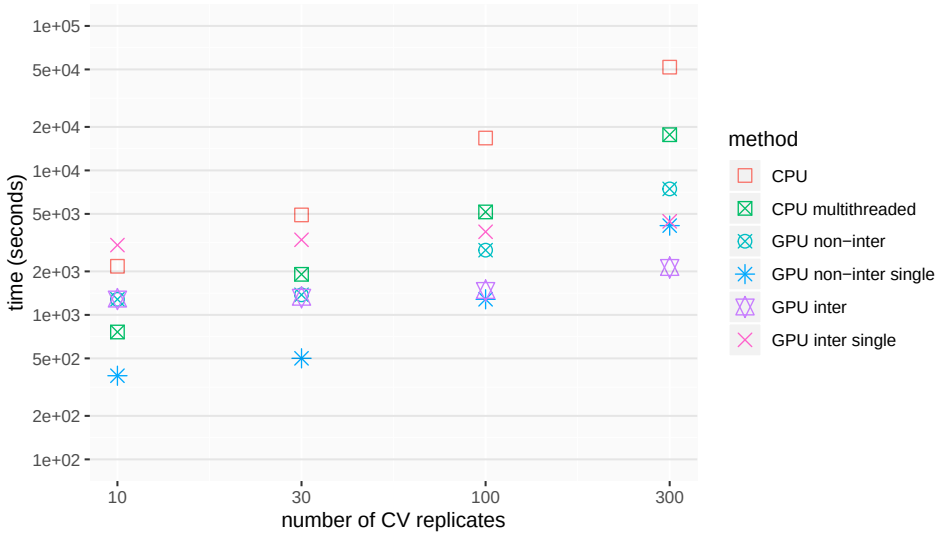


Figure 7.4: Cross validated GPU vs sparse CPU runtimes for anticoagulants dataset, n = 77,122, p = 12,392. GPU to CPU speedup shown as ratio

7.4.3 Data Representation Format

The previous demonstrations all use data in indicator representation for both CPU and GPU. This data representation stores only nonzero indices, because all nonzero data are 1.

In contrast, sparse representation stores both nonzero indices and their values, and dense representation stores values for all indices. We perform a comparison among indicator, sparse, and dense representation using simulated data with $n = 100,000$ and $p = 1,000$. The underlying data are either sparse, with 2% nonzero, or dense, with 20% nonzero. As shown in Figure 7.5, sparse and indicator representations have very similar runtimes for GPU, GPU single, and CPU methods, with indicator slightly faster. Surprisingly, using dense data representation is slowest for both sparse and dense data, and for all three methods. While GPU single (using indicator representation) is the fastest method for sparse data, GPU (using indicator representation) is the fastest method for dense data.

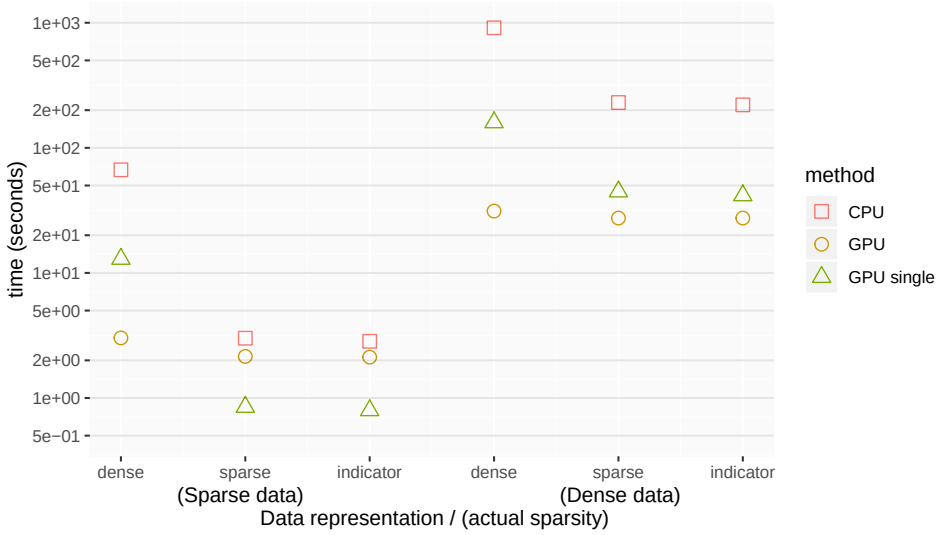


Figure 7.5: GPU vs CPU runtimes for variable n = 100,000, p = 1,000. Sparse data has 2% sparsity, while dense data has 20% sparsity

7.4.4 Comparison to Glmnet

The R package glmnet [132] provides efficient software for fitting lasso and ridge regularized CCD algorithms for logistic regression and other generalized linear models. Glmnet achieves great efficiency in exploiting warm starts between successive λ values in a grid search, that typically contains close to 100 search values. In comparison, our cross-validation search method typically only searches 3-6 values for λ, but in a more adaptive manner than grid search [123]. For a fair comparison in regards to total amount of work performed by 10-fold

cross-validation, we compare the default cv.glmnet call to our implementation with 20 repetitions, or $R = 200$. We use simulated data for which the true effect sizes $\boldsymbol{\beta}$ are known, and compare the prediction accuracy through the c-statistic and the estimation accuracy through the mean absolute difference between the estimated and true effect sizes $\boldsymbol{\beta}$.

We perform tests with $n = 100,000$ and $p = 1,000$. $\boldsymbol{\beta}$ are drawn from a normal distribution with 0 mean and 2 standard deviation. Results for sparse $\boldsymbol{\beta}$, with 80% zeros, are shown in Table 7.1. Our GPU code in Cyclops runs approximately 3x faster than glmnet, and achieves almost 10 times smaller mean absolute difference of the coefficients $\boldsymbol{\beta}$. However, glmnet's lasso identifies significantly more of the 0's in $\boldsymbol{\beta}$. All tested methods achieve high predictive accuracy (c-statistic) of almost 0.95. We additionally compared our GPU code to glmnet on data generated from dense $\boldsymbol{\beta}$, shown in Table 7.2. Both Cyclops and glmnet, with both lasso and ridge regression, achieve very high c-statistic of almost 0.99. Runtimes were similar throughout, with little benefit achieved from the GPU code. However, Cyclops performed significantly better on estimating the underlying coefficients, with approximately 0.100 average absolute error, while glmnet with lasso had 2.162 average absolute error and ridge had 1.696 average absolute error. Considering that the $\boldsymbol{\beta}$ are drawn from a distribution with only 2 standard deviation, glmnet is providing very little accuracy for estimating the model coefficients.

One reason that our GPU implementation does not achieve substantial speedup compared to glmnet is because glmnet uses successive quadratic approximations to the likelihood function that are easier to optimize than using CCD directly on logistic regression [132]. However, these quadratic approximations do not provide descent guarantees, and can converge to different solutions as optimizing logistic regression directly without approximations. We use the built-in glm{stats} function in R as a gold standard to which we compare Cyclops and glmnet without regularization. In Table 7.3 we simulate dense data with nonzero $\boldsymbol{\beta}$ under a variety of n and p. Despite Cyclops and glmnet both running with a threshold of 10^{-7}, Cyclops comes several orders of magnitude closer to the output of glm, which runs iteratively reweighted least squares. Not only does Cyclops come closer to the output of glm, we have seen in Tables 7.1 and 7.2 that Cyclops also comes an order of magnitude closer to the true

$\boldsymbol{\beta}$ values when running regularization.

	GPU - lasso	Glmnet - lasso	GPU - ridge	Glmnet - ridge
time (s)	11.827	31.995	16.903	55.707
% zeros found	22.5%	61%	—	—
average diff	0.0497	0.615	0.0622	0.492
c-statistic	0.948	0.947	0.948	0.947

Table 7.1: Cyclops vs Glmnet for sparse data

	GPU - lasso	Glmnet - lasso	GPU - ridge	Glmnet - ridge
time (s)	51.013	80.222	47.858	55.752
average diff	0.0918	2.162	0.101	1.696
c-statistic	0.989	0.989	0.989	0.988

Table 7.2: Cyclops vs Glmnet for dense data

	Cyclops	Glmnet
Test 1	$8.85 * 10^{-7}$	$1.90 * 10^{-3}$
Test 2	$4.84 * 10^{-6}$	$1.61 * 10^{-2}$
Test 3	$9.07 * 10^{-6}$	$2.97 * 10^{-2}$

Table 7.3: Mean absolute error of coefficients between Cyclops / Glmnet and Glm. Test 1: $n = 1,000, p = 10$, Test 2: $n = 5,000, p = 50$, Test 3: $n = 10,000, p = 100$

7.5 Discussion

Logistic regression has relatively simple gradient and hessian calculations that render it relatively resistant to GPU optimization, because the amount of arithmetic is small compared to the amount of global memory transactions. Nevertheless, we are able to find fruitful areas for GPU optimization. In non cross-validated computations, running the entire CCD loop in a single GPU kernel saves on GPU kernel overhead and global memory operations, and is preferred for smaller data sample sizes n. The GPU is better suited for large sample

sizes n than for large covariate counts p, because we cannot escape the serial nature of CCD. For cross-validation, we are able to simultaneously update all replicates by synchronizing their progression through CCD. We also find impressive performance with interleaved memory at high replicate counts R, reflecting the benefit of coalesced memory access. In the anticoagulants dataset, the runtime of the interleaved GPU method increases by less than $2\times$ between $R = 10$ replicates and $R = 300$ replicates. This enables us to perform many repetitions of k-fold cross-validation to reduce sampling variability [131], a process that produces a prohibitively linear increase in runtime on CPU. In a sense, adding more replicates is almost "free," coming with only a nominal increase in the running time.

In our demonstrations, we have tested replicate counts up to $1,000$, which begs the question: why would we want to repeat 10-fold cross-validation 100 times? Many studies and popular software don't even perform any cross-validation repetitions. In practice, we don't need to perform so many cross-validation repetitions, but our machinery allows us to perform other parallel tasks that are of interest. Our GPU code allows us to fit multiple models synchronously that share underlying data $\boldsymbol{X}$ but have different weights and/or other parameters. With 1,000 "replicates", we could perform 10-fold cross-validation on a grid of 100 λ regularization hyperparameter values. Or, we could perform searches over multiple regularization hyperparameters λ_i for different groups of covariates, or as in an elastic net. Or, we could use 1,000 replicates to fit 1,000 bootstrap samples that each have different weights $\boldsymbol{w}^l$. The GPU programming principles are the same: synchronize the replicates to them simultaneously, and interleave their memory to achieve coalesced access.

Our results favor different GPU methods in regimes of less and more overall work. For non cross-validated problems, using a single kernel is favorable except when the sample size is very large, when it becomes advantageous to distribute work across multiple work groups. For cross-validated problems, interleaved memory only outperforms non-interleaved memory at high replicate counts, suggesting that noncoalesced memory operations is tolerable when the total amount of work is small. When the replicate count increases, and the GPU becomes saturated with memory requests, interleaved memory dominates. Interestingly, using a single kernel improves the performance of interleaved memory but decreased performance

of noninterleaved memory. If we were to recommend a single GPU method to use across different problems, it would be the interleaved, multiple-kernel method. While less favorable at smaller problem sizes and replicate counts, this method has tolerable runtimes. When the problem sizes grow in sample size n or replicate count R, to a regime most in need for GPU optimization, this method shines and delivers the largest absolute decreases in runtime compared to CPU.

We compare our GPU implementation in Cyclops to the popular R package glmnet for running generalized linear models. We find several times speedup for cross-validated regressions on sparse data, and marginal speedup for dense data. While both programs achieve high predictive accuracy as measured by the c-statistic, Cyclops provides much better estimation of the underlying coefficients $\boldsymbol{\beta}$. We demonstrate this better estimation in both regularized and regularized simulations, and show orders of magnitude improvement both in relation to the underlying $\boldsymbol{\beta}$ and the output of the glm{stats} function. We hypothesize that this improvement in estimation comes from our CCD algorithm more accurately approaching the optimum parameter vector $\hat{\boldsymbol{\beta}}$ than glmnet which uses imprecise quadratic approximations for computational speedup. The difference in estimation is significant for regularized regression on dense data, in which the average absolute error of the estimated coefficients from glmnet is approximately the same size as the standard deviation of the underlying coefficients.

The massive parallelization potential of GPUs seems at first glance well suited to MM algorithms [133], which seek to optimize a series of functions tangential to the target likelihood function. MM algorithms are able to achieve parameter separation in CCD, such that updating each coordinate is an independent task and all coordinates can be updated simultaneously on a GPU across different thread blocks. MM algorithms for logistic regression have been proposed [133], and we have explored them to preliminary extents. We find that although parameter separation is indeed massively parallelizable and takes advantage of GPU hardware, the increased number of iterations to convergence for MM algorithms is immense and outweigh the benefits of parameter separation. MM in fact produces quadratic approximations that, unlike that used in glmnet, provide descent guarantees, but the resul-

tant objective function has contours too shallow for fast (or even moderate) convergence. Perhaps generalized linear models other than logistic regression may benefit more from the combination of MM parameter separation and GPU parallelization.

Listing 7.1: updateGradHess

```
__kernel void computeGradHess(
  // For sparse and dense data:
  __global const real* X,        // data values

  // For sparse and indicator data:
  __global const int* K,         // nonzero indices
  const uint offX,           // offset in X
  const uint offK,           // offset in K

  const uint N,          // # nonzero indices
  __global const real* xBeta,          // linear predictors
  __global real* buffer,           // output buffer
  const uint index) {          // covariate index

  // Define shared memory for thread-block reduction
  __local real sGradient[TPB], sHessian[TPB];

  // Partial sums for this thread
  real tSumGradient = 0.0; tSumHessian = 0.0;

  int lid = get_local_id(0);     // Thread id within block
  const uint loopSize = get_global_size(0); // loop size
  int i = lid;         // first element index for thread

  while (i < N) {
    // Dense data access:
    real x = X[offX+i],        // x from dense X
    real xb = XBeta[i];      // contiguous access of XBeta

    // Sparse data access:
    int k = K[offK + i];     // real index of subject
    real x = X[offX + i],    // x from sparse X
    real xb = XBeta[k];      // noncontiguous access of XBeta

    // Indicator data access:
    int k = K[offK + i];     // real index of subject
    real x = 1;      // x is 1
    real xb = XBeta[k];      // noncontiguous access of XBeta

    // Intercept data access:
    real x = 1;      // x is 1
    real xb = XBeta[i]; // contiguous access of XBeta

    real exb = exp(xb) ;
    real numer = x * exb;
    real denom = 1.0 + exb;
    real tGradient = numer / denom;
    real tHessian = numer * x / denom - tGradient * tGradient;
    tSumGradient += tGradient;
    tSumHessian += tHessian;
          i += loopSize;
    }

  sGradient[lid] = tSumGradient;
  sHessian[lid] = tSumHessian;
  // Reduce across all threads in block, leaves total in first element
  parallelReduction(sGradient);
  parallelReduction(sHessian);

  if (lid == 0) {
          buffer[get_group_id(0)] = sGradient[0];
          buffer[get_group_id(0) + get_num_groups(0)] = sHessian[0];
    }
}
```

Listing 7.2: updateDelta

```
__kernel void updateDelta(
  __global const REAL* buffer,       // input grad and hess components
  __global real* deltaVector,      // output step size delta
  const uint TBS,

  __global real* boundVector,      // adaptable trust-region bound
  __global const real* priorParams,    // regularization hyperparameter
  __global const real* XjYVector,    // constant gradient component
  __global real* betaVector,       // current beta values

  const uint index) {

  // Define shared memory for thread-block reduction
  __local real sGradient[TPB], sHessian[TPB];

  int lid = get_local_id(0);     // Thread id within block

  // copy gradient and hessian components into local memory
  while (lid < TBS) {
    sGradient[lid] += buffer[lid];
    sHessian[lid] += buffer[lid+TBS];
  }

  // Reduce across all threads in block, leaves total in first element
  parallelReduction(sGradient);
  parallelReduction(sHessian);

  if (lid == 0) {
    __local REAL grad, hess, beta, delta;
    grad = sGradient[0] - XjYVector[index];    // include constant gradient component
    hess = sHessian[0];
    beta = betaVector[index];
    real hyper = priorParams[index];

    // Compute delta according to regularization type
    delta = computeDelta(grad, hess, beta, hyper);

    // Apply then update adaptable trust-region bound
    real bound = boundVector[index];
    if (delta < -bound) {
      delta = -bound;
    } else if (delta > bound) {
      delta = bound;
    }
    real intermediate = max(fabs(delta)*2, bound/2);
    intermediate = max(intermediate, 0.001);
    boundVector[index] = intermediate;

    deltaVector[index] = delta;
    betaVector[index] = delta + beta;
  }
}
```

Listing 7.3: updateXBeta (sparse)

```
__kernel void updateXBeta(
  __global const REAL* X,        // data values
  __global const int* K,         // nonzero indices
  const uint offX,             // offset in X
  const uint offK,             // offset in K
  const uint N,              // # nonzero indices

       __global const real* deltaVector,
       __global real* xBeta,
  const uint index) {

  uint i = get_global_id(0);
  real delta = deltaVector[index];
  const uint loopSize = get_global_size(0);
  while (i < N) {
    const uint k = K[offK+i];
    const real x = X[offX+i];
    const real inc = delta * x;
    real xb = xBeta[k] + inc;
    xBeta[k] = xb;
    i += loopSize;
  }
}
```

Listing 7.4: updateGradHessCV (sparse)

```
__kernel void computeGradHessCV(
  __global const real* X,        // data values
  __global const int* K,         // nonzero indices
  const uint offX,           // offset in X
  const uint offK,           // offset in K

  const uint N,          // # nonzero indices
  __global const REAL* xBetaVector,
  __global REAL* buffer,
  const uint cvStride,
  const uint kStride,
  const uint syncCVFolds,
  __global int* allZero) {       // for skipping updateXBeta

  if (get_global_id(0) == 0) allZero[0] = 1;  // reset flag

  // Define shared memory for thread-block reduction
  __local real sGradient[TPB], sHessian[TPB];

  // Partial sums for this thread
  real tSumGradient = 0.0; tSumHessian = 0.0;

  uint lid0 = get_local_id(0);     uint lid1 = get_local_id(1);
  uint cvIndex = get_global_id(0);       // cv replicate
  uint gid1 = get_global_id(1);         uint i = gid1;      // index for loop
  uint loopSize = get_global_size(1);   // loop size

  if (cvIndex < syncCVFolds) {
    while (i < N) {
      int k = K[offK + i];
      real x = X[offX + i],

      // interleaved layout
      uint offset = k * cvStride + cvIndex;

      // noninterleaved layout
      uint offset = k + kStride * cvIndex;

      real xb = xBetaVector[offset];
      real exb = exp(xb);
      real numer = x * exb;
      real denom = (real)1.0 + exb;
      real g = numer / denom;
      real w = weightVector[offset];     // CV weight
      real tGradient = w * g;
      real tHessian  = w * g * (x - g);
      tSumGradient += tGradient;
      tSumHessian += tHessian;
      i += loopSize;
    }

    sGradient[mylid] = tSumGradient;
    sHessian[mylid] = tSumHessian;

    // Reduce across all threads in block, leaves total in elements where lid1 == 0
    uint lid = lid1*TPB0 + lid0;     // index for parallel reduction
    parallelReduction(sGradient);
    parallelReduction(sHessian);

    if (lid1 == 0) {
      buffer[cvIndex * get_num_groups(1) + get_group_id(1)] = sGradient[lid];
      buffer[(cvIndex + syncCVFolds) * get_num_groups(1) + get_group_id(1)] = sHessian[lid];
    }
  }
}
```

Listing 7.5: updateDeltaCV

```
__kernel void updateDeltaCV(
  __global const REAL* buffer,       // input grad and hess components
  __global real* deltaVector,      // output step size delta
  const uint TBS,
  const uint stride,
  const uint syncCVFolds,
  const uint cvStride,

  __global real* boundVector,      // adaptable trust-region bound
  __global const real* priorParams,   // regularization hyperparameter
  __global const real* XjYVector,    // constant gradient component
  __global real* betaVector,       // current beta values

  const uint index,
  __global uint* allZero,
  __global const int* doneVector) {      // indicators for completed CV replicates

  // Define shared memory for thread-block reduction
  __local real sGradient[TPB], sHessian[TPB];

  uint cvIndex = get_group_id(0);
  int lid = get_local_id(0);     // Thread id within block

  // copy gradient and hessian components into local memory
  if (lid < TBS) {
    sGradient[lid] = buffer[lid + TBS * cvIndex];
    sHessian[lid] = buffer[lid + TBS * (cvIndex + syncCVFolds)];
  }

  // Reduce across all threads in block, leaves total in first element
  parallelReduction(sGradient);
  parallelReduction(sHessian);

  if (lid == 0) {
    __local uint offset;
    offset = index + J * cvIndex;
    __local REAL grad, hess, beta, delta;
    grad = sGradient[0] - XjYVector[offset];     // include constant gradient component
    hess = sHessian[0];
    beta = betaVector[offset];
    real hyper = priorParams[index];

    // Compute delta according to regularization type
    delta = computeDelta(grad, hess, beta, hyper);

    // Apply then update adaptable trust-region bound
    real bound = boundVector[offset];
    if (delta < -bound) {
      delta = -bound;
    } else if (delta > bound) {
      delta = bound;
    }
    real intermediate = max(fabs(delta)*2, bound/2);
    intermediate = max(intermediate, 0.001);
    boundVector[offset] = intermediate;

    deltaVector[index*cvStride+cvIndex] = delta;
    betaVector[offset] = delta + beta;
    if (delta != 0.0) {
      allZero[0] = 0;          // if any nonzero, disable flag
    }
  }
}
```

Listing 7.6: updateXBetaCV (sparse)

```
__kernel void updateXBetaCV(
  __global const REAL* X,        // data values
  __global const int* K,         // nonzero indices
  const uint offX,             // offset in X
  const uint offK,             // offset in K
  const uint N,           // # nonzero indices

       __global const real* deltaVector,
       __global real* xBetaVector,
  const uint index,
  const uint cvStride,
  const uint kStride,
  const uint syncCVFolds,
  __global const int* allZero) {


  uint lid0 = get_local_id(0);
  uint lid1 = get_local_id(1);
  if (allZero[0] == 0) {
    uint i = get_global_id(1);
    uint cvIndex = get_global_id(0);
    uint loopSize = get_global_size(1);
    if (cvIndex < syncCVFolds) {
      real delta = deltaVector[index * cvStride + cvIndex];
      if (delta != 0) {
        while (i < N) {
          int k = K[offK + i];
          real x = X[offX + i],
          // interleaved layout
          uint offset = k * cvStride + cvIndex;

          // noninterleaved layout
          uint offset = k + kStride * cvIndex;

          real inc = delta * x;
          real xb = xBetaVector[vecOffset] + inc;
          xBetaVector[vecOffset] = xb;
          i += loopSize;
        }
      }
    }
  }
}
```

CHAPTER 8

Comparative Safety and Effectiveness of Alendronate Versus Raloxifene in Women with Osteoporosis

8.1 Introduction

Osteoporosis is a chronic, progressive disorder characterized by unbalanced bone resorption, decreased bone mass, and deterioration of the bone microarchitecture, leading to decreased bone strength and increased fracture susceptibility [134]. The already significant global health burden of osteoporosis continues to increase alongside human longevity [135]. Postmenopausal women are especially at risk, with an associated osteoporosis prevalence ranging from approximately 20% in the United States and the European Union to nearly 40% in South Korea and Japan [136, 137, 138].

Osteoporotic fractures, the most serious being those of the hip and vertebrae, are of foremost concern to osteoporosis patients and fracture prevention is the primary target of pharmacologic treatment. The bisphosphonate (and frequent first-line therapy) alendronate and the selective estrogen receptor modulator (SERM) raloxifene are among the most popular antiresorptive agents for the prevention and treatment of postmenopausal osteoporosis [139, 140]. Based on existing randomized studies that compare alendronate and raloxifene separately to placebo [141, 142], alendronate seems to have superior fracture prevention benefits. However, few randomized studies consider head-to-head comparative effectiveness of osteoporosis drugs that should inform patient treatment decisions [143]. Some studies find improved bone mineral density in alendronate vs raloxifene users [144, 145], but improved bone mineral density has not been proven to decrease fracture risk [139, 142]. Observational studies can provide evidence missing from the randomized study literature, especially re-

garding rare but serious adverse events that require large study populations to detect. Two existing observational studies perform propensity score (PS) adjusted comparative effectiveness analysis on insurance claims databases and find no difference in both vertebral and nonvertebral fracture rates between alendronate and raloxifene patients [137, 146], but do not address suspected serious adverse events of long-term alendronate use such as atypical femoral fractures, esophageal cancer, and osteonecrosis of the jaw.

In this paper, we conduct a retrospective database cohort study investigating comparative risks of fractures and select adverse events among first-time initiators of alendronate and raloxifene. We utilize the extensive research network of the Observational Health Data Sciences and Informatics (OHDSI) collaborative [3] to conduct our study in nine clinical data sources including insurance claims sets and electronic medical records (EMR), representing a diversity of patient populations. We implement a suite of methods and analyses to address confounding and bias inherent to observational studies. We construct PS models using a large set of patient features that we believe offer more comprehensive confounding control than the limited models traditionally used in observational studies. We also conduct negative control experiments, an emerging tool in observational analytics [10], to adjust for residual systematic study biases that are unaccounted for by measured confounders.

8.2 Methods

8.2.1 Data Sources

We conduct a new-user cohort study comparing first-time users of alendronate with new users of raloxifene in nine clinical data sources encoded in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) version 5 [2] from participating research partners across the OHDSI community [3, 38]. Three data sources are electronic medical records: the University of Texas Cerner Health Facts Database (total of 2.4 million [M] patients), Columbia University Medical Center/NewYork-Presbyterian Hospital (4.5M) and Stanford University Hospital (2M). Six data sources are claims records: OptumInsight's

Clinformatics^{TM}Datamart (Eden Prairie, MN) (CEDM, 40.7M), Truven MarketScan Commercial Claims and Encounters (CCAE, 122M), Truven MarketScan Multi-State Medicaid (MDCD, 17.3M), Truven MarketScan Medicare Supplemental Beneficiaries (MDCR, 9.3M), IQVIA PharMetrics Plus (P-Plus, 105M), and the Korean National Health Insurance Service - National Sample Cohort (NHIS NSC, 1.1M). OHDSI network studies are carried out through a federated model, where the access to data and statistical testing executes inside the firewall of the research partners' infrastructure on de-identified patient information, and the research coordinators collect aggregate results absent of patient-level information for meta-analysis, interpretation, and manuscript generation. Each partner has obtained the necessary Institutional Review Board (IRB) approval or exemption to participate.

8.2.2 Study Design

This study follows a retrospective, observational, comparative cohort design [147]. We include women over 45 years old who are first time users of alendronate or raloxifene, and who have a diagnosis of osteoporosis in the year prior to treatment initiation. Patients are required to have continuous observation in the database for at least one year prior to treatment initiation and 90 days after. We exclude patients with a previous diagnosis of hip fracture, high-energy trauma, or other diseases related to pathological fractures (including Paget's disease), as well as patients with prior hip replacements or exposure to any bisphosphonate (including alendronate) or the SERMs raloxifene and bazedoxifene. We restrict the study time from January 2001 to February 2012 because relative drug utilization rates are more stable during that period across data sources. We use raloxifene as the reference treatment. Full cohort details, including concept codes, are provided in the Supplementary Material.

The primary outcome of interest is osteoporotic hip fracture, while secondary outcomes include vertebral fracture and suspected adverse events of long-term alendronate therapy: atypical femoral fracture (AFF), osteonecrosis of the jaw (ONJ), and esophageal cancer. We begin the outcome risk window at 90 days after treatment initiation, and exclude patients with prior occurrence of that outcome before the risk window. As our primary analysis, we

have elected before executing the study to end the outcome time-at-risk window when the patient is no longer observable in the database, analogous to an intent-to-treat design. In addition, to assess the sensitivity of our results to this decision, we consider an alternative analysis in which we end the time-at-risk window at first cessation of the continuous drug exposure, analogous to an on-treatment design. Continuous drug exposures are constructed from the available longitudinal data by considering sequential prescriptions that have fewer than 30 days gap between prescriptions. Due to database encoding difficulties in constructing continuous drug exposure periods, we exclude the PharMetrics, Cerner, and NHIS NSC data sources from the alternative analysis.

8.2.3 Statistical Analysis

We conduct our cohort study using the Open-Source OHDSI CohortMethod R package [87], with large-scale analytics achieved through the Cyclops R package [28]. We use propensity scores – estimates of treatment exposure probability conditional on pre-treatment baseline features in the one year prior to treatment initiation – to control for potential confounding and improve balance between the target (alendronate) and reference (raloxifene) cohorts [21]. We include all available patient demographic and drug exposure, medical condition and procedure codes as covariates in the PS model as potential confounders instead of a prespecified set of investigator-selected confounders. Detailed covariate information is provided in the Supplementary Material.

We fit the PS model using an L1-regularized large-scale logistic regression model [20, 33], with L1 penalty hyperparameter selected through 10-fold cross-validation. We transform the PS to preference scores that account for differences in drug prevalence and availability [64], trim these preference scores <0.25 and >0.75, and create five equally-sized strata. To assess successful confounding control, we present the preference score distributions and covariate balance metrics.

We estimate comparative alendronate vs raloxifene hazard ratios (HR) using a Cox proportional hazards model stratified by the preference scores strata. As confounding control is

addressed by PS adjustment, we include treatment exposure as the sole covariate in the outcome model. We report the estimated HR for each outcome along with their associated 95% confidence intervals (CI) obtained from the profile likelihood [62]. We combine estimates from data sources into a summary HR using a random effects model meta-analysis [148]. Finally, we present Kaplan-Meier survival plots for the primary outcome to characterize the contour of hip fracture risk over time.

Propensity score adjustment addresses confounding from measured covariates, while residual bias after PS adjustment derives from unmeasured and systematic sources within our data and study design. To estimate such residual bias, we conduct negative control outcome experiments with 147 negative control outcomes [15], identified through a data-rich algorithm [61]. Negative control outcomes, separate of our study outcomes, are events believed to be unaffected by the studied treatments, thus having a presumed true HR of 1. See the Supplementary Material for the list of included negative controls. The distribution of the negative control estimates characterizes the study residual bias and is an important artifact from which to assess the study design [11]. Fitting an empirical null distribution to these negative control estimates allows us further to calibrate the p-values for the outcomes of interest [10].

8.3 Results

8.3.1 Population Characteristics

Across all data sources, we identify 283,586 alendronate patients and 40,463 raloxifene patients for the primary hip fracture analysis, totaling 1,076,597 and 156,080 patient-years of observation, respectively; corresponding cohort sizes for all study outcomes are similar (Table 8.1). For the hip fracture outcome, Table 8.2 further partitions these patients by data source. Approximately, 98% of the patient come from claims databases, and the relatively few raloxifene users from the Columbia and Stanford EHRs suggests that these data sources contribute only modest information. Figure 8.1 presents the distributions of study entry year

and age at study entry for each data source. By these two characteristics, the data sources span a diversity of patient populations. The on-treatment alternative analysis yields similar cohort sizes for included data sources (Supplementary Tables 7-8).

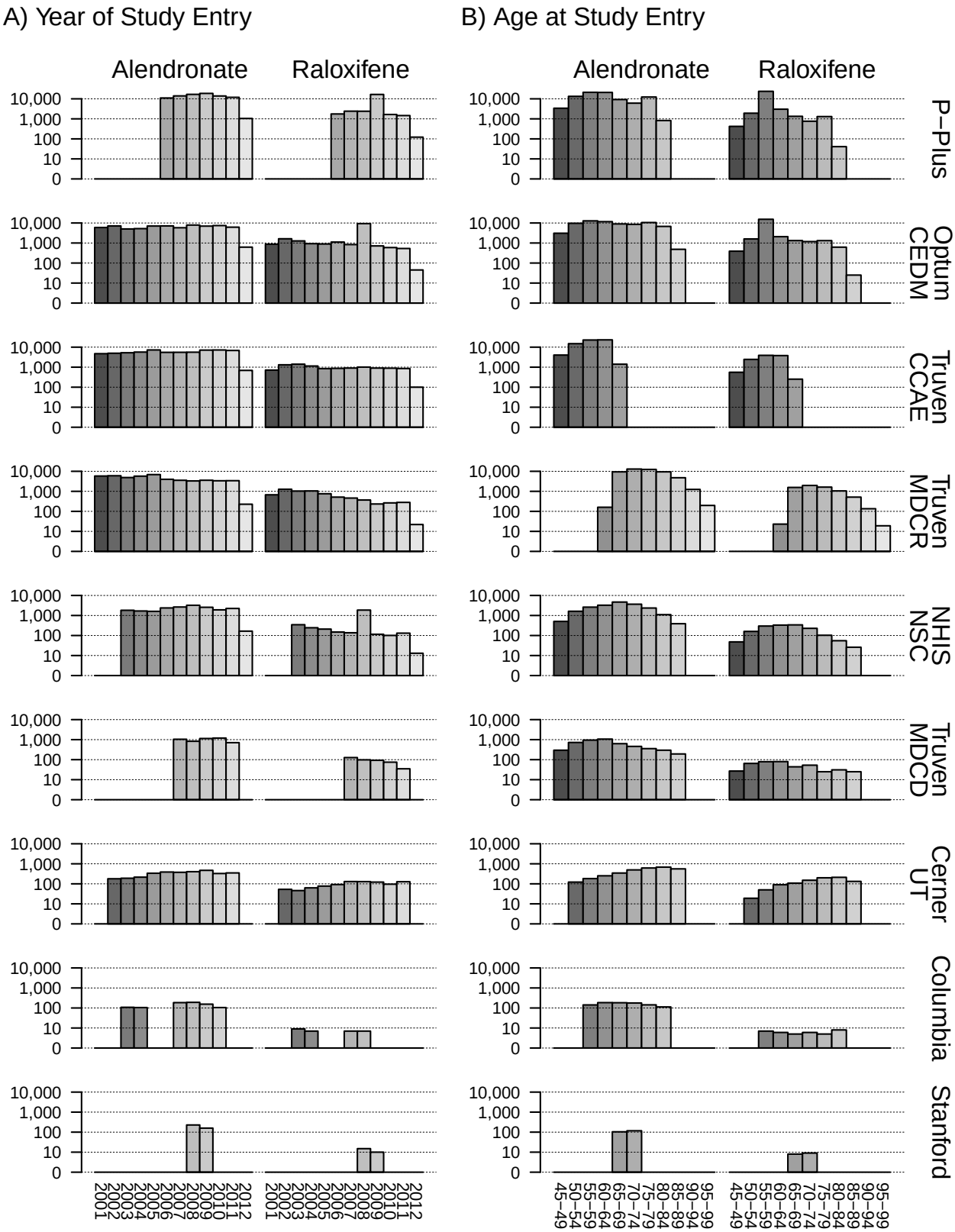


Figure 8.1: A) year of and B) age at study entry, stratified by drug exposure and data source. Note patient counts are on the log-scale

Outcome	Alendronate				Raloxifene			
	Patients	Years	Events	Rate	Patients	Years	Events	Rate
Hip fracture	283,586	1,076,597	8,051	7.48	40,463	156,080	1,033	6.62
Vertebal fracture	279,497	1,058,734	8,659	8.18	40,051	154,031	1,134	7.36
Atypical femural fracture	283,894	1,094,049	1,244	1.14	40,503	158,722	109	0.69
Esophageal cancer	283,981	1,096,983	234	0.21	40,482	158,858	35	0.22
Osteonecrosis of the jaw	284,079	1,097,499	101	0.09	40,511	158,972	9	0.06

Table 8.1: Size of study cohorts for each outcome of interest in primary analylsis. Rate: incidence per 1,000 person-years

Data source	Alendronate			Raloxifene		
	Patients	Years	Events	Patients	Years	Events
P-Plus	78,155	245,336	1,216	10,742	34,711	117
Optum CEDM	67,100	262,467	2,495	10,167	40,528	323
Truven CCAE	64,003	228,085	432	10,534	38,655	63
Truven MDCR	47,576	210,908	3,247	6,459	29,840	457
NHIS NSC	17,766	94,139	313	1,314	7,823	26
Truven MDCD	4,570	16,454	209	369	1,340	19
Cerner UT	2,644	8,867	100	787	2,740	23
Columbia	1,131	7,696	24	49	298	<6
Stanford	641	2,645	15	42	145	<6
Total	283,586	1,076,597	8,051	40,463	156,080	1,033

Table 8.2: Number of patients, observation years, and number of hip fracture events in study cohort by data source in primary analysis

8.3.2 Primary Outcome Assessment

In the primary analysis, there are 8,051 and 1,033 total hip fractures in the alendronate and raloxifene cohorts, corresponding to incidence rates of 7.48 and 6.62 events per 1,000 person-years (Table 8.1). The respective on-treatment alternative incidences are expectedly lower, at 5.35 and 5.32 (Supplementary Table 9). Neither the primary analysis (summary HR 1.03, 95%CI: 0.94 - 1.13) (Figure 8.2a) nor the on-treatment alternative (summary HR 0.88, 95% CI: 0.71 - 1.11) (Figure 8.2b) demonstrate a statistically significant difference between treatments. Figure 8.3 presents a representative Kaplan-Meier plot from the Optum CEDM data source absent of stratification. While the plot seems to show slower raloxifene user hip

fracture development, the PS stratified effect size estimate is statistically insignificant (HR 1.03, 95%CI: 0.92-1.16).

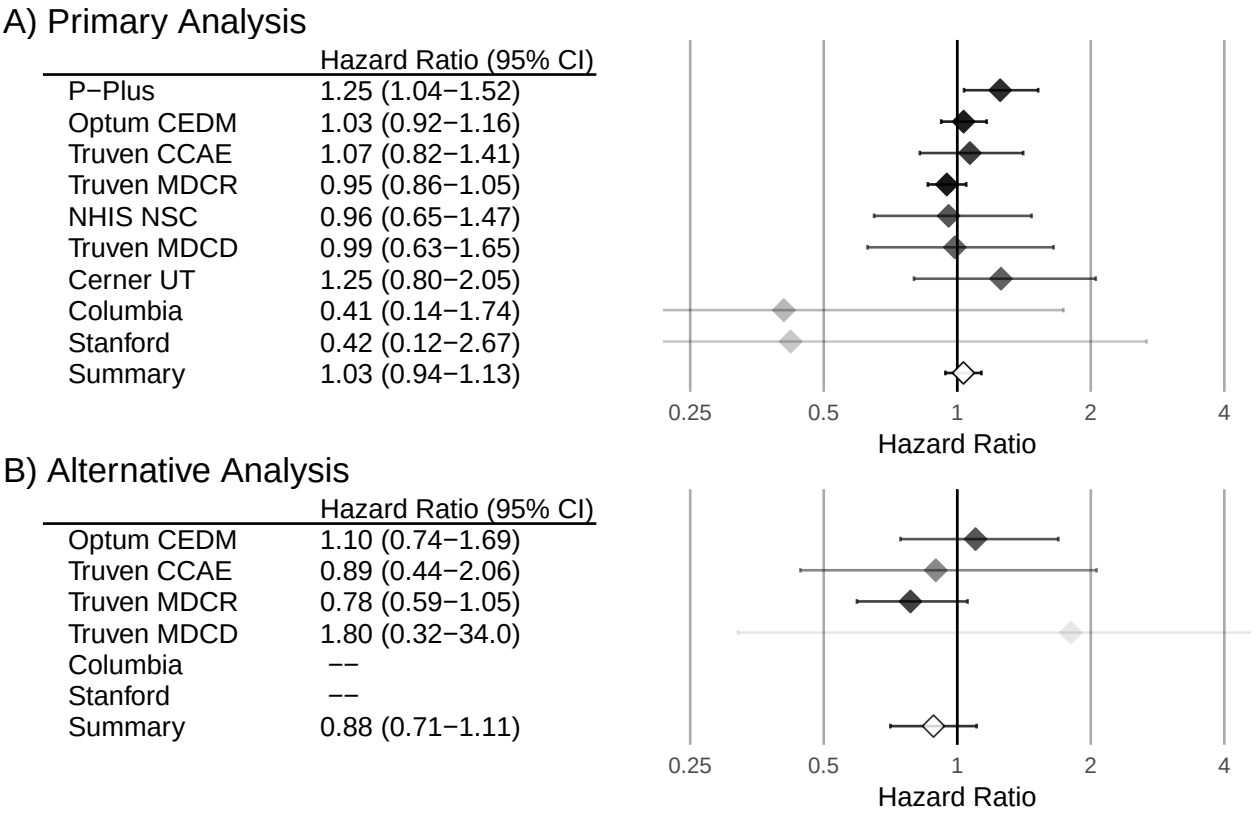


Figure 8.2: A) Primary and B) alternative analysis hazard ratios for hip fracture. More precise estimates have greater opacity. Missing HR from data sources with 0 raloxifene events

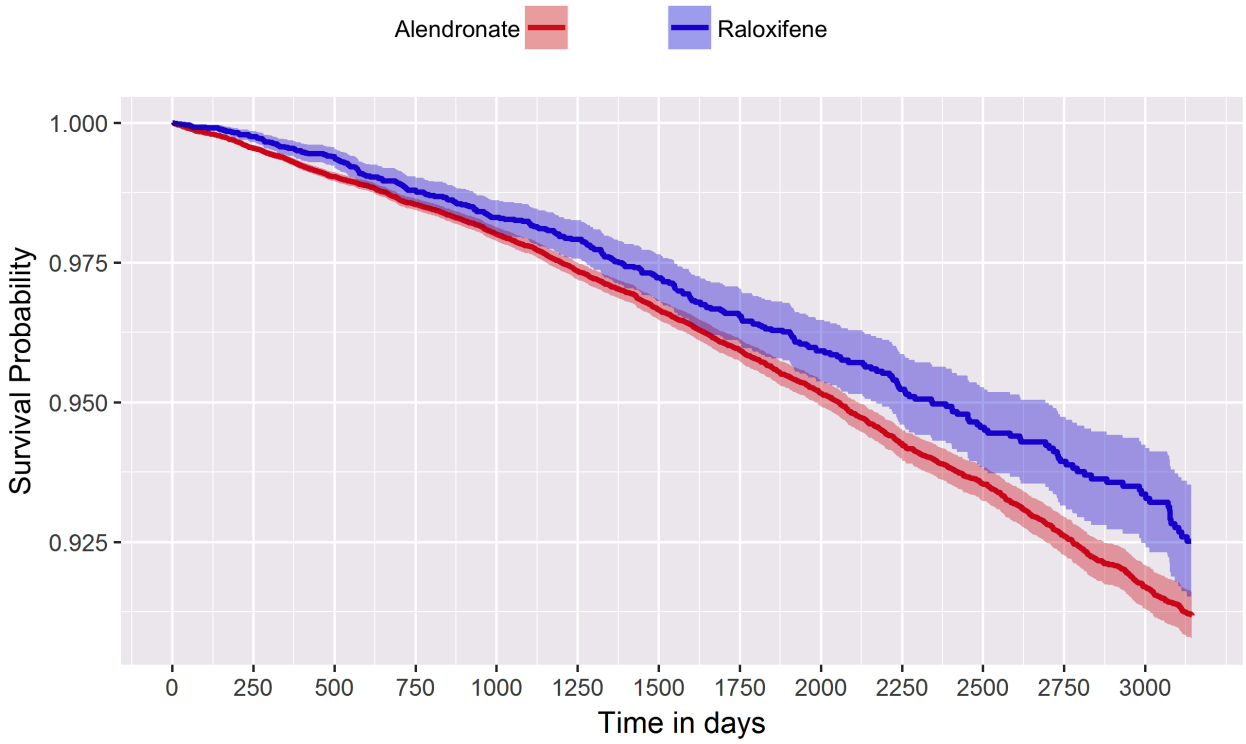


Figure 8.3: Kaplan-Meier plot for hip fracture outcome in Optum CEDM data source

8.3.3 Secondary Outcome Assessment

In the primary analysis, there are 8,659 vertebral fracture, 1,244 AFF, 234 esophageal cancer and 101 ONJ events among alendronate users, with corresponding crude incidence rates of 8.18, 1.14, 0.21 and 0.09 events per 1,000 person-years (Table 8.1). Among raloxifene users, there are 1,134 vertebral fracture, 109 AFF, 35 esophageal cancer and 9 ONJ events (incidence rates: 7.36, 0.69, 0.22 and 0.06). Alendronate users show a slightly higher vertebral fracture risk with statistical significance (summary HR 1.07, 95% CI: 1.01 - 1.14) (Figure 8.4a), and a markedly higher AFF risk (summary HR 1.51, 95% CI: 1.23 - 1.84) (Figure 8.4b). There is no significant difference in esophageal cancer risk (summary HR 0.95, 95% CI 0.53 - 1.70) (Figure 8.4c) or ONJ risk (summary HR 1.62, 95% CI 0.78 - 3.34) (Figure 8.4d).

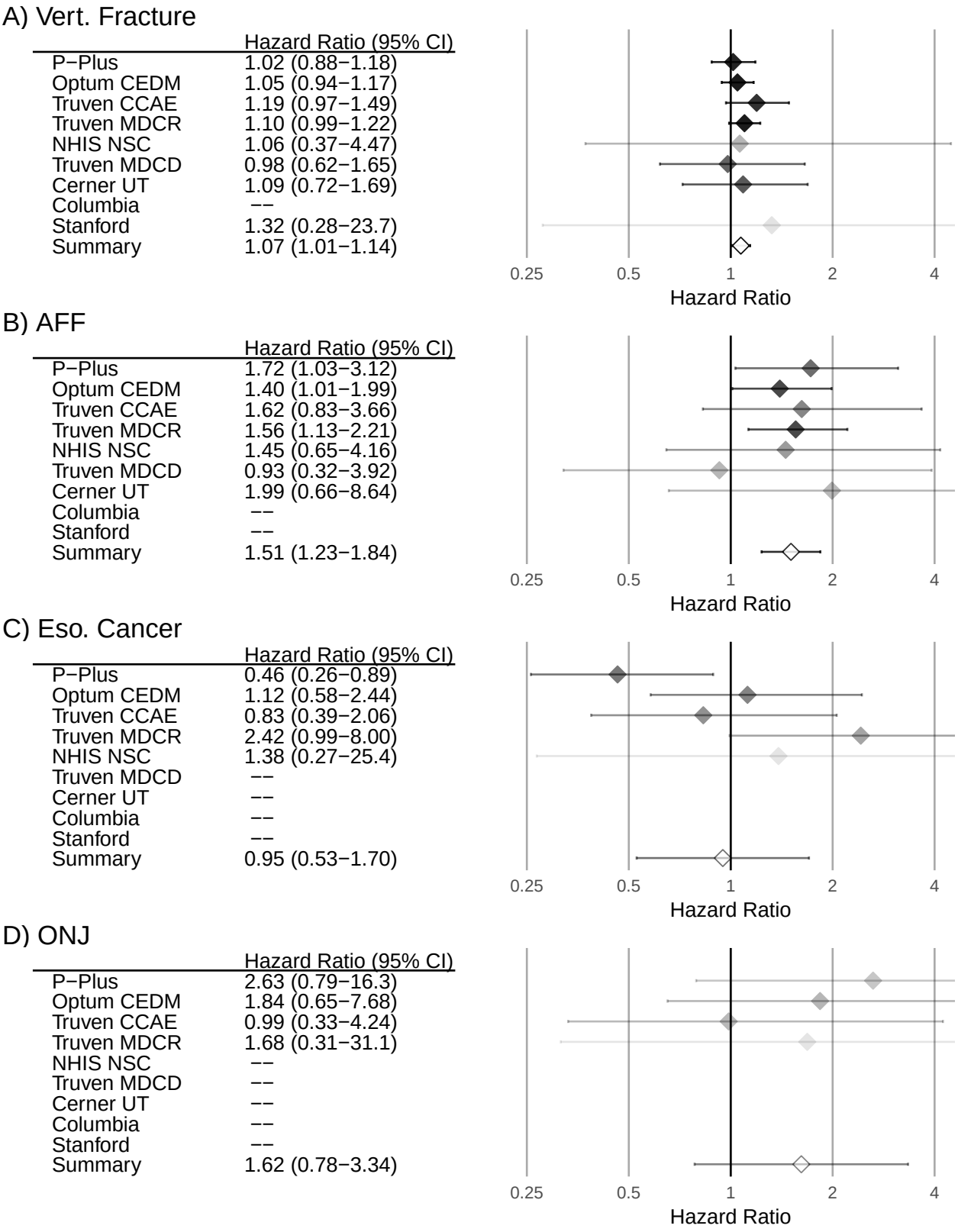


Figure 8.4: Primary analysis hazard ratios for A) vertebral fractures, B) atypical femoral fracture, C) esophageal cancer, and D) osteonecrosis of the jaw. More precise estimates have greater opacity. Missing HR from data sources with 0 raloxifene events

In the on-treatment alternative analysis, the respective rates for the four secondary outcomes are 6.28, 0.73, 0.11, 0.03 among alendronate users and 6.56, 0.35, 0.23, 0.00 among raloxifene users (Supplementary Table 9). Some data sources have 0 events among one or both treatment groups, and consequently have nonexistent HR estimates. We find no significant vertebral fracture risk (summary HR 0.87, 95% CI: 0.71 - 1.07) and lose all power in the other three hypotheses, with extremely wide confidence intervals for AFF and esophageal cancer and 0 raloxifene cohort outcomes for ONJ (see Supplementary Material: Analysis Results).

8.3.4 Cohort Balance

Treatment groups from real-world data require reasonable propensity score overlap to meaningfully conduct a comparative effectiveness study. Across all data sources, preference score distributions are generally similar, suggesting comparable prescription practices (Supplementary Figure 12). A large majority of patients have intermediate preference scores, and all data sources except Cerner and NHIS NSC display at most 10% loss to preference trimming to 0.25-0.75 (Table 8.3). Figure 8.5 shows a representative preference score distribution in the Optum CEDM data source.

Data source	Alendronate	Raloxifene	Total
P-Plus	10%	11%	10%
Optum CEDM	7.2%	5.9%	7%
Truven CCAE	3.4%	3.8%	3.4%
Truven MDCR	5.8%	6.5%	5.9%
NHIS NSC	12%	17%	13%
Truven MDCD	7.9%	14%	8.4%
Cerner UT	21%	19%	21%
Columbia	0%	0%	0%
Stanford	0%	0%	0%

Table 8.3: Percentage of cohort eliminated by trimming to 0.25-0.75 preference score

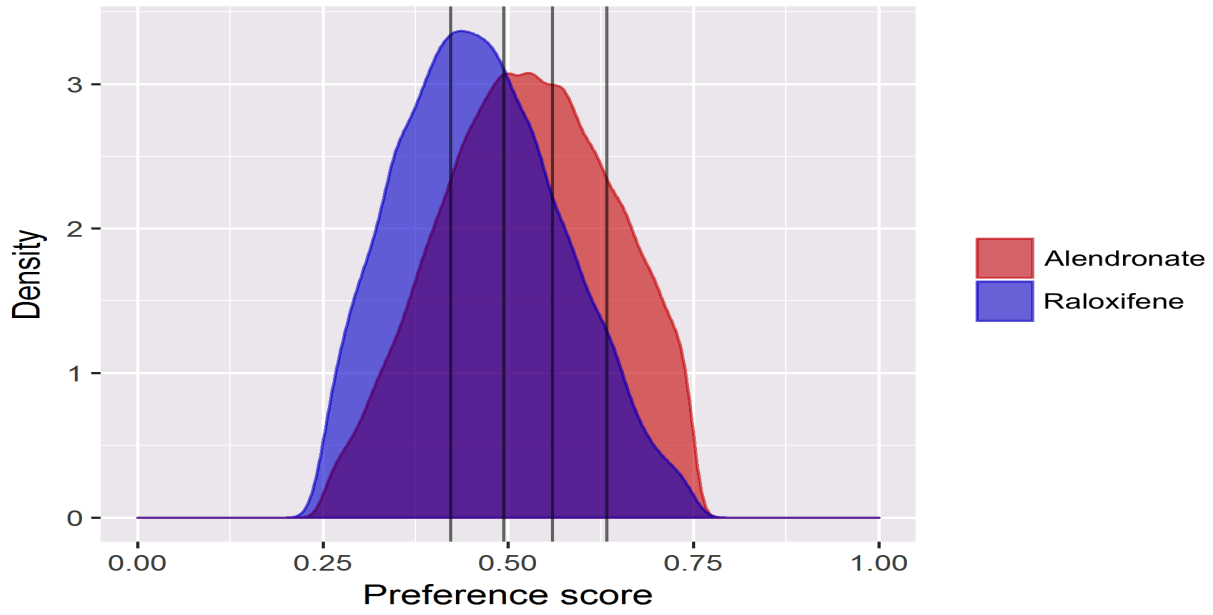


Figure 8.5: Preference score distribution of study subjects in Optum CEDM data source. Trimmed to 0.25-0.75, with black lines indicating stratification thresholds

We assess the covariate balance achieved through PS adjustment by comparing the standardized difference between treatment groups for all covariates before and after PS trimming and stratification, as shown graphically for Optum CEDM (Figure 8.6) and all data sources (Supplementary Figure 1), with summary statistics for all data sources shown in Table 8.4. In all but one data source (Stanford) that has poor PS differentiation, there are large decreases from PS adjustment in both the mean standardized difference and the proportion of covariates with standardized difference greater than 0.05.

Data source	Before PS	After PS
P-Plus	0.23	0.04
Optum CEDM	0.20	0.05
Truven CCAE	0.16	0.05
Truven MDCR	0.20	0.06
NHIS NSC	0.36	0.13
Truven MDCD	0.32	0.21
Cerner UT	0.46	0.13
Columbia	0.73	0.44
Stanford	0.45	0.44

Table 8.4: Mean standardized difference of all covariates before and after propensity score trimming and stratification, by data source

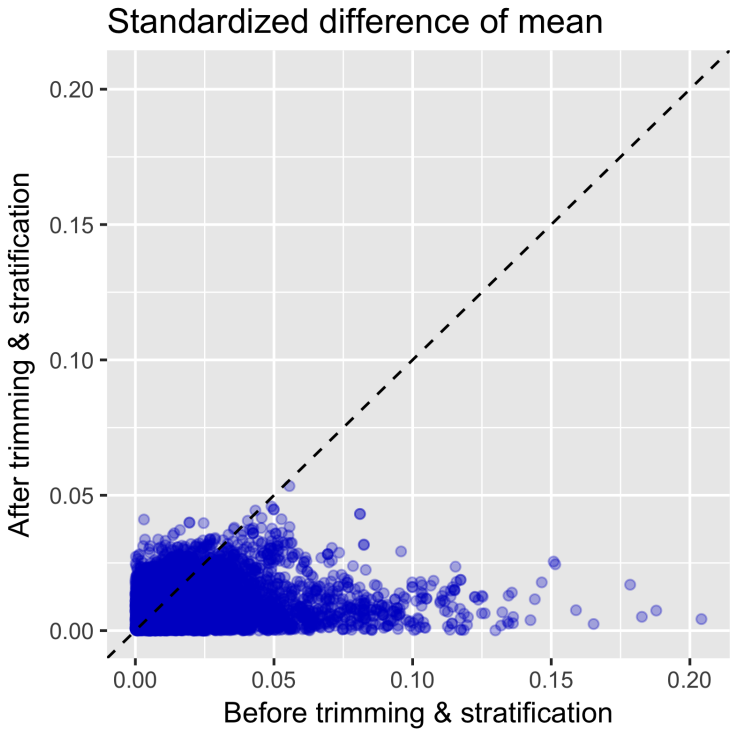


Figure 8.6: Standardized difference of covariates (1 dot = 1 covariate) in Optum CEDM study population before and after propensity score trimming and stratification

We analyze the covariate balance at the covariate-specific level by focusing on the top 20 originally unbalanced covariates from the Optum CEDM data source (Figure 8.7). Before PS adjustment, the alendronate group has higher proportions of bone disorders and higher

mortality risk as measured by the Charlson Comorbidity Index; the raloxifene group has higher proportions of gynecologic examinations and procedures, and gastrointestinal contraindications to alendronate. These unbalanced clinical covariates have been previously reported and are important potential confounders [149]. All of these top covariates become balanced through the PS adjustment process to absolute standardized differences below 0.05. In addition, all but one covariate have after-PS adjustment absolute standardized difference less than 0.05, and that one covariate (dependence on respiratory device in 365 days prior to treatment initiation) is still an improvement compared to the unadjusted cohort. Analysis of the top unbalanced covariates for the other data sources also show similar balance improvements for potential confounders (Supplementary Figures 2-10). Overall, the PS adjustment process produces large improvements in covariate balance that reduces the impact of potential confounding in our effect estimates.

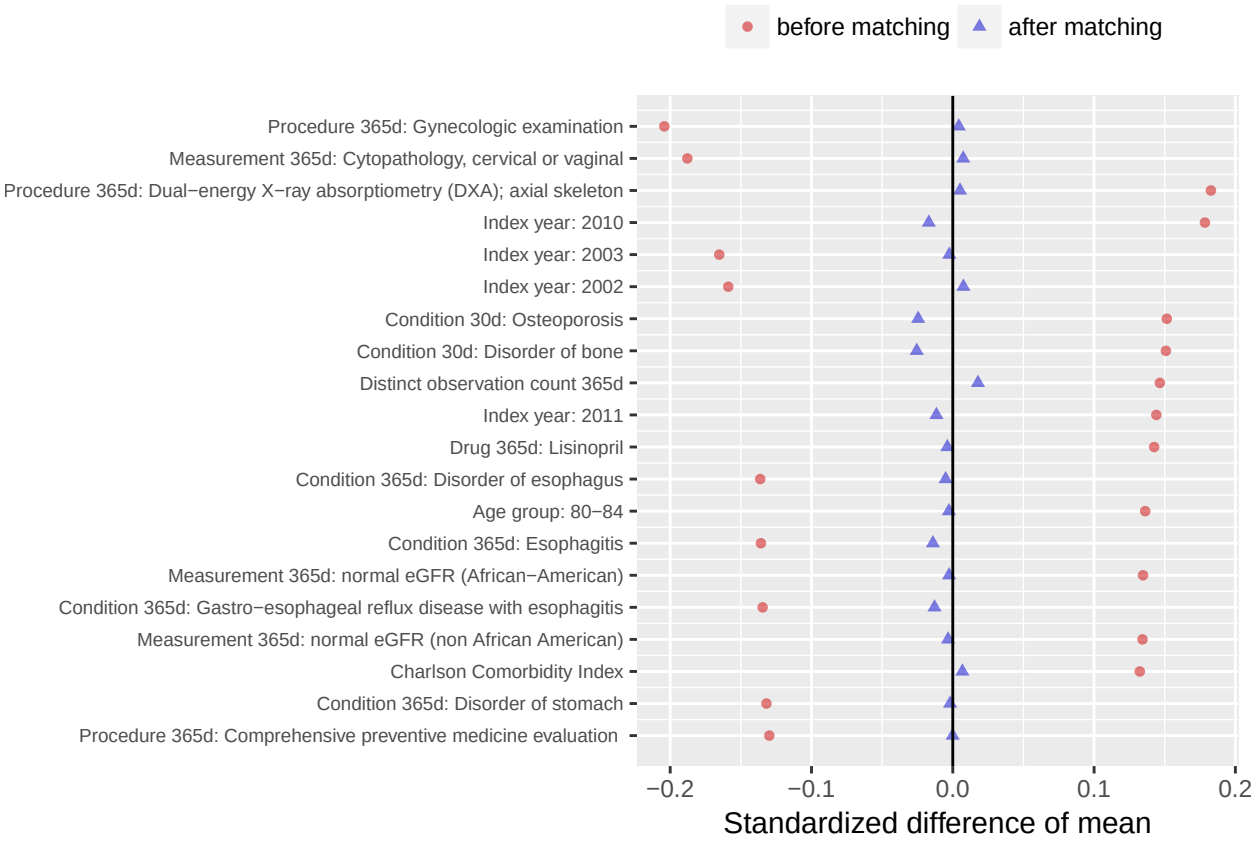


Figure 8.7: Top 20 covariates by absolute standardized difference between alendronate and raloxifene groups in Optum CEDM study. Positive difference indicates higher alendronate group frequency

8.3.5 Negative Control Outcomes

In the absence of bias, 95% of the negative controls estimates' 95% confidence intervals are expected to include the presumed null HR of 1. In the Optum CEDM study primary analysis, we see that 141/147 (96%) of the CIs do so (Table 8.5). Figure 8.8a shows the corresponding distribution of HR estimates and their associated standard errors from each of the 147 negative control outcomes for the same study. 143/147 (97%) of the estimates lie above the dotted line that represents the theoretical p-values; the slight difference from the 141 CIs containing 1 is due to asymptotic p-value assumptions. The orange region in Figure 8.8a represents the 95% threshold for calibrated p-values [10]; in the Optum CEDM study 142/147 of the estimates lie above this region and thus accept the null effect hypothesis. The negative control estimates are closely distributed around the presumed null value, and Figure 8.8b reaffirms the similarity between the negative control p-values under the theoretical calculation and under the calibrated empirical calculation. Table 8.6 gives the calibrated p-values for the hip fracture outcome primary analyses for all data sources; no estimate changes statistical significance as a result of calibration. Overall, negative control results show low residual bias across data sources for both primary and alternative analyses (Supplementary Figures 13 - 16), giving further credence to the relative unbiasedness of our treatment effect estimates.

	Empirical Null Dist.			Coverage of Null Effect		
Data source	Mean	SD	Controls	Empirical CI	Theoretical p	Calibrated p
P-Plus	-0.00803	0.0352	147	135 (92%)	135 (92%)	138 (94%)
Optum CEDM	-0.0106	0.0157	147	141 (96%)	143 (97%)	142 (97%)
Truven CCAE	-0.0221	0.014	146	139 (95%)	139 (95%)	141 (97%)
Truven MDCR	-0.0345	0.0201	146	133 (91%)	133 (91%)	135 (92%)
NHIS NSC	-0.00491	0.0162	122	117 (96%)	119 (98%)	118 (97%)
Truven MDCD	-0.0462	0.0247	126	120 (95%)	123 (98%)	122 (97%)
Cerner UT	0.0627	0.0373	105	99 (94%)	103 (98%)	102 (97%)
Columbia	-0.542	0.0178	53	51 (96%)	51 (96%)	52 (98%)
Stanford	-0.964	0.0816	35	32 (91%)	32 (91%)	35 (100%)

Table 8.5: Empirical null distribution constructed from negative controls, and the number of estimates that do not reject the null effect hypothesis. Empirical confidence intervals are from the profile likelihood, theoretical p-values are from the likelihood asymptotic distribution, and calibrated p-values are from the negative control calibrated standard errors. For the calibrated p-value, a leave-one-out design was used. Results by data source for primary analysis

	Original Estimate			Calibrated p-value		
Data source	Mean	SD	p-value	p-value	95% lb	95% ub
P-Plus	0.223	0.0982	0.0233	0.0316	0.016	0.0731
Optum CEDM	0.0328	0.0601	0.585	0.5	0.358	0.675
Truven CCAE	0.0657	0.137	0.631	0.528	0.441	0.624
Truven MDCR	-0.0543	0.0508	0.286	0.738	0.513	0.959
NHIS NSC	-0.0449	0.209	0.83	0.849	0.748	0.957
Truven MDCD	-0.0125	0.246	0.96	0.887	0.7	0.994
Cerner UT	0.227	0.241	0.346	0.519	0.316	0.765
Columbia	-0.9	0.645	0.163	0.581	0.386	0.798
Stanford	-0.865	0.795	0.276	0.867	0.6	0.994

Table 8.6: Original estimates and p-values for hip fracture primary analysis, with negative control calibrated p-values. Bounds on calibrated p-values calculated from the 95% bounds of original estimate

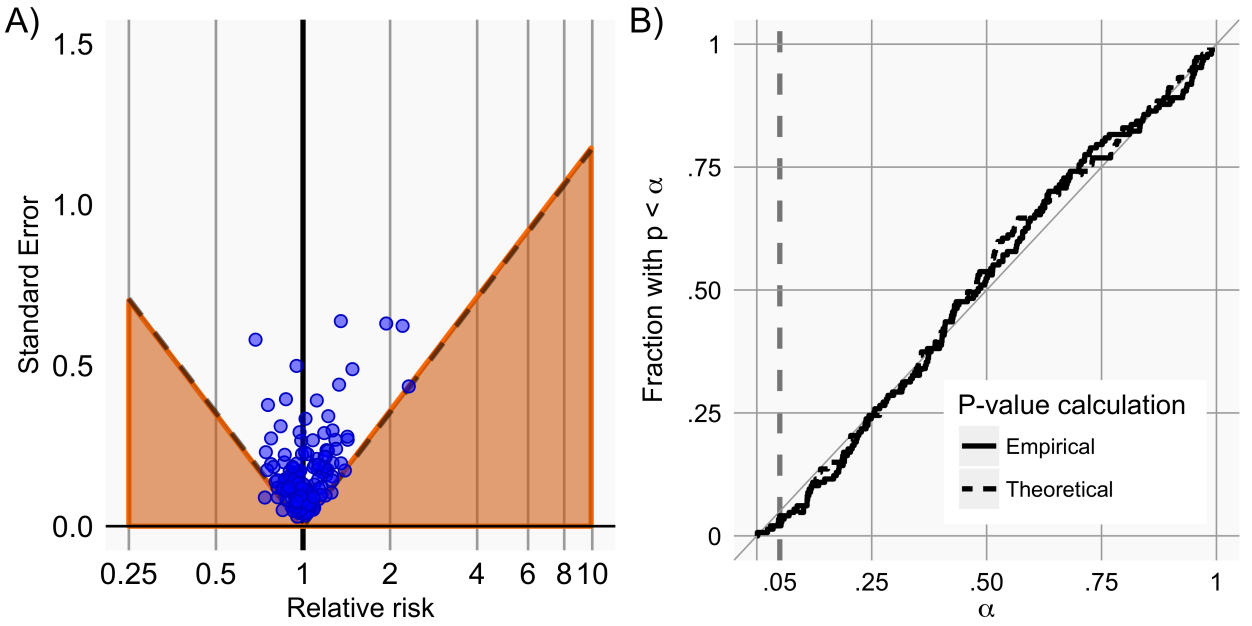


Figure 8.8: Negative control results from Optum CEDM primary analysis. A) Traditional and calibrated significance testing. Estimates below the dashed line have p $<$ 0.05 using traditional p-value calculation. Estimates in the orange areas have p $<$ 0.05 using the calibrated p-value calculation. Blue dots indicate negative controls. B) Calibration plot showing the fraction of negative controls with p $<$ α, for different levels of α. Both traditional p-value calculation and p-values using calibration are shown. For the calibrated p-value, a leave-one-out design was used

8.4 Discussion

Prevailing clinical wisdom favors alendronate as the first-line treatment option for osteoporosis patients against fracture [150, 151, 152, 153, 154]. However, head-to-head randomized studies of alendronate vs raloxifene have only shown increased bone mineral density with alendronate [144, 145], which do not necessarily relate to clinically observed fracture risk [139, 142]. Our results find little difference in hip fracture risk between new users of alendronate and raloxifene, and also find a small but statistically significant higher vertebral fracture risk with alendronate. Foster et al report non-significantly higher alendronate vertebral fracture risk compared to raloxifene using Truven CCAE and Truven MDCR data [137]. Our data sources are similarly individually non-significant, but together they reveal a statistically significant effect favoring raloxifene.

Growing concern over long-term bisphosphonate use has contributed to steep declines

in their prescription [155]. Previous studies report conflicting non-significant [156, 157] and positively significant [158, 159] estimates for AFF risk as a result of bisphosphonate-related suppression of bone remodeling [160]. We find that compared to raloxifene, alendronate does lead to increased AFF risk. Importantly, this well-known and statistically significant risk difference demonstrates that our data sources and study design furnish sufficient statistical power to detect a true difference in the hip fracture HR if one were to exist, given that the rates of AFF are almost an order-of-magnitude less than of hip fracture in our data.

Further, upper gastrointestinal mucosa stimulation is a common bisphosphonate adverse event [161, 162, 163, 164], but association with the related esophageal cancer is less established [165, 13, 166, 167, 168]. We find very similar esophageal cancer incidence between alendronate and raloxifene users, and no difference in hazard ratio. We similarly find no difference for osteonecrosis of the jaw, although our study is likely underpowered for this very rare adverse event.

Many sources of bias unique to retrospective, non-randomized data require attention in order to confidently interpret observational study results. Firstly, results may vary from database to database because of differences in study population, and the generalizability of a single study is low [169]. Our study benefits from a large population (over 300,000 patients) ranging from a diversity of data sources held by multiple data partners in the OHDSI community. Secondly, results from different observational studies are hard to compare to one another due to differences in study implementation details. Our OHDSI network study utilizes a common data vocabulary, standard research protocol, and shared implementation software to reduce study heterogeneity from implementation specifications.

Thirdly, observational studies necessarily suffer from confounding due to non-random treatment assignment. Propensity scores that model the treatment assignment probability are a popular tool to address such confounding. While there are many different ways to build pre-treatment covariates and to construct a PS model, the predominant approach involves the investigator's manual selection of suspected confounders. However, this approach may introduce bias into the treatment effect estimate [35], and different investigators often arrive at different expert-selected PS models. Our PS approach builds an expansive model that

includes all available pre-treatment patient features and selects relevant confounders through an automated regularization procedure. We additionally validate the performance of our PS adjustment by studying the preference score distributions and covariate balance metrics in each data source. Overall, we demonstrate large covariate balance improvements, suggesting promising control of observed confounders in our study.

A common finding of meta-analyses, either of observational studies or randomized trials, is that different studies attempting to estimate the same quantity produce entirely non-overlapping confidence intervals. Typically constructed through statistical asymptotic theory, reported confidence intervals only capture the element of random error, which becomes smaller with larger sample size. The remaining differences among different studies arises from non-random error, including study population differences, heterogeneous measurement error, implementation discrepancies, and systematic differences between data sources. Combining divergent study results without addressing these latter sources of bias defeats the purported benefit of meta-analyses to leverage the larger aggregate sample sizes across studies to reduce random error. In addition to demonstrating confounding control that should limit estimate deviations from study population differences and using standard research protocols and tools that should limit implementation discrepancies, our study addresses systematic error in each data source through negative control analyses. We use negative controls to quantify systematic bias for this alendronate vs raloxifene comparative effectiveness study, and use the empirical null distribution of negative control estimates to adjust the individual study p-values for our actual outcomes of interest. In this study, we find low amounts of systematic bias across data sources, providing credibility to our meta-analysis summary hazard ratio estimates.

Recent 2017 guidelines from the American College of Physicians have expressed alarm for raloxifene and other SERMs over cerebrovascular and thromboembolic event concerns borne out of multiple randomized studies comparing raloxifene to placebo [170]. In the context of the general lack of comparative effectiveness evidence for osteoporosis pharmacologic agents, our study focuses on fracture outcomes and select adverse events associated with alendronate therapy. Further comparative effectiveness research should additionally focus on the adverse

events associated with raloxifene.

Our study carries several limitations. Bias from measured and unmeasured sources cannot be ruled out of any observational study, this one included. Data derived from electronic medical records and insurance claims are naturally noisy with missing and misclassified values, and unknown patient histories prior to database entry; our negative control experiments are just one approach to address systematic study bias. Additionally, several of our insurance claims data sources provided much larger study populations that proportionately dominate the smaller data sources in the meta-analysis. As electronic medical records differ in fundamental ways from claims databases, either separate analyses or more complex meta-analysis weighting schemes may accentuate their unique differences. Having said that, several of our participating electronic medical record data sources have very little treatment or outcome data, and may not be as suitable for comparative effectiveness studies.

8.5 Conclusion

In a retrospective, head-to-head comparative effectiveness study across nine data sources, we find that raloxifene users have a similar hip fracture risk, slightly decreased vertebral fracture risk, and fewer adverse atypical femoral fractures as compared with alendronate users.

CHAPTER 9

Safety and Effectiveness of Recombinant Human BMP-2 in Spinal Fusion Surgeries

9.1 Introduction

Back and neck pain are among the most common symptoms encountered in clinical practice, and contribute to high morbidity and excess health care expenditures in the range of $100 billion per year in the United States [171]. While management of spinal pain is primarily medical, over 100,000 patients yearly undergo spinal fusion surgery to treat underlying conditions such as disk herniation or degenerative disease [172]. Bone-morphogenetic proteins (BMPs) are growth factors that promote bone formation [173] and offer an alternative to iliac crest bone grafts (ICBG) typically used in spinal fusion surgeries. Of the many BMP subtypes identified, recombinant human (rh)BMP2 and rhBMP7 have been used in orthopaedic surgery, and rhBMP2 (which we refer to as just "BMP") is widely used in spinal fusion surgeries [174]. In 2002, the US Food and Drug Administration (FDA) approved a recombinant rhBMP2 device for anterior lumbar spine surgery. Supported by numerous industry funded studies confirming its efficacy and safety, BMP use skyrocketed to double digit percentages of all spinal fusion surgeries, both in the lumbar spine and for off-label use in the cervical and thoracic spine [175, 176, 177]. However, in 2008 the FDA released a "black box warning" regarding life-threatening soft tissue swelling in anterior cervical fusions with BMP [178, 179, 180]. Medtronic, the owner of the commercial BMP/Infuse product, released individual-patient data from its clinical trials and subsequently multiple systemic reviews found serious misrepresentations of adverse events (AE) in previous industry-funded publications [181, 182, 183, 184]. The reviews also found improved but non-significant effec-

tiveness benefits for BMP over bone graft. Due to concerns over BMP complications, BMP use peaked around 2007-2008 and fell dramatically afterwards [185, 186, 187].

BMP has been shown to improve primary outcomes in spinal fusion, producing as good or better fusion compared to ICBG and similar to lower reoperation rates [174, 188, 189, 190, 191, 192, 193]. However, BMP has been associated with numerous suspected adverse events (AE), the most serious among them postoperative wound complications, soft tissue swelling, cancer, radiculopathy, and ectopic bone formation / heterotopic ossification [194]. Nevertheless, many of these AE have not been confirmed through high quality, large-scale studies that adjust for potential confounding [182], perhaps explaining previously wary surgeons' slow re-embrace of off-label BMP use. Large-scale databases offer an effective tool for observational research in orthopaedic surgery [195], and several large-scale observational studies examine BMP complication concerns. Two studies, Cahill et al. [175] and Williams et al. [196], examine postoperative complications and find significantly increased BMP complications only with anterior cervical fusion. Veeravagu et al. [197] reports lower revision surgeries with BMP but also an increase in overall complications. Both Hindoyan et al. [198] and Savage et al. [199] find that BMP is associated with lower complication rates for lumbar fusion surgeries. However, these studies lack comprehensive control for potential confounding. Several report unadjusted odds ratios, or match BMP to non-BMP patients based only on a small number of patient characteristics.

Among the numerous suspected AE associated with BMP usage, cancer risk is arguably the most concerning. In some small randomized trials, BMP was associated with surprisingly high effect sizes with cancer, especially at high doses [200, 201]. However, this association with cancer is not seen in several small retrospective studies [202, 203], including studies focused on high dose BMP patients [204, 205]. Three large-scale retrospective studies on national databases regarding BMP cancer risk have been published, and they reach conflicting conclusions, with two finding no difference [206, 207], and a third finding lower BMP cancer risk [208]. These large-scale retrospective studies also lack comprehensive control for potential confounding. One [208] controls only for patient demographics, while the other two control for a patient's entire medical history only through a single Charlson Comorbidity

Index and diagnosis leading to the fusion surgery.

In this paper, we conduct a retrospective database cohort study across five longitudinal observational databases investigating repeat surgery and AE rates between BMP and non-BMP spinal fusion surgeries. For AE, we focus on postoperative infection, postoperative seroma and hematoma formation, radiculitis, heterotopic ossification, and cancer. We additionally focus on cancer outcomes by malignant and benign subtypes. Using research tools developed in the Observational Health Data Sciences and Informatics (OHDSI) community [3], we conduct our study in four claims databases and one electronic medical records database using a common research protocol and data vocabulary. We implement a suite of methods and analyses to address confounding and bias inherent to observational studies.

9.2 Methods

9.2.1 Data Sources

We conduct a new-user cohort study [38] comparing first-time recipients of spinal fusion surgery with and without BMP administration in five clinical data sources encoded in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) version 5 [2]. The four insurance claims databases are IBM MarketScan Commercial Claims and Encounters (US employer-based private payer; patient aged 65 years or older), Optum ClinFormatics (US private payer; primarily aged 65 years or younger), IBM MarketScan Medicare Supplemental Beneficiaries (US retirees; patients aged $>$ 65 years), IBM MarketScan Multistate Medicaid (US Medicaid enrollees; all ages). The electronic health record database is Optum Pan-Therapeutic (US health systems; all ages). In order mentioned, we abbreviate these five databases as CCAE, Optum, MDCR, MDCD, and PanTher.

9.2.2 Study Design

This study follows a retrospective, observational, comparative cohort design [38], comparing a target group (T) to a comparator group (C) for the risk of an outcome (O). Our target

group (T) consists of first-time recipients of any spinal fusion procedure, of at least 18 years of age at time of procedure, who had administration of bone morphogenetic protein around the time of their procedure. Our comparator group (C) consists of first-time recipients of any spinal fusion procedure, of at least 18 years of age, who did not have administration of bone morphogenetic protein. We restrict the study dates from January 1, 2003, around the time of BMP introduction to the market, to December 31, 2017. Full cohort details, including CDM concept codes, are provided in the Supplementary Material.

We conduct two different analyses in our study. In our primary analysis, we compare BMP users to non-users with regards to a primary outcome of interest (refusion surgery), and five secondary outcomes of interest that are suspected adverse events associated with BMP (radiculitis, postoperative infection, postoperative seroma/hematoma, heterotopic ossification, and cancer). Here we define refusion surgery as any subsequent spinal fusion surgery, regardless of type. Cancer includes any new neoplasm, benign or malignant, regardless of type. Table 9.1 details the risk windows for these primary analysis outcomes.

In our secondary analysis, we compare BMP users and non-users who have no history of any cancer with regards to 18 benign and malignant neoplasms – all benign neoplasms, all malignant neoplasms, and benign and malignant neoplasms of the following categories based on the International Classification of Diseases (ICD9): lip, oral cavity, and pharynx; digestive; thoracic and respiratory; connective (including bone, skin, and breast); genitourinary; lymphatic and hematopoietic; nervous; endocrine. The risk window for these cancer outcomes extends from 14 days after index date to the end of patient observation in the database.

9.2.3 Statistical Analysis

To control for measured confounding in our comparisons of target and comparator cohorts, we use propensity scores (PS), a predominant tool in retrospective studies [21, 33]. The PS is an estimate of the treatment assignment probability, which is unknown in observational studies. We build the PS model using a data-driven process through regularized

Outcome	Risk-window start (days from index)	Risk-window end (days from index)	Exclude previous outcome?
Refusion surgery	14	end of observation	NO
Postoperative infection	0	60	NO
Postoperative seroma/hematoma	0	60	NO
Radiculitis	14	end of observation	NO
Heterotopic ossification	14	end of observation	NO
Cancer	14	end of observation	YES

Table 9.1: Outcome cohort definitions for primary analysis. Outcomes are only counted within the risk window, defined relative to index date. When analyzing cancer, we exclude patients with prior recorded neoplasms.

regression [19], using pre-treatment patient characteristics as model covariates. These covariates include all observed clinical aspects, including demographics, previous conditions, drug exposures, procedures, clinical measurements and observations, and morbidity scores. See the Supplemental Material for detailed covariate descriptions. With the estimated propensity score, we perform variable length matching [59] with a maximum ratio of 10:1 and a standardized propensity score caliper of 0.20, and use a greedy matching algorithm [60]. We then estimate the hazard ratio (HR) of treatment to comparator for each outcome using a stratified Cox proportional hazard model. We combine estimates from data sources into a summary HR using a random effects model meta-analysis [209]. We report the HR instead of an odds ratio because we have access to longitudinal time-to-outcome data.

To control for the effect of unmeasured confounding, we employ negative and positive controls to quantify the systemic bias in our system and compute adjusted HR estimates and confidence intervals. Negative controls are outcomes *a priori* believed to not be differentially affected by the compared exposures, thus having a presumed true hazard ratio of 1 [15, 14]. The distribution of estimated hazard ratios from a set of negative controls serves as an estimate of the systemic bias present in a study, and can be used to calibrate p-values [10]. Furthermore, using synthetic positive controls constructed from negative controls, we are able to calibrate confidence intervals out of outcome hazard ratios [11]. We identify 100 negative controls for comparing BMP to non-BMP users through a combination of a data-rich algorithm [61] and consideration of the potential adverse events listed in the INFUSE BMP product manual. We use the empirical null distributions and synthetic positive controls to

calibrate each HR estimate and its confidence interval, and p-value. Statistically significant estimates have 95% confidence intervals (CI) that do not include the null hypothesis of no-effect, which corresponds to a calibrated p-value of less than 0.05 without correcting for multiple testing [210].

9.2.4 Software

This study was conducted using the software suite developed by the OHDSI community for conducting high-quality, large-scale observational clinical studies. The study design was formulated using the ATLAS open source software (`https://github.com/ohdsi/atlas`), which allows efficient specification of study parameters and automated construction of a R study package. The study package heavily uses the R CohortMethod package [27] for all high-level analysis. Underlying statistical regressions, including building the computationally intensive propensity score model, are serviced by the R Cyclops package [28]. These open source software greatly facilitate the process of conducting an observational clinical study using state-of-the-art methodology.

9.3 Results

For the refusion outcome, the proportion of patients who received BMP ranges from 12.1% in the MDCR database to 16.7% in the Panther database. The sole EMR database, Panther, has comparable cohort sizes to the two larger insurance claims databases CCAE and Optum. MDCR, the database with primarily patients over the Medicare age of 65 (Figure 9.1), has the largest decrease (approximately 65.6%) in cohort size when excluding patients with prior neoplasm records for the cancer outcome analysis. Figure 9.2 shows the proportion of spinal fusions with BMP by year for each database. In all five databases, the proportion of BMP surgeries began decreasing between 2007 and 2011, and has continued declining to very low levels (under 2%) by the end of our study period in 2017.

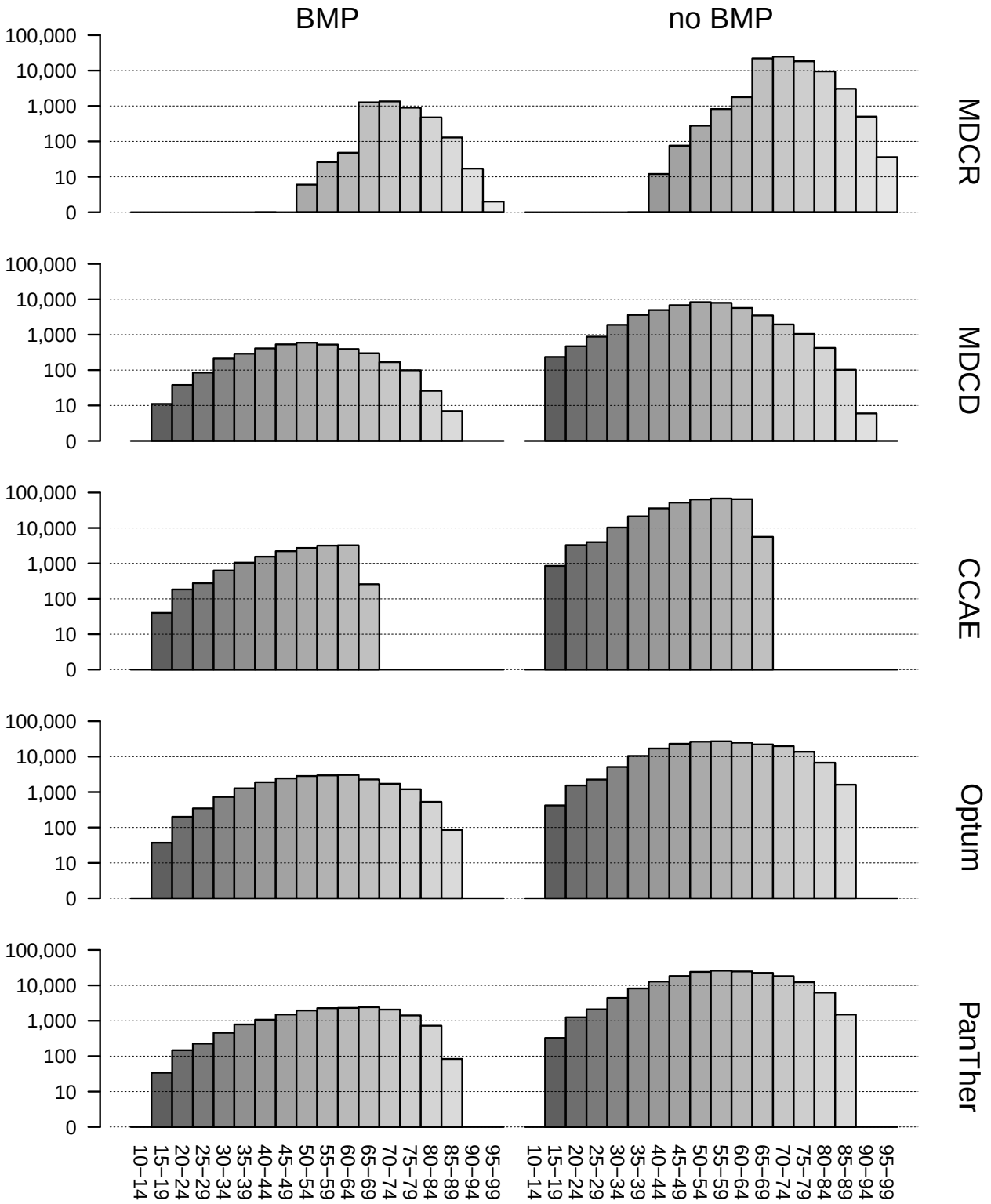


Figure 9.1: Age demographics by database

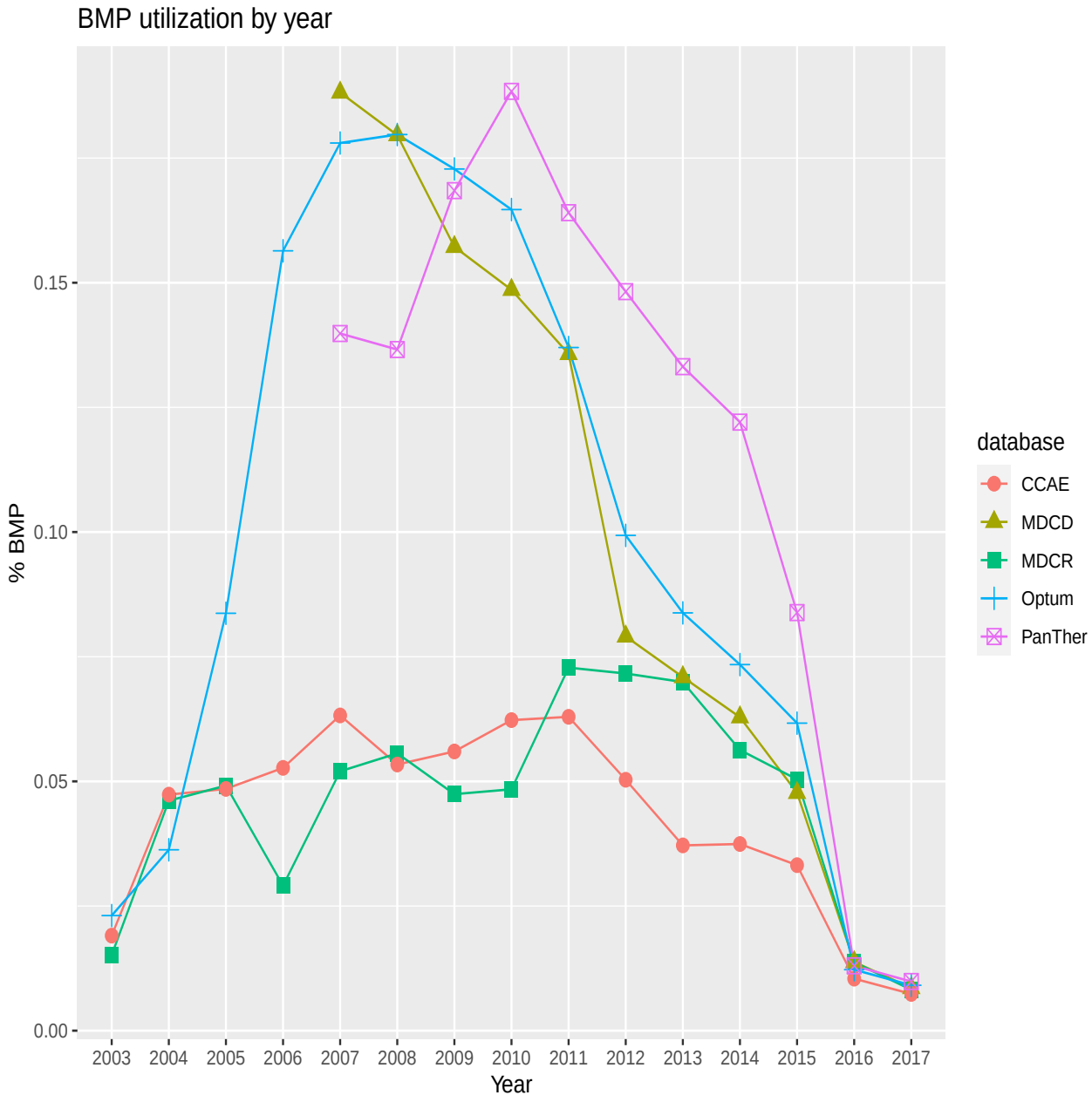


Figure 9.2: Proportion of spinal fusion surgeries with BMP by year and database

Across all data sources, we identify 60,427 patients with BMP and 349,771 patients without BMP for the primary refusion analysis, totaling 161,213 and 934,822 years of patient observation, respectively (Table 9.2). Corresponding population sizes are similar for the other outcomes, except for the new neoplastic disease outcome that excludes patients with prior cancer codes, that has 34,332 and 193,435 patients with and without BMP, respectively, with 80,790 and 444,507 years of patient observation. Note the much lower periods of patient

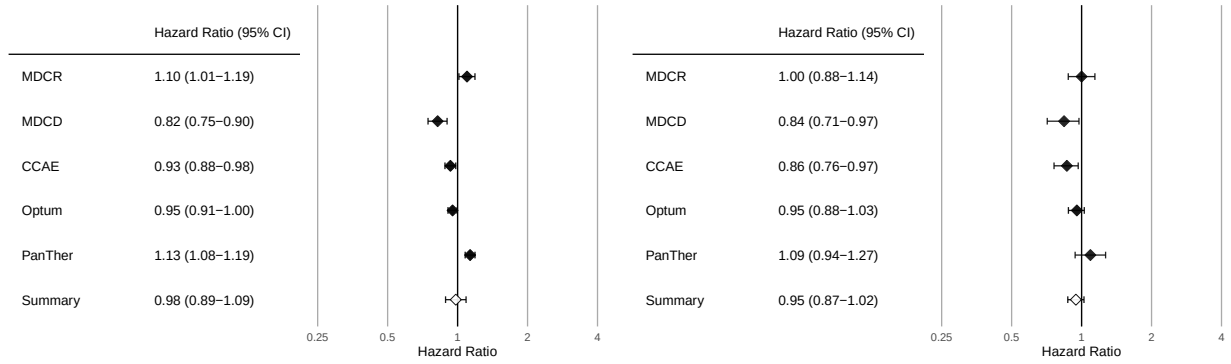
observation for the two postoperative outcomes, which have only a 60 day risk window. The rates for heterotopic ossification outcomes are very low, below 0.50 per 1,000 person-years.

	BMP				no BMP			
Outcome	Patients	Years	Events	Rate	Patients	Years	Events	Rate
subsequent fusion	60,427	161,213	8,829	54.77	349,771	934,822	50,822	54.37
seroma/hematoma	61,198	9,729	1,645	169.08	354,121	56,666	7,525	132.80
postoperative infection	61,198	9,761	1,671	171.19	354,121	56,613	9,398	166.00
radiculitis	60,427	161,482	7,917	49.03	349,771	906,607	52,609	58.03
heterotopic ossification	60,427	184,301	70	0.38	349,771	1,065,272	467	0.44
new neoplastic disease	34,332	80,790	9,396	116.30	193,435	444,507	53,744	120.91

Rate: incidence per 1,000 person-years

Table 9.2: Incidence for primary analysis

Figure 9.3 shows the results for the refusion outcome in the primary analysis. The uncalibrated hazard ratios (HR) are displayed on the left. All five confidence intervals are individually statistically significant and do not cross 1, and have small standard errors. The summary HR is 0.98 (95% CI: 0.89-1.09). After empirical calibration, only two of the confidence intervals are statistically significant, and the standard errors are larger, leading to wider confidence intervals. The summary HR is not statistically significant at 0.95 (95% CI: 0.98-1.02), and actually has a narrower CI than the unadjusted HR, despite the five component confidence intervals all being wider. The empirical calibration reduces the heterogeneity across databases.

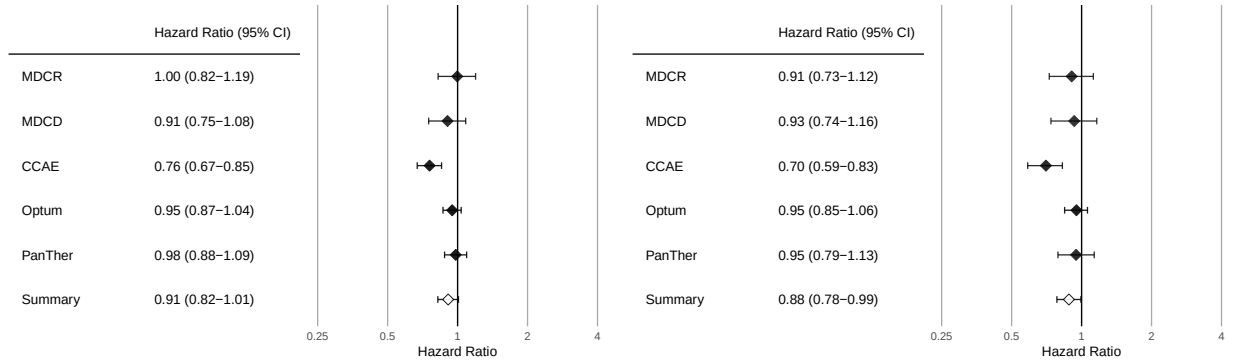


(a) Before calibration

(b) After calibration

Figure 9.3: Refusion outcome hazard ratios, primary analysis

Among the six outcomes in the primary analysis, only postoperative infection has a statistically significant calibrated summary HR is postoperative infection, for which BMP patients have lower rates than non-BMP patients, at 0.88 (95% CI: 0.78-0.99)(Figure 9.4, Table 9.3). Heterotopic ossification has a calibrated summary HR of 1.14 (95% CI: 0.86-1.51), but due to very low outcome counts this outcome has wide confidence intervals, and is nonsignificant.



(a) Before calibration

(b) After calibration

Figure 9.4: Postoperative infection outcome hazard ratios, primary analysis

Outcome	HR	lower	upper
Primary analysis			
subsequent fusion	0.95	0.87	1.02
seroma/hematoma	0.97	0.80	1.17
postoperative infection	0.88	0.78	0.99
radiculitis	0.99	0.93	1.06
heterotopic ossification	1.14	0.86	1.51
new neoplastic disease	0.98	0.93	1.03
Cancer analysis			
malignant all	0.96	0.90	1.04
malignant oral	1.01	0.77	1.32
malignant digestive	0.97	0.78	1.19
malignant thoracic	1.00	0.84	1.20
malignant connective	0.96	0.87	1.05
malignant genitourinary	0.85	0.70	1.03
malignant lymphoid/hematopoietic	0.93	0.80	1.08
malignant nervous	1.03	0.69	1.52
malignant endocrine	1.13	0.83	1.54
benign all	0.99	0.94	1.04
benign oral	0.94	0.74	1.18
benign digestive	0.95	0.88	1.02
benign thoracic	0.85	0.65	1.12
benign connective	1.01	0.96	1.07
benign genitourinary	1.00	0.89	1.11
benign lymphoid	1.36	0.88	2.11
benign nervous	0.89	0.71	1.11
benign endocrine	1.07	0.90	1.27

Table 9.3: Calibrated summary hazard ratios

In the secondary analysis of cancer outcomes by subtype, we have 33,447 and 188,478 patients with and without BMP for the analysis looking for all malignant neoplasms, with 97,893 and 543,468 years of patient observation, respectively (Table 9.4). The cohort sizes for the other cancer outcomes are similar. The raw, unadjusted rates of all malignant neoplasms are 26.16 per 1,000 person years among BMP patients and 27.00 among non-BMP patients. For benign neoplasms, the respective rates are 85.91 and 90.06. No outcome among the 18 studied subtypes demonstrates a statistically significant result (Table 9.3). The calibrated summary HR for all malignant neoplasms is 0.96 (95% CI: 0.90-1.04) and for all benign neoplasms is 0.99 (95% CI: 0.94-1.04).

Outcome	BMP				no BMP			
	Patients	Years	Events	Rate	Patients	Years	Events	Rate
malignant all	33,447	97,893	2,561	26.16	188,478	543,468	14,675	27.00
malignant oral	33,592	104,334	163	1.56	189,273	579,480	886	1.53
malignant digestive	33,587	103,797	442	4.26	189,227	576,895	2,374	4.12
malignant thoracic	33,588	104,033	340	3.27	189,224	578,236	1,810	3.13
malignant connective	33,567	101,375	1,236	12.19	189,069	562,078	7,443	13.24
malignant genitourinary	33,564	103,343	511	4.94	189,120	574,039	2,860	4.98
malignant lymphoid/hemato.	33,535	103,653	365	3.52	189,003	575,846	2,089	3.63
malignant nervous	33,587	104,438	104	1.00	189,220	580,277	566	0.98
malignant endocrine	33,593	104,487	61	0.58	189,270	580,390	396	0.68
benign all	33,388	84,204	7,234	85.91	188,010	461,674	41,579	90.06
benign oral	33,595	104,299	141	1.35	189,276	578,741	938	1.62
benign digestive	33,580	96,211	3,144	32.68	189,199	530,023	18,834	35.53
benign thoracic	33,593	104,396	83	0.80	189,273	579,684	604	1.04
benign connective	33,456	92,192	4,425	48.00	188,440	509,642	25,135	49.32
benign genitourinary	33,580	102,729	713	6.94	189,190	570,068	4,014	7.04
benign lymphoid	33,594	104,491	44	0.42	189,260	580,628	232	0.40
benign nervous	33,575	104,162	154	1.48	189,153	578,692	933	1.61
benign endocrine	33,582	103,934	248	2.39	189,214	577,752	1,303	2.26

Rate: incidence per 1,000 person-years

Table 9.4: Incidence for secondary cancer analysis

Through our propensity score adjustment, covariate imbalance is greatly reduced for our primary and secondary analyses. In the primary analysis (Figure 9.5), no covariate in any database has a post-matching standardized difference greater than 1, an encouraging sign for confounding control. In the secondary analysis, the covariate balance is similarly improved by propensity score matching, although two covariates in the MDCR database do have post-matching standardized differences greater than 0.1. One of these covariates is the Index year for 2011, which is the year of dramatic decline in BMP usage due to published articles about its safety. The other covariate is an indicator for Measurement of column chromatography (includes mass spectrometry) in the 365 days prior to index date, which does not seem to be related to BMP vs non-BMP usage.

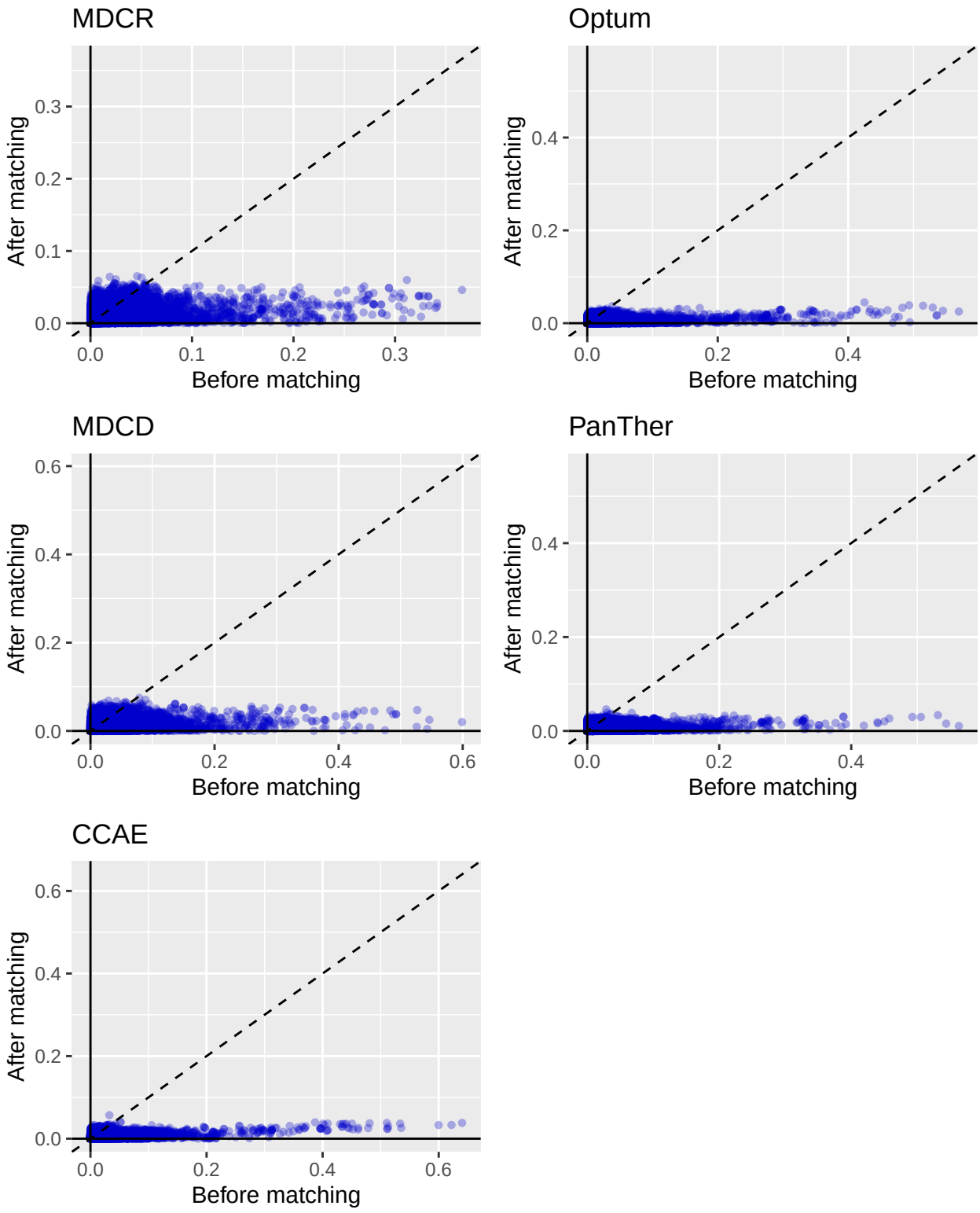


Figure 9.5: Covariate balance before and after matching, primary analysis. Each point represents the covariate balance for a single covariate.

The results of our negative and positive control experiments are demonstrated in Figure

9.6 for the primary analysis in the CCAE database. For the negative controls (true relative risk = 1), only 77% of the estimates' 95% confidence intervals include 1, and the null distribution fitted to the estimates has a mean slightly greater than 1. We see that the areas of statistical significance for the fitted null distribution, as delineated by the orange boundary, deviates from the regions of statistical significance for the unadjusted distribution (dotted lines). This is true for the three plots showing positive controls as well, demonstrating the need for empirical calibration of our estimates. As we increase the true relative risk from 1 to 4, the proportion of estimates that fall under the nominal 95% region decreases from 77% to 55.6%, mostly due to the estimates spreading out in their point estimates without an increase in their standard error. As seen in Figures 9.3 and 9.4, this has the effect of increasing the standard error of the calibrated estimates for the outcomes of interest.

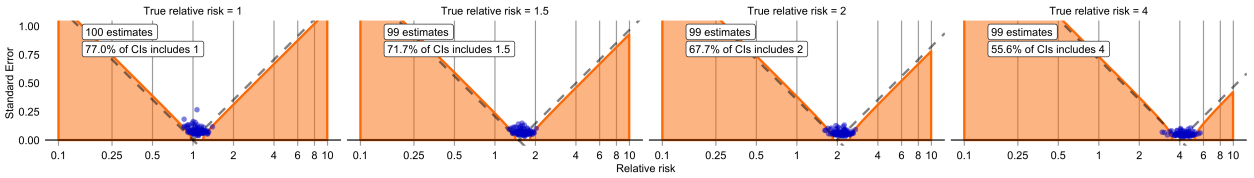


Figure 9.6: Negative and positive control distributions for CCAE database, primary analysis.

9.4 Discussion

Concerns over BMP adverse events borne out of misrepresented industry sponsored research lead to a large decline in the utilization of BMP starting around 2008. According to our study, this decrease in BMP utilization continued in dramatic fashion through 2017, the latest year studied, to only around 2% of spinal fusion surgeries. However, we find that BMP is safe to use in spinal fusion surgeries. Compared to spinal fusion surgeries without the use of BMP, surgeries with BMP have similar, non-statistically significant, hazard ratios for refusion surgeries, postoperative seromas/hematomas, radiculitis, heterotopic ossification, and cancer. In addition, across our five studied databases, BMP has a lower summary HR for postoperative infection, at 0.88 (95% CI: 0.78-0.99). In a detailed secondary analysis of 18 different categories of benign and malignant neoplasms, we find that BMP does not have

significantly different hazard ratios compared to non-BMP for any category.

Because of our generous definition of refusion operations as any subsequent spinal fusion surgery, we observe much higher rates of refusion than in other studies. Our observed rates of refusion are 14.6% in BMP patients and 14.5% in non-BMP patients. This compares to low-to-mid single digits reported in other studies [197, 198]. Our observed rates of radiculopathy are also higher than those reported in [198]. However, our observed rates of postoperative infection [197] and postoperative seroma/hematoma [199] do approximate that seen elsewhere. We observe heterotopic ossification to have an extremely low incidence of approximately 0.1%; this outcome is not well estimated in existing large-scale longitudinal observational studies. Radiologic heterotopic ossification due to BMP is commonly observed but rarely symptomatic, and thus not recorded in electronic medical records or insurance claims [211, 212, 213]. One reason for the discrepancy in outcome rates for refusion and radiculopathy is that they are dependent on follow-up time, which varies from study to study. While we also report rates per 1,000 person-years, these rates are not reported in the vast majority of BMP observational studies. Postoperative outcomes, even if they may vary in exact definition, are more consistently measured across studies, and lend themselves to more comparable rates. We advocate for increased reporting of rates per person-year for longitudinal observational studies.

Our secondary analysis finds a malignant neoplasm rate of 7.7% in BMP patients and 7.8% in non-BMP patients, over a mean follow-up of 2.89 years. These rates are comparable to that reported in two longitudinal observational studies of BMP cancer risk [207, 208] and lower to that in a third [206]. Our observed rates of benign neoplasms, 21.67% in BMP patients and 22.12% in non-BMP patients, are significantly higher than the mid-single digits reported in [208]. This discrepancy is largely due to our inclusion of benign neoplasms of the skin, which (categorized under connective) account for a majority of our benign neoplasm cases. Overall, our cancer rates agree with that reported elsewhere, lending credence to our analysis that finds no difference in cancer risk.

Our study benefits from a large combined sample size across five different databases. Across 24 outcomes in our primary and secondary analyses, only a single, postoperative

infection, has a significant summary HR, favoring BMP. However, there are many more significant results among the individual databases. For example, for the refusion outcome, two out of the five databases have a calibrated HR that significantly favors BMP. For the postoperative seroma/hematoma outcome (see Supplementary Material), one calibrated HR favors BMP and another favors non-BMP. Our meta-analysis combines these estimates into a single HR that reflects the full combined populations across databases, and produces a more credible estimate for the population-level effect size. Compared to a meta-analysis that combines estimates across studies, our study benefits from a consistent study design across databases. This reduces error arising from disparate study implementations, so that the remaining differences across databases more accurately reflect the underlying differences among study populations.

The excellent control of covariate balance in our primary and secondary analysis points to successful control of confounding through propensity score matching. Compared to other studies that adjust for baseline patient characteristics with or without propensity scores, our methodology includes a much larger set of covariates that reflects the totality of patient medical histories. The large models with many covariates that we use for our propensity scores have been shown to be superior to smaller models in regards to achieving covariate balance [19]. In addition, our negative and positive control experiments revealed appreciable amounts of residual bias that we control for by calibrating our hazard ratio confidence intervals. This calibration has an observed effect of shifting individuals HR estimates and also increasing standard errors, leading to fewer significant signals. For example, without calibration all five databases have a significant HR for refusion, but after calibration only two have a significant HR (Figure 9.3). We believe that our statistical methodology reduces spurious results that may be due to measured or unmeasured confounding, and more accurately estimates outcome effect sizes.

CHAPTER 10

Comprehensive Comparative Effectiveness of Antidepressant Treatments in Preventing Suicide and Suicidal Ideation

10.1 Introduction

Depression is the largest contributor to global disability, with over 300 million worldwide patients [214]. Among the many sequelae to clinical depression, suicide and suicidal ideations are arguably the most severe, and depression is a major factor in the close to 800,000 worldwide suicides that occur annually [214]. Antidepressant medications and psychotherapy are first-line treatments for clinical depression [215], though the association between drug classes and suicidality is not well understood in real-world settings [216]. While it is clear that pharmacological treatment for depression generally reduces suicide risk [217, 218], there are only few studies that perform comparative effectiveness among treatments with regard to suicide and suicidal ideation.

Clinical trials routinely assess suicide and suicidal ideation outcomes, but there are practical and ethical concerns in prospective studies of suicide [219]. In addition, clinical trials typically aim to certify a specific treatment's efficacy versus placebo, and rarely conduct comparative effectiveness analyses among multiple efficacious treatments. For example, the warnings against antidepressants in pediatric populations for fear of increasing suicidality are based in placebo-controlled trials, not comparative effectiveness studies [220]. In the absence of randomized and prospective trials comparing depression treatments, observational data offer a valuable resource in determining which antidepressant treatments pose less sui-

cidality risk. Existing comparative effectiveness studies are consistently retrospective and observational [221, 222, 223, 224].

Despite a possible (and controversial) association between antidepressants and suicidality [225], the clinical benefits of pharmacotherapy outweigh the risks [226], and some treatment is recommended [227]. But which treatment should it be? Individual studies contribute individual yet sometimes contrasting pieces to the puzzle. For example, fluoxetine has been shown in some studies to increase suicidality [228], while another study found that venlafaxine has higher risk than fluoxetine [223]. In [222], SSRIs and SNRIs are reported to have a similar risk of self-harm, while [224] found SNRIs have higher risk. Each published study investigates a single hypothesis, and is subject to its own biases arising from study implementation details, baseline population characteristics, and residual bias.

In prior work, we introduced a new paradigm for conducting high-throughput observational studies [229]. This high-throughput paradigm employs consistent and standardized methods to generate calibrated estimates for a large number of hypotheses, thus minimizing the reproducibility concerns of conducting single-hypothesis studies. We conducted a large-scale comparative effectiveness study analyzing 17 antidepressant treatments with regards to 22 clinical outcomes across 4 observational databases, thus generating thousands of hypotheses answered using a consistent methodology. In this paper we report on the clinical findings of our previous work [229] focusing on suicide and suicidal ideation as a primary clinical outcome of interest. We conduct all pairwise comparisons of 17 depression treatments and draw conclusions with regard to individual treatments and treatment classes.

10.2 Methods

10.2.1 Data Sources

For each pairwise comparison of depression treatments in each database, we conduct a new-user cohort study [38]. We conduct our study in four longitudinal insurance claims databases encoded in the Observational Medical Outcomes Partnership (OMOP) Common

Data Model (CDM) version 5 [2]. These databases are IBM MarketScan Commercial Claims and Encounters (US employer-based private payer; patient aged 65 years or older), Optum ClinFormatics (US private payer; primarily aged 65 years or younger), IBM MarketScan Medicare Supplemental Beneficiaries (US retirees; patients aged > 65 years), IBM MarketScan Multi-state Medicaid (US Medicaid enrollees; all ages).

10.2.2 Study Design

We follow a retrospective, observational, comparative cohort design [38] comparing first time users of a target treatment (T) to first time users of a comparator treatment (C) with regards to an outcome (O). Our target and comparator treatments are drawn from the 17 depression treatments listed in Table 10.1. In each comparison, the T and C cohorts are restricted to the years in which both treatments are observed in the database. Our outcome of interest (O) is suicide and suicidal ideation, defined using CDM5 concept codes (see Supplementary Material). Subjects with prior outcome are excluded from the analysis. We define the time-at-risk for the outcome to start on the day of treatment initiation and end on the last day of treatment administration, allowing for 30 day gaps in treatment continuation. We additionally conduct a sensitivity analysis in which the time-at-risk ends on the last day of patient observation in the database.

10.2.3 Statistical Analysis

Observational studies inherently contain confounding due to baseline differences among compared treatment populations. To adjust for measured confounding from encoded patient characteristics, we perform propensity score stratification. Propensity scores are estimates of treatment assignment that are ubiquitous tools for confounding control in observational analyses [21, 33]. We construct a propensity score for each comparison pair and data source using a data-driven process through regularized regression [19]. We utilize an expansive propensity score model that includes covariates for all conditions, procedures, drug exposures, etc. and allow the data-driven regression to select an ideal model. Variables with fewer than

Treatment	Class
Amitriptyline	Tricyclic antidepressant
Doxepin	Tricyclic antidepressant
Nortriptyline	Tricyclic antidepressant
Bupropion	Atypical antidepressant
Mirtazipine	Atypical antidepressant
Trazodone	Atypical antidepressant
Vilazodone	Atypical antidepressant
Citalopram	Selective seronotinin reuptake inhibitor
Escitalopram	Selective seronotinin reuptake inhibitor
Fluoxetine	Selective seronotinin reuptake inhibitor
Paroxetine	Selective seronotinin reuptake inhibitor
Sertraline	Selective seronotinin reuptake inhibitor
Desvenlafaxine	Serotonin norepinephrine reuptake inhibitor
Duloxetine	Serotonin norepinephrine reuptake inhibitor
Venlafaxine	Serotonin norepinephrine reuptake inhibitor
Electroconvulsive therapy	Other
Psychotherapy	Other

Table 10.1: Studied depression treatments

100 non-zero values are excluded from the model. See the Supplementary Material for a description of included propensity score model covariates.

With the propensity score, we stratify the target and comparator cohorts into 10 equally sized strata, and condition our outcome model on the strata. We use a Cox proportional hazards model for estimating out outcome effect size to obtain hazard ratios of the target versus comparator treatments. For each target-comparator-database combination, we report diagnostics including covariate balance before and after propensity score stratification, and plotted propensity score distributions. These diagnostics evaluate the effectiveness of our propensity score analysis in removing confounding and generating comparable stratified target and comparator cohorts.

Even after controlling for confounding from encoded variables, residual and systematic bias exists in observational studies [10, 230]. It is not possible to estimate this bias using hazard ratio estimates from our outcome of interest, that has an unknown true effect size. Instead, we employ negative control outcomes, which have a presumed null effect size, to quantify the size of the residual bias [15, 14]. We identify 52 negative controls for our study through a data-rich algorithm [61]. Using these negative control estimates, and positive

controls that we build using the negative controls, we are able to calibrate each outcome estimate, its confidence interval, and its p-value for rejecting the null hypothesis of no differential effect [10, 11]. Calibrated estimates with 95% confidence intervals that do not include 1, or equivalently a p-value less than 0.05, are considered statistically significant.

10.3 Results

Table 10.2 represents the average cohort sizes for each treatment in each database. Each treatment is independently compared to all other treatments, and in each individual comparison the cohorts represent first time users of the two treatments, and the patients could have had prior exposures to the uncompared treatments. Several treatments are not present in the MDCD and MDCR databases, namely vilazodone, paroxetine, and electroconvulsive therapy for MDCD and MDCR, and additionally doxepin and desvenlafaxine for MDCR. These treatments are also the ones with the smallest cohort sizes in the CCAE and Optum databases. Among the four medication classes (TCAs, Atypicals, SSRIs, SNRIs), TCAs have the smallest cohort sizes across all four databases. In all four databases, psychotherapy has larger cohort sizes than any other individual treatment, and is comparable to all SSRIs combined. Electroconvulsive therapy has by far the smallest cohort size among all treatments.

We computed 822 estimates for all available pairwise comparisons of the 17 depression treatments across the 4 databases. Figure 10.1 presents these data as class by class comparisons. Each individual point represents the hazard ratio for one comparison in one database, with the target treatment from the first class (ex: TCA) and the comparator treatment from the second class (ex: Atypicals). Under the null hypothesis of no differential effects, 95% of the estimates should have statistically nonsignificant 95% confidence intervals. Compared to all other treatment classes (atypicals, SSRIs, SNRIs, ECT, and psychotherapy), TCAs have noticeably more than 5% of statistically significant estimates, and the significant estimates almost all favor TCAs as having fewer suicide and suicidal ideation outcomes. After TCAs, the next most favorable class of treatments is the atypical antidepressants, that have

	CCAE	MDCD	MDCR	Optum
Amitriptyline	42480	9146	4067	22558
Doxepin	16912	2908	0	8207
Nortriptyline	22999	2617	3070	11507
Bupropion	195458	17928	12140	95990
Mirtazapine	56275	12773	17960	39216
Trazodone	152665	27162	14701	78680
Vilazodone	16691	0	0	6666
Citalopram	118253	25667	14109	71627
Escitalopram	161654	11760	15979	93405
Fluoxetine	125589	19017	7201	64019
Paroxetine	7322	0	0	3452
Sertraline	150763	19890	13971	85329
Desvenlafaxine	34345	3095	0	14671
Duloxetine	109124	12551	12306	54937
Venlafaxine	100857	9955	9511	58417
Electroconvulsive Therapy	3234	0	0	2359
Psychotherapy	509303	50946	34713	227345

Table 10.2: Cohort size within each database

many statistically significant estimates to the left compared to SSRIs, SNRIs, ECT, and psychotherapy. The SSRIs vs SNRIs subplot does not display a marked preference for either treatment class. All four medication classes are superior to ECT and to psychotherapy, while ECT vs psychotherapy was not available for comparison.

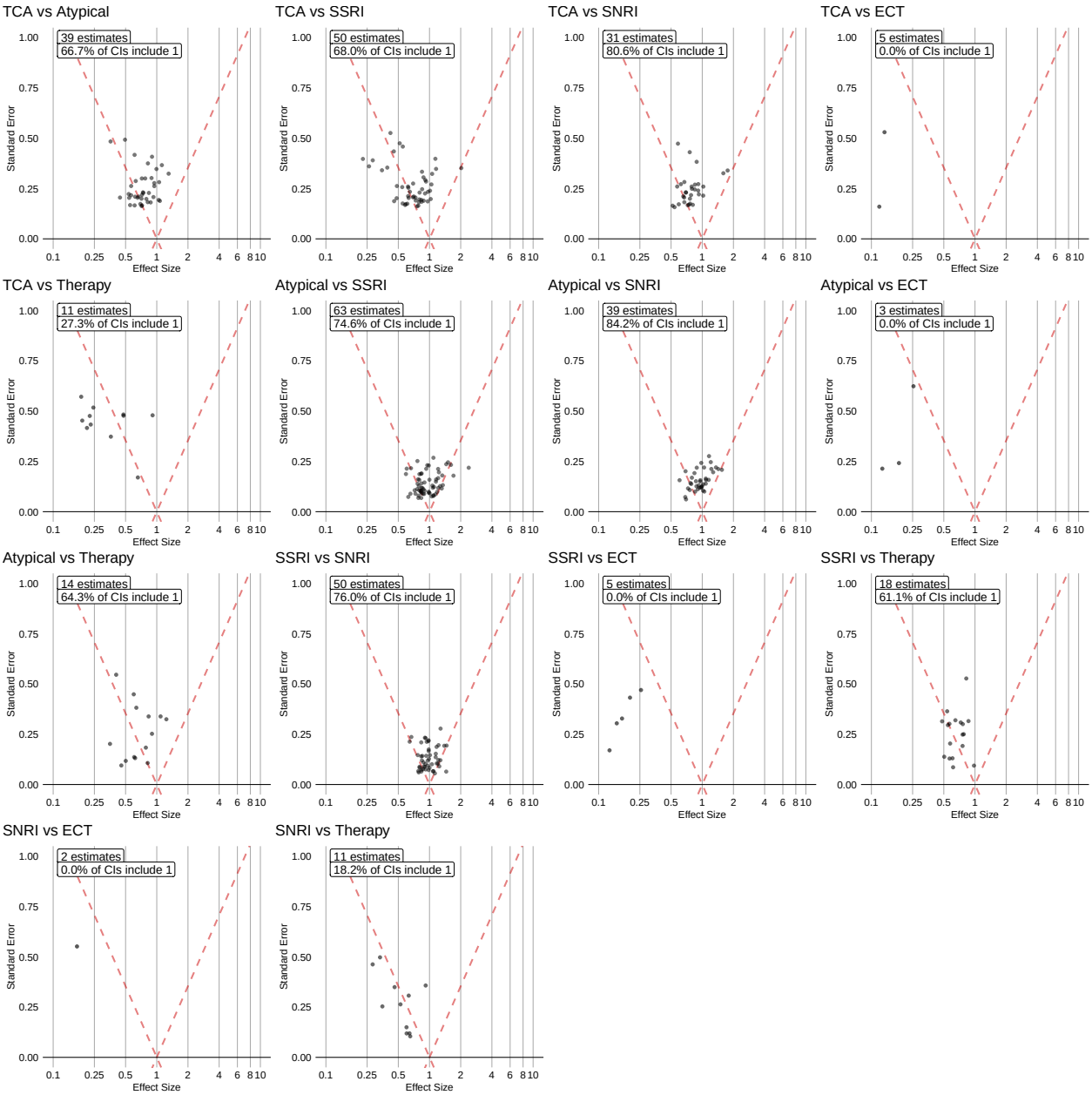


Figure 10.1: Class-class comparisons of hazard ratio estimates. Each individual point represents the comparison of one treatment from the first class to one treatment from the second class in one database. Effect sizes greater than 1 represent more suicide and suicidal ideation outcomes in the first treatment class, and vice versa. Points above the dashed line are not statistically significant, while points below the dashed line are statistically significant.

We delve into the individual treatment-treatment comparisons in Figure 10.2, which represents for each comparison the net number of databases that have a statistically significant result favoring the target or the comparator treatment. The darker shade of red, the more

databases have significantly fewer suicide and suicidal ideation in the target treatment, and vice versa for shades of blue. Each comparison is performed independently, so for example doxepin-nortriptyline has a different result than nortriptyline-doxepin due to slightly different constructed propensity scores. Among all comparisons, only a single one, amitriptyline-mirtazapine and mirtazapine-amitriptyline, have all four databases demonstrate statistically significant results, favoring amitriptyline over mirtazapine. Amitriptyline, nortriptyline, and bupropion have the most number of favorable statistically significant comparisons to other treatments, while mirtazapine, fluoxetine, electroconvulsive therapy, and psychotherapy have the most number of unfavorable comparisons. We see that the results favoring TCAs seem in Figure 1 come from amitriptyline and nortriptyline, and less so doxepin. Furthermore, the results favoring atypical antidepressants over SSRIs and SNRIs come almost entirely from bupropion. The mixed results comparing SSRIs to SNRIs come from unfavorable comparisons involving fluoxetine (for SSRIs) and venlafaxine (for SNRIs).

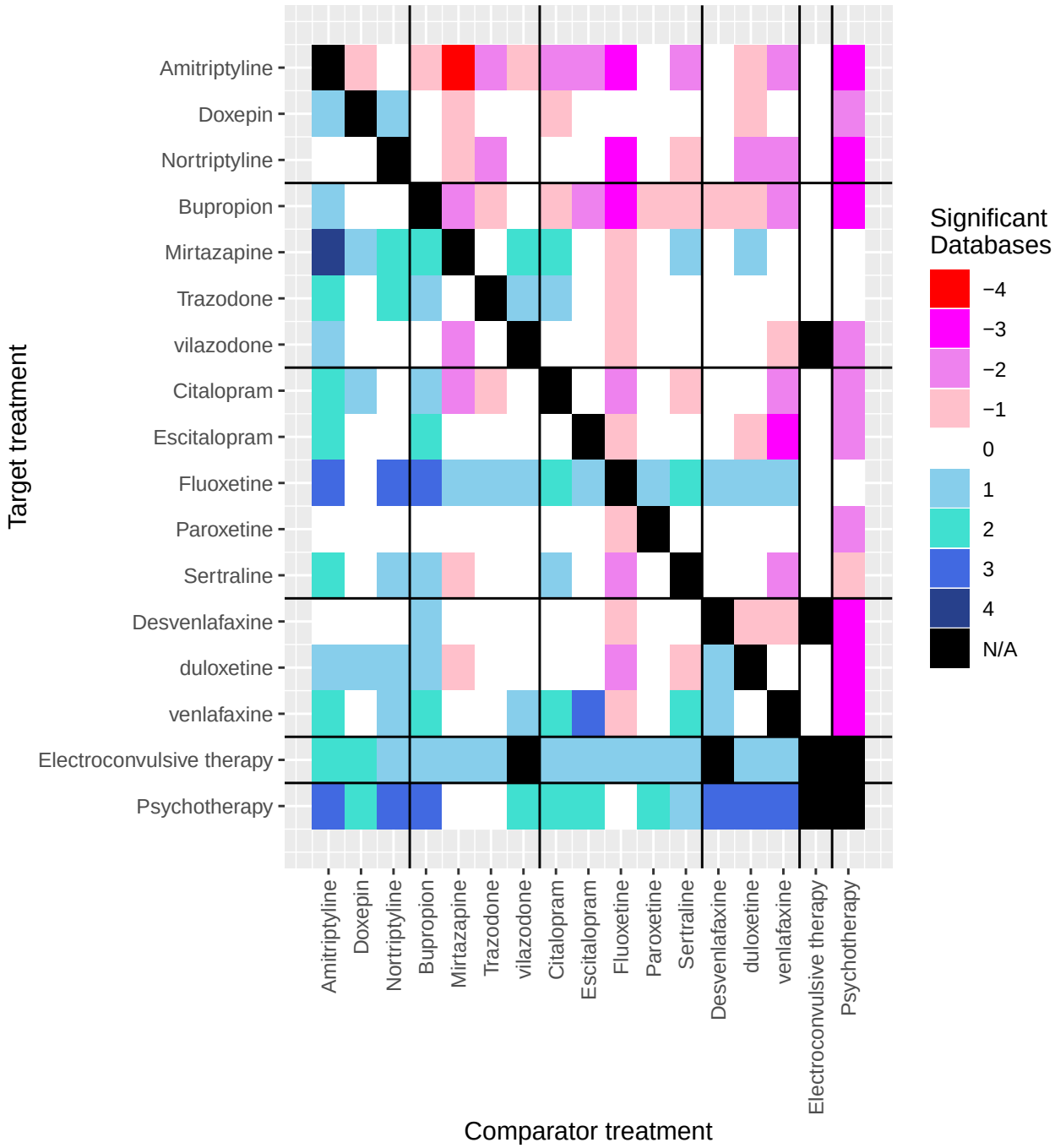


Figure 10.2: Comparison of individual treatments across four databases. Each cell displays the number of databases for each treatment-treatment comparison giving a significant hazard ratio estimate that is greater than (positive) or less than (negative) 1. For example, there is a net of 1 (out of 4) databases that have a statistically significant HR estimate less than 1 for the comparison of amitriptyline to doxepin, a result that favors amitriptyline.

We utilize propensity scores to control for measured confounding and reduce covariate imbalance between compared treatment groups. Often an after-stratification standardized

difference threshold of 0.1 is used to evaluate successful propensity score adjustment. Figure 10.3 displays the maximum after-stratification standardized mean difference (SMD) among all covariates for all pairwise comparisons of treatments in the CCAE database. Overall, most analyses have successful propensity score adjustment, but five treatments have more than four comparisons with unbalanced covariates: amitriptyline, nortriptyline, vilazodone, paroxetine, and psychotherapy. No data were available for electroconvulsive therapy. Amitriptyline, paroxetine, and psychotherapy have the most maximum SMD greater than 0.3, while most of the unbalanced comparisons for nortriptyline and vilazodone have maximum SMD between 0.1-0.3. Interestingly, these difficult-to-balance treatments include three (amitriptyline, nortriptyline, and psychotherapy) of the treatments that also have strong evidence with regards to the outcome, suicide and suicidal ideation (Figure 10.2).

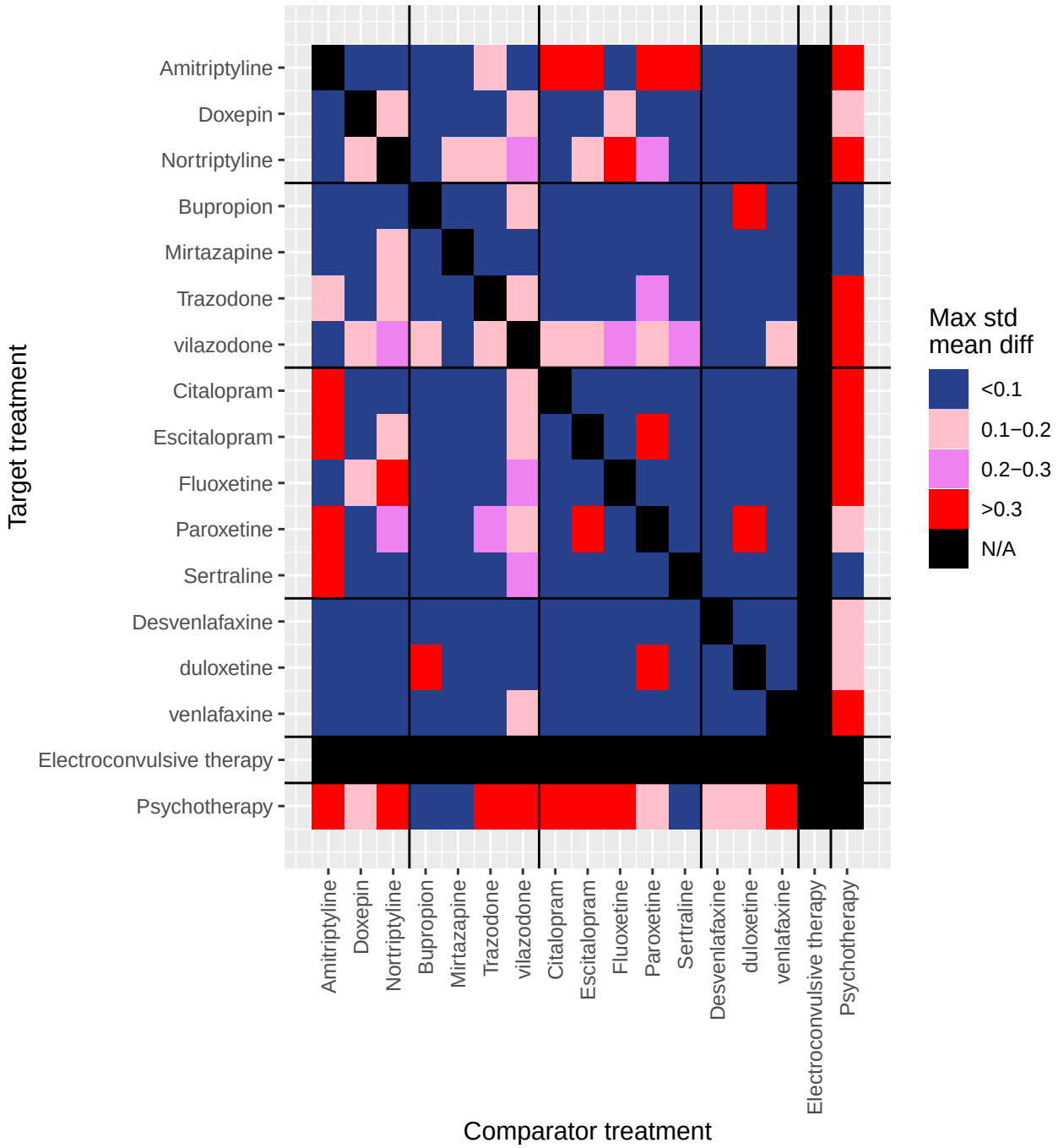
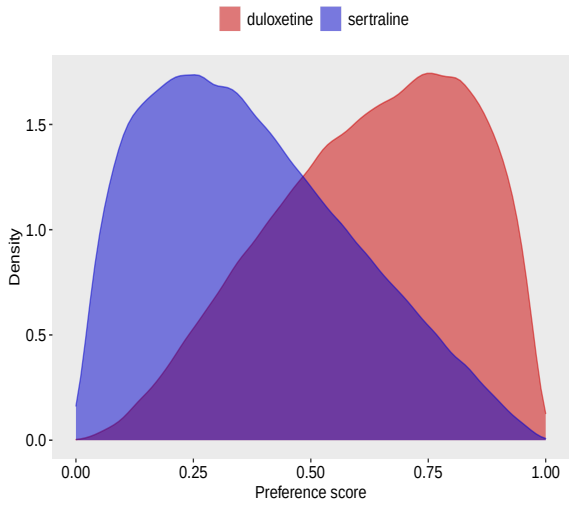


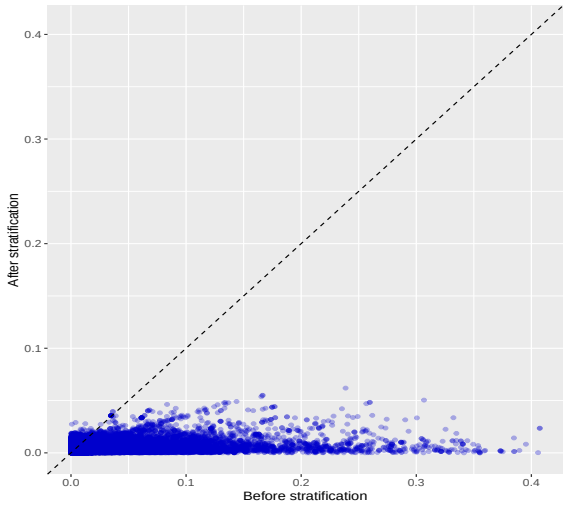
Figure 10.3: Maximum post-PS matching absolute standardized mean difference (SMD) among all covariates in treatment-treatment comparisons in the CCAE database. Displays whether this maximum SMD is greater than or less than 1.

On further inspection, the large maximum SMDs for some treatment comparisons are due to increasingly disjoint PS distributions that are difficult to stratify. Figure 10.4a shows the PS distribution for the duloxetine-sertraline comparison in the CCAE database. There

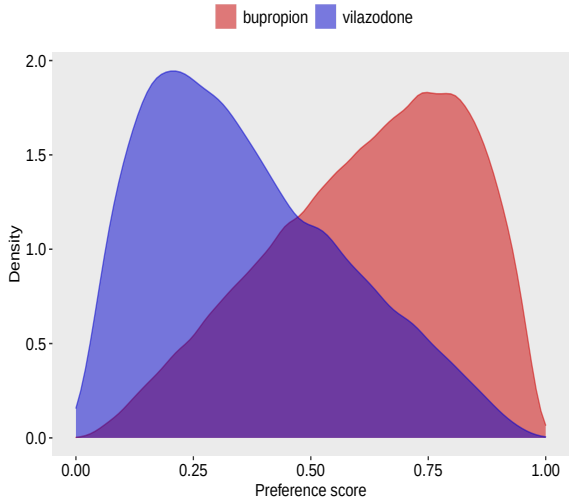
is both reasonable separation between the cohorts and reasonable overlap, allowing for successful stratification and covariate balancing (Figure 10.4b). The bupropion-vilazodone PS distribution has slightly less overlap (Figure 10.4c) and more separation between cohorts. Only a single covariate has slightly greater than 0.1 SMD (Figure 10.4d), and although the overall covariate balance is worse than in duloxetine-sertraline, it is still reasonable. In contrast, trazodone and paroxetine have little overlap between their PS distributions (Figure 10.5a), and although some of the most extreme before-stratification SMDs are reduced, almost as many covariates seem to have improved as worsened covariate balance (Figure 10.5b). Finally, the amitriptyline-citalopram comparison has complete separation of PS distributions (Figure 10.5c), allowing for no stratification and no progress on covariate balance (Figure 10.5d).



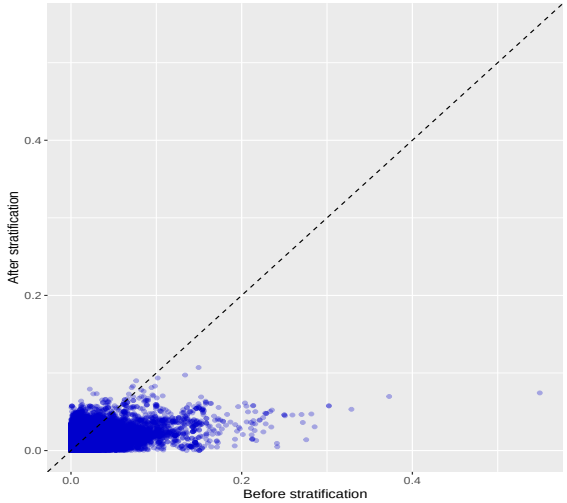
(a) PS plot for duloxetine-sertraline



(b) Balance plot for duloxetine-sertraline

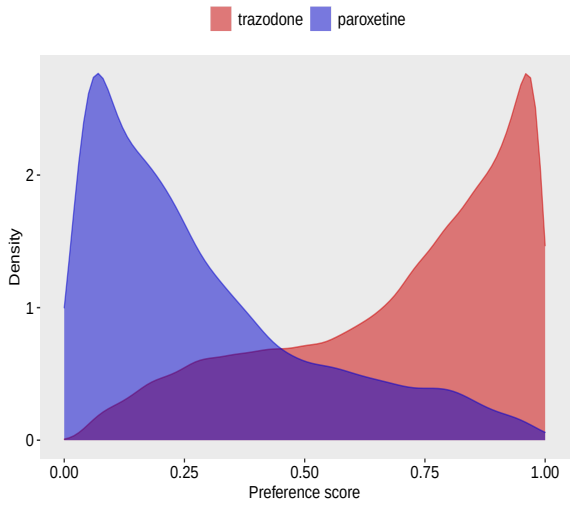


(c) PS plot for bupropion-vilazodone

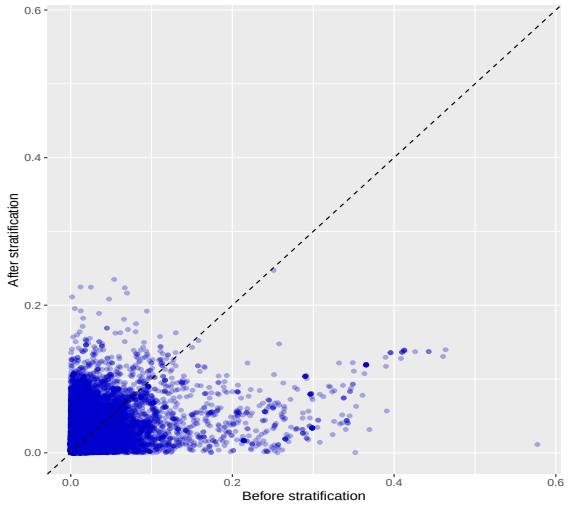


(d) Balance plot for bupropion-vilazodone

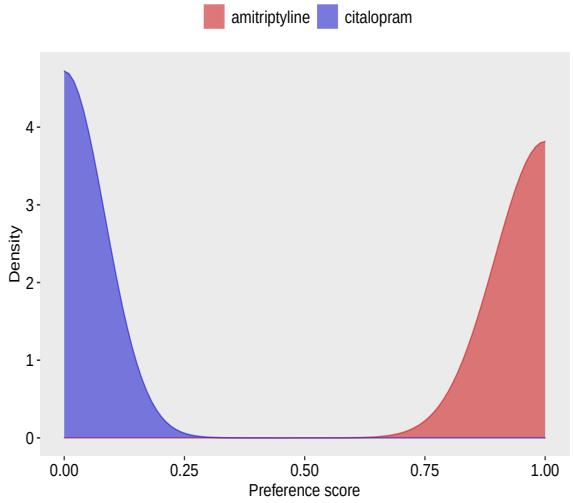
Figure 10.4: PS distributions and covariate balance plots for duloxetine-sertraline and bupropion-vilazodone comparisons in the CCAE database. Each covariate balance plot point represents a single covariate's before and after stratification standardized mean difference. Points below the dotted line have improved covariate balance.



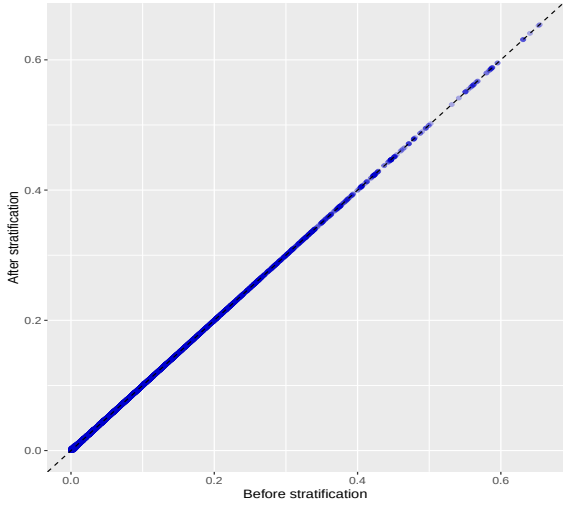
(a) PS plot for trazodone-paroxetine



(b) Balance plot for trazodone-paroxetine



(c) PS plot for amitriptyline-citalopram



(d) Balance plot for amitriptyline-citalopram

Figure 10.5: PS distributions and covariate balance plots for trazodone-paroxetine and amitriptyline-citalopram comparisons in the CCAE database. Each covariate balance plot point represents a single covariate's before and after stratification standardized mean difference. Points below the dotted line have improved covariate balance.

10.4 Discussion

Randomized trials remain the gold standard of evidence in clinical medicine, but they have areas of deficiency in generating evidence for real-world applications. Controlled studies can be prohibitively expensive and have insufficiently large sample sizes to detect adverse events or differential effects in comparative effectiveness studies. One review of clinical trials

studying pharmacological effects on self-harm finds that relevant studies are too few and too small to generate firm conclusions, and larger trials are needed [216]. Retrospective observational studies fill in this void for evidence by utilizing the staggering amount of information available in longitudinal databases on millions of patients [3]. In this paper, we embrace a new paradigm for generating observational evidence by conducting a comprehensive all-by-all comparison of 17 different antidepressant treatments across 4 databases [229]. Each individual clinical hypothesis is studied using the same consistent methodology and study design, including cohort definitions, statistical analysis parameters, and decisions regarding presentation of results.

We present a novel class-by-class comparison of all major antidepressant medication classes. In our results, tricyclic antidepressants compare favorably to other treatment classes. However, these favorable TCA comparisons are mostly from amitriptyline and nortriptyline (Figure 10.2) that also have disjoint PS distributions and high after-stratification SMD in comparison to several other treatments, indicating that these comparisons are ineffective. Atypical depressants are the next most favorable class, with favorable signals coming mostly from bupropion and some unfavorable signals coming from mirtazapine. SSRIs and SNRIs are two of the more popular drug classes, and existing research using PS analysis shows both no differential effect [222] and results favoring SNRIs [224]. We find mixed results comparing these two classes, with most statistically significant signals coming from unfavorable comparisons involving fluoxetine (an SSRI) and venlafaxine (an SNRI) (Figure 10.2).

In our results, both ECT and psychotherapy lead to more suicide and suicidal ideation outcomes compared to all medication classes (Figure 10.1). However, we employed a new-user cohort design in which patients are only on a single treatment. Psychotherapy is recommended in conjunction with pharmacological treatment [215], and not as monotherapy, which our study partially captures. We do not believe that psychotherapy is ineffective for preventing suicide and suicidal ideations; our results mainly suggest against psychotherapy-only treatment. Although we employ PS stratification with large-scale PS models [19] that should remove measured confounding, there may still be residual confounding present in our study. In particular, our ECT results suggest there may be channeling bias in our study,

as ECT is reserved for serious patients unresponsive to other therapy [231], and our results show strong evidence against ECT compared to other therapies.

The lack of direct comparative effectiveness research regarding suicide and suicidal ideation presents challenges in prescribing one medication over another. Rubino et al. [223] reports venlafaxine has higher suicide risk compared with citalopram and fluoxetine, a result we see with citalopram, but not with fluoxetine (we observe fluoxetine has a higher risk). However, that study relies on a small outcome model with 24 covariates to avoid saturation of the statistical model, while we employ thousands of covariates along with regularized regression [19] for model selection. Jick et al. [221] compares amitriptyline, fluoxetine, and paroxetine separately to dothiepin, and not to each other, and find no differences in suicidal behavior. By doing an exhaustive pairwise comparison of antidepressant drugs, we are able to see that amitriptyline, fluoxetine, and paroxetine in fact compare very differently to other treatments. Amitriptyline is our most favorable compared treatment, fluoxetine is our least favorable, and paroxetine is squarely in the middle.

All-by-all comparisons of treatments within a clinical domain are a new way of conducting observational research and utilizing the full scale of data available in longitudinal databases [229]. Using a consistent methodology across hundreds of individual hypotheses and four databases, our study reveals the benefits of bupropion and the risk of fluoxetine with regards to suicide and suicidal behavior. In depth analysis of PS distributions and covariate balance are able to reveal treatment-treatment combinations that are incomparable and unsuitable for comparative effectiveness analysis. Our study results and methodology can inform the treatment decision making process among multiple medications for clinical depression.

CHAPTER 11

Comparative Effectiveness of Branded Versus Generic Versions of Antihypertensive, Lipid-Lowering and Hypoglycemic Substances

11.1 Introduction

Generic medications offer potential for substantial health care cost savings compared to their branded drug counterparts [232, 233, 234], but their adoption is hindered by doubts among physicians and patients regarding their efficacy and safety [235, 236, 237, 238, 239]. Some of the concerns arise from a lack of knowledge or misinformation regarding the concept of bioequivalence and/or from marketing efforts of branded drug manufacturers. Frequently, the argument is that while pharmaceutical companies need to conduct extensive clinical trials to bring an innovator branded drug to market, they are required only to demonstrate biologic equivalence for new generic drugs, and not equivalence in clinical outcomes [240]. However, randomized controlled trials of generic drugs vs originators are rarely feasible or required by regulators, unless bioequivalence cannot be shown with pharmacokinetic studies, e.g. for drugs not administered systemically. As a result, randomized trials comparing generic to branded drugs feature relatively small sample sizes that are sufficient to show bioequivalence, but are by their very nature not powered to find significant differences in clinical efficacy [241, 242, 243].

In the absence of randomized trials, retrospective data are a crucial resource to collect clinical data on generic drugs [244]. Observational studies conducted using longitudinal health databases that contain millions of patient records could discern clinically meaningful

differences between branded and generic drugs. However, these studies require careful control for potential confounding that plagues all observational research. Previous studies in this field have highlighted the need to control not only for patient medical history, but also additional factors such as socioeconomic status [245] and medication adherence [246].

There has been extensive observational research on antiepileptic drugs, including several narrow therapeutic index drugs that are particularly concerning for introducing generic alternatives [247, 248, 249, 250]. There are relatively fewer studies on generic medications for chronic metabolic diseases such as hypertension or heart failure, hyperlipidemia, and diabetes mellitus that offer significant opportunities for cost savings owing to their widespread use [251]. In this study, we compared death and cardiovascular outcomes between branded and generic formulations of 17 antihypertensive, cholesterol-lowering, and oral hypoglycemic drugs using national pharmacy and hospitalization data representing nearly all insured persons in Austria.

11.2 Methods

11.2.1 Study Population and Data

We analyzed all filled prescriptions that were submitted for reimbursement to 13 Austrian social security institutions, including all nine provincial sickness funds as well as four nationwide institutions (federal employees, farmers, independent business owners, and railroad and mining employees). In total, these institutions cover 98.5% of all insured persons in Austria. Prescription data were available for 9,413,620 insured persons from 2007 to 2012, and each record contained a pseudonymized unique patient identifier, volume (number of packages), package size (number of units per package), strength (dose per unit), the pharmacy article identifier of the dispensed drug, and patient co-payment waiver status (yes or no). In addition, linked through the pseudonymized patient identifier we obtained patient birth months, sex, date of deregistration from the social security institution (if applicable), date of death (if applicable), and all hospitalizations in the study period including admission date, length

of stay, and discharge diagnoses. These data have been utilized and described in previous studies comparing generic and branded drug costs [234] and investigating double medication rates [252].

11.2.2 Investigated Drug Classes

For each of the investigated chronic diseases (hypertension or heart failure, hyperlipidemia, and diabetes mellitus) we compiled a list of therapeutic substances in terms of the World Health Organization (WHO)'s fifth-level (seven digit) Anatomical Therapeutic Chemical (ATC) code [253] as previously reported [234]. From this list, we selected the 17 substances with the highest potential monetary savings that could be achieved by generic substitution [234], and for which generic and branded versions were simultaneously available in the same combination of package size and strength. This list comprised twelve single substances or substance combinations for hypertension or heart failure treatment (metoprolol, bisoprolol, nebivolol, carvedilol, amlodipine, enalapril, lisinopril, ramipril, enalapril and diuretics, lisinopril and diuretics, ramipril and diuretics, losartan and diuretics), two lipid-lowering substances (simvastatin, fluvastatin), and three oral hypoglycemic substances (metformin, gliclazide, repaglinide). A database supplied by the Austrian Agency for Health and Food Safety (AGES) provided information on the branded versus generic status of each pharmacy article. For each drug class we defined a start date of the study period as the date at which a generic version of the drug was first reimbursed in our database. Thus, data from branded medicines were only considered starting with the month when a generic was also available. Only specific package size/strength combinations for which both generic and branded products were available were considered.

11.2.3 Patient Inclusion

Patients were included when they filled a new prescription of any of the investigated substances. This index date was used to determine subsequent study outcomes and ascertain preceding covariates. Only patients who were at least 18 years old at the index date were

considered in the analysis. A wash-out period of at least 180 days with no prescription of the substance of interest was required to define a "new prescription" and to harvest covariates, and therefore, patients were excluded if the wash-out period was not fully covered by our database (Fig. 11.1). If a patient was simultaneously eligible for inclusion for multiple substances, we randomly selected one substance for that patient and hence included the patient only once in our study.

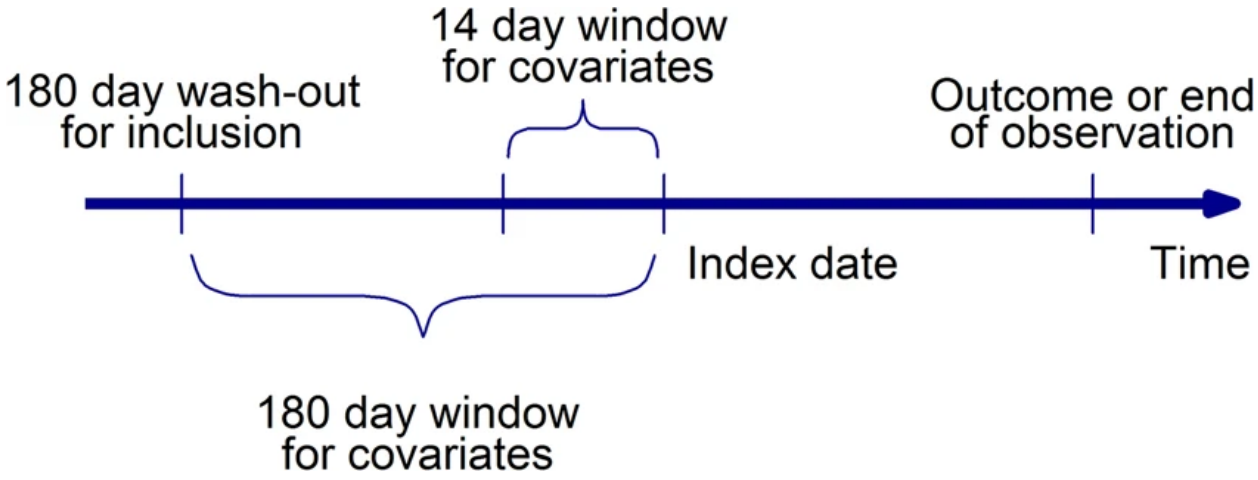


Figure 11.1: Data harvesting for the study. The inclusion date was the date of the first prescription or hospital admission of a patient in the data base. The index date was the date of first prescription of a study medication after a wash-out period of at least 6 months with no prescriptions of medicines of the same ATC code. All patients with an index date occurring at least 6 months after the inclusion date were included. Covariates (hospital discharge diagnoses, prescriptions, hospitalization days) were harvested during the 6 months preceding the index date. The patients were followed-up in the data base until an outcome event (death, MACCE), until deregistration from the insurer or until 31 December 2012, whichever occurred earlier.

11.2.4 Ascertainment of Study Outcomes

As primary outcomes, we considered time to all-cause death and time to major cardiac or cerebrovascular events (MACCE). We defined MACCE as any myocardial infarction, stroke, transient ischemic attack, or all-cause death. Ascertainment of MACCE was based on the following ICD10 codes recorded in hospital discharge diagnoses: I20, I21, I60, I61, I62, I63, I64, I65, I66, I69, G45. The starting point of these analyses was the time of index prescription. If patients had no events recorded in the database, we censored them at the date of their last observation date or at the date of deregistration from insurance,

whichever occurred first. Treatment discontinuation was defined as no refill of initial type of prescription (branded or generic) within 180 days, conditional on survival and follow-up of at least 180 days.

11.2.5 Ascertainment of Covariates

The following covariates were harvested at the date of index prescription: age at prescription, sex, insurer, copayment waiver status, specialty of prescriber (general practitioner, internal medicine specialist, hospital, other), and year of prescription. Within the time period of 14 days preceding the index prescription, we extracted binary variables indicating whether any hospitalization ended in that period, whether a "long" hospitalization (duration of more than 14 days) ended in that period, an indicator for each discharge diagnosis recorded, and indicators for each drug class prescribed (ATC2 level). The same set of variables was also extracted for the time period of 180 to 14 days preceding the index prescription.

11.2.6 Statistical Analyses

A high-dimensional propensity score as outlined in Schneeweiss et al. [18] describing the probability of receiving branded versus generic medication was fitted to calculate inverse probability of received treatment weights (IPTW). Specifically, we included the main descriptors (age, a quadratic age term $(\text{age}/100)^2$, sex, any hospitalization in 180 day and 14 day windows, any discharge diagnosis indicating myocardial infarction or cerebrovascular events), the 200 variables with the highest potential to correct for bias [18, 83], two-way interactions among the main descriptors, and interactions between these main descriptors and diagnoses and prescriptions. We applied the least angle shrinkage and selection operator (lasso) to regularize and perform model selection among the interaction terms [20]. We used IPTW to equalize differences in the characteristics between patients receiving branded drugs and patients receiving generic drugs as index prescription. We evaluated success of propensity score weighting by comparing standardized mean differences in all covariates before and after weighting.

Kaplan-Meier curves were used to describe time to death and time to MACCE. Ninety-five percent confidence intervals (CI) for incidence rates were computed using a Poisson distribution. Cox proportional hazards regression models were used to estimate unadjusted and adjusted hazard ratios and 95% confidence intervals (95%CI) for all-cause death and MACCE, and logistic regression was used to compute unadjusted and adjusted relative risks and 95%CI for treatment discontinuation.

For each substance, we estimated hazard ratios for mortality as well for MACCE with the following adjustment levels:

- unadjusted,
- minimally adjusted (adjusted for age, $(\text{age}/100)^2$, sex, and copayment waiver status),
- adjusted by an extended set of covariates (minimal adjustment set plus calendar year of index prescription, specialty of prescriber, previous hospitalizations, recent MI or cerebrovascular events, any diagnosis in group of endocrine, nutritional or metabolic diseases (ICD10 code E) or in group of diseases of circulatory system (ICD10 code I), and any previous use of antihypertensive, lipid-lowering or hypoglycemic drugs),
- fully adjusted by IPTW weighting (aHR). IPTW-adjusted models were also subgrouped by sex, by age ($\leq$ or $>$ 70 years), by any history of cardiovascular or diabetes disease (CVDD) as evidenced by previous diagnosis codes or relevant drug prescriptions, and by diabetes treatment status (no diabetes vs. oral hypoglycemic drugs prescribed but no insulin vs. insulin prescribed).

As a sensitivity analysis for potential unmeasured confounding we calculated E-values for point estimates and confidence limits according to VanderWeele and Ding [1]. E-values quantify the minimum strength of an association between a hypothetical unmeasured confounder and both treatment and outcome that could account for the observed treatment effect after controlling for measured covariates. We investigated time-dependency of treatment effect estimates by estimating adjusted hazard ratios during the first six months (censoring later

events) and during the period after six months (conditional on surviving six months). These analyses were accompanied by interaction testing.

To investigate the role of treatment discontinuation, three additional models were fitted for each outcome (IPTW adjusted):

- a landmark model, conditional on survival of 6 months, including treatment discontinuation before 6 months as covariate,
- a subgroup landmark model with landmark set at 6 months including only patients who continued treatment within 6 months from initial prescription,
- and a subgroup landmark model (6 months) including only patients who discontinued treatment within 6 months from initial prescription.

All hazard ratios were estimated separately for each substance and were then pooled across all substances of the same indication (antihypertensive drugs, lipid-lowering drugs, hypoglycemic drugs) using random-effects meta-analysis. If weighted models were estimated, then a robust covariance matrix was used. All models were stratified for package size/strength combination at initial prescription.

Data were analyzed using PostGreSQL [254] and R [255].

11.2.7 Ethics, Data Protection and Data Availability

The protocol of the study was created in compliance with the Guidelines for Good Pharmacoepidemiology Practices [256]. According to the Austrian Federal Act concerning the Protection of Personal Data ('Datenschutzgesetz', DSG) the study was exempted from the need to obtain informed consent from the participants as the research data base which was provided by the Main Association of the Austrian Social Security Institutions was already irreversibly pseudonymized and the identities of the participants could not be established. As this was a retrospective study, participation in the study did not alter any risks of the participants. The study protocol and the exempt from the need to obtain informed consent was

approved by the Ethics Committee of the Medical University of Vienna (ECS 1533/2013). The data that support the findings of this study are available from the Main Association of the Austrian Social Security Institutions but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Main Association of the Austrian Social Security Institutions.

11.3 Results

11.3.1 Patients

During the study period from 2007 to 2012, 986,149; 47,359; and 201,038 patients with index prescriptions for antihypertensive, lipid-lowering and hypoglycemic drugs were included, respectively, with follow-up totaling to 1,920,544; 93,952; and 383,460 patient years. Figure 11.2 shows patient counts for each evaluated substance, grouped by branded or generic medicines. Characteristics of patients at their first index prescription are displayed in Table 11.1. In general, patients receiving branded medicines were older, more often had recent (within past 14 days) or previous (within past 180 days) hospitalizations, more often had used antihypertensive, lipid-lowering and hypoglycemic drugs before and more often received their index prescriptions from hospitals compared to patients treated with generic medicines. For lipid-lowering and hypoglycemic drugs, we also observed that men received branded medicines more often than women.

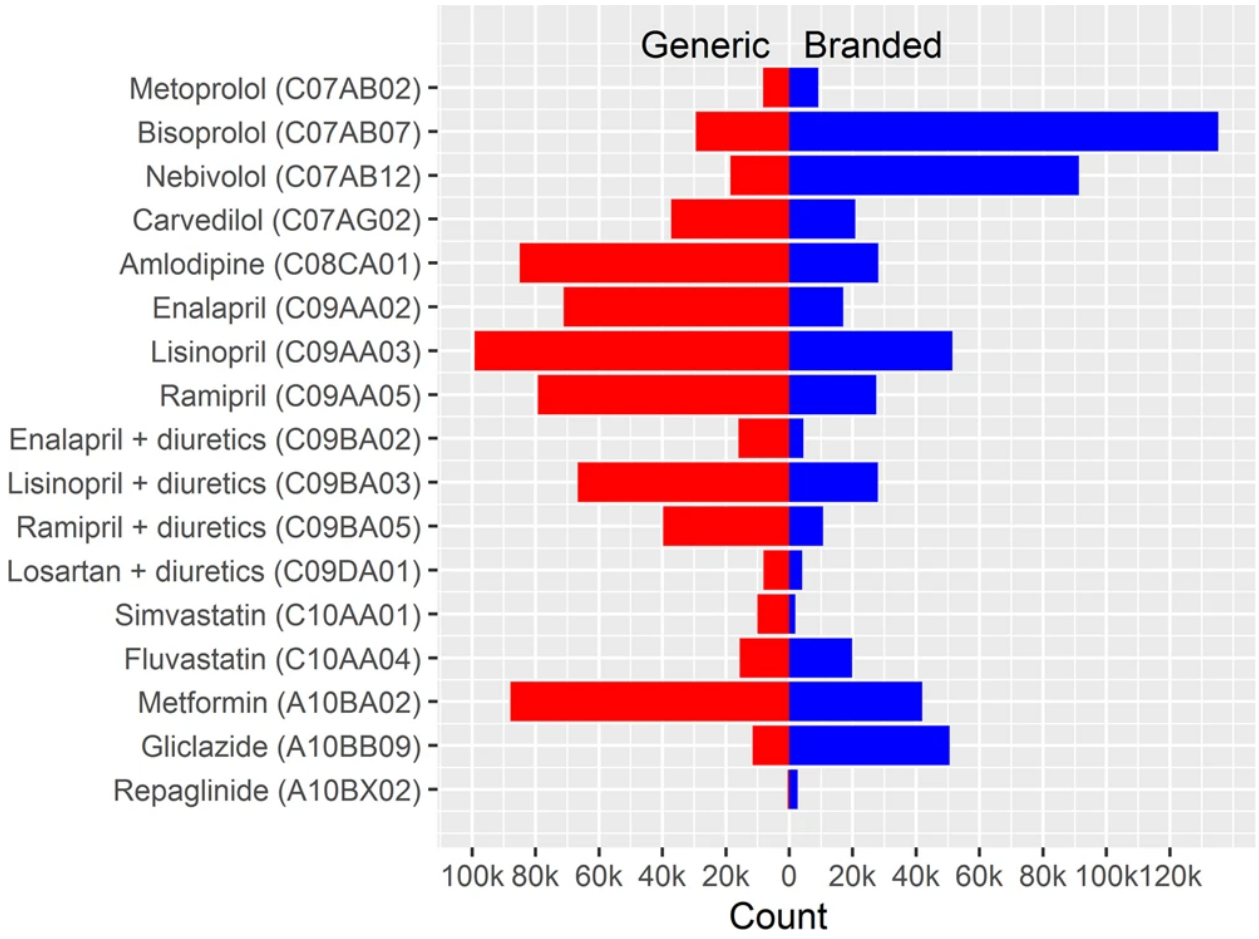


Figure 11.2: Patient counts (1k = 1,000) for each substance evaluated.

Propensity models achieved concordance indices in the range of 0.639 (enalapril, C09AA02) to 0.854 (losartan and diuretics, C09DA01). After IPTW weighting, maximum standardized mean differences across all high-dimensional propensity score covariates were below 10% for all substances, except for repaglinide, A10BX02 (17.3%) and bisoprolol, C07AB07 (12.4%, Supplementary Table 1). The means of the standardized mean differences across all covariates were always $< 3\%$, and were $< 1\%$ for 15 of the 17 studied substances.

11.3.2 Antihypertensives: Primary Time-to-Event Outcomes

Across all 12 antihypertensive substances, 53.8 (95% CI; 53.3, 54.3) deaths per 1000 patient-years were observed for branded medicines, while the corresponding figure was 30.2 (95% CI; 29.9, 30.5) for generic medicines. After IPTW adjustment, the estimated incidence

rates were 45.8 (95% CI; 45.5, 46.1) deaths per 1000 patient years for branded medicines and 40.6 (95% CI; 40.4, 40.9) for generic medicines. Crude cumulative five-year survival rates in branded and generics users were 77.8% (95% CI; 77.3%, 78.4%) and 85.9% (95% CI; 85.5%, 86.2%), respectively, and the corresponding IPTW-adjusted rates were 79.8% (95% CI; 79.4%, 80.1%) and 82.7% (95% CI; 82.4%, 83.0%) (Fig. 11.3). The unadjusted pooled branded vs. generic hazard ratio (HR) of 1.75 (95% CI; 1.56, 1.98) was attenuated after IPTW adjustment to an aHR of 1.15 (95% CI; 1.06, 1.26), favoring generics. Table 11.2 illustrates the aHR resulting from different adjustments, demonstrating a continuously decreasing aHR with increasing covariate adjustment. Table 11.3 compares fully adjusted aHRs across different substances. Interestingly, while results for most substances favored of generics, the direction of association was reversed for bisoprolol (C07AB07) and nebivolol (C07AB12). Among all subgroup analyses conducted, we only found significant treatment effect modification with history of CVDD (interaction $p < 0.001$). In patients without CVDD history, the aHR was 1.47 (95%CI; 1.31, 1.64), while being only 1.10 (95%CI; 1.01, 1.20) in patients with CVDD history.

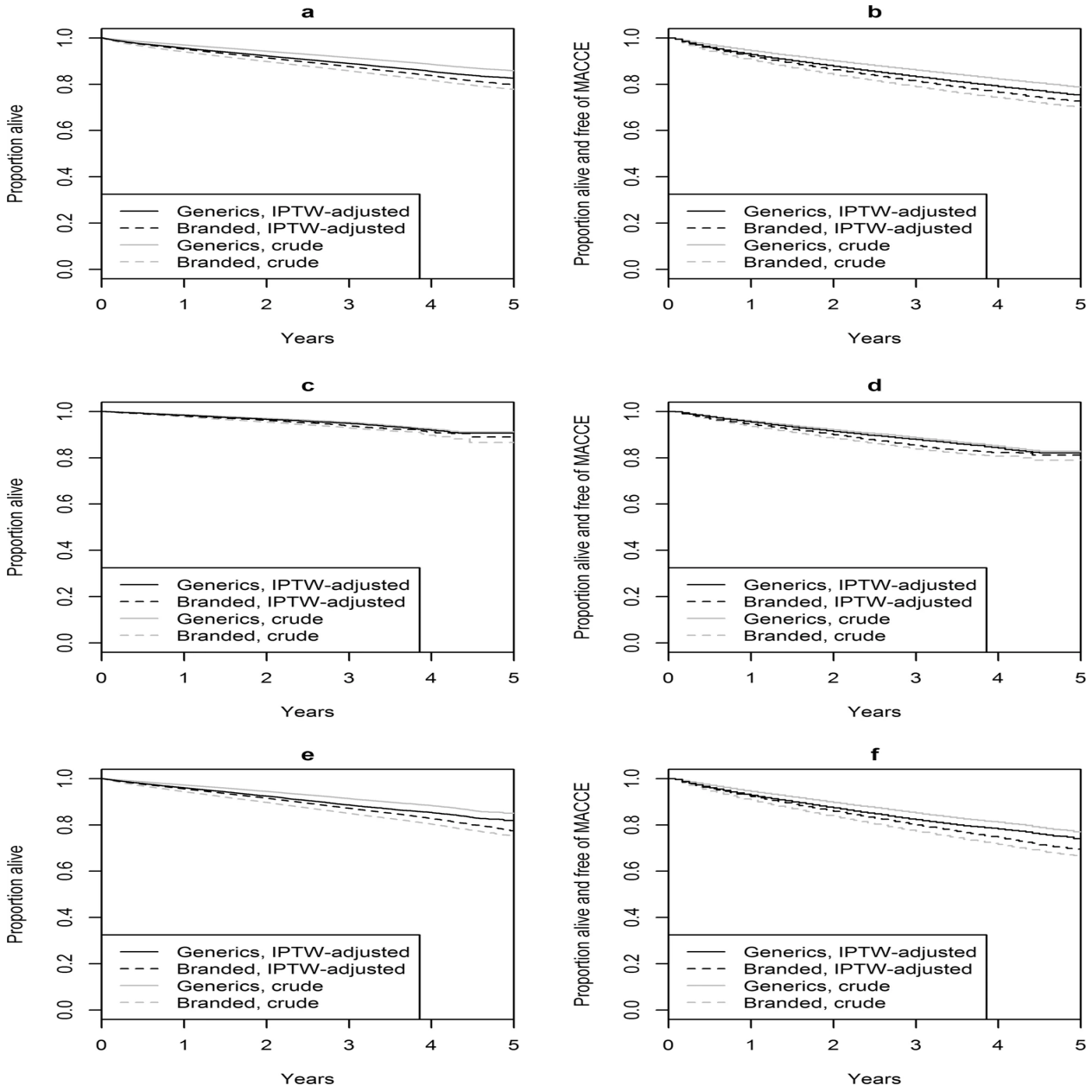


Figure 11.3: Survival curves and curves of cumulative MACCE-free survival. (a) Overall survival for patients with index prescriptions for antihypertensive drugs. (b) MACCE-free survival for patients with index prescriptions for antihypertensive drugs. (c) Overall survival for patients with index prescriptions for lipid-lowering drugs. (d) MACCE-free survival for patients with index prescriptions for lipid-lowering drugs. (e) Overall survival for patients with index prescriptions for hypoglycemic drugs. (f) MACCE-free survival for patients with index prescriptions for hypoglycemic drugs.

In patients receiving branded medicines, we observed a rate of 83.6 (95% CI; 82.9, 84.2) major cardiac and cerebrovascular events (MACCE) per 1000 patient-years, compared to 51.3 (95% CI; 50.9, 51.8) in patients using generic medicines. The IPTW-adjusted MACCE

incidence rates were 72.3 (95% CI; 72.0, 73.3) and 64.1 (95% CI; 63.8, 65.0). After IPTW adjustment, the hazard ratio for MACCE was 1.13 (95% CI; 1.07, 1.20). While most individual substances favored generic drugs, an opposite effect estimate was observed for bisoprolol (C07AB07) with aHR 0.91 (95% CI; 0.85, 0.98), and no significant benefit was observed for nebivolol (C07AB12), enalapril and diuretics (C09BA02) and losartan and diuretics (C09DA01) (Table 11.4). Nominally significant treatment effect modification was detected for age (interaction p-value 0.042), with the estimated treatment effects being stronger in patients aged 70 years or younger, aHR 1.20 (95% CI; 1.13, 1.28) than in relatively older patients, aHR 1.10 (95% CI; 1.04, 1.17), and for history of CVDD (interaction $p < 0.001$) (Table 11.4).

11.3.3 Antihypertensives: Treatment Discontinuation

In 26.7% of all index prescriptions of branded medicines and also in 26.7% of all index prescriptions of generic medicines, no refill was observed within the first six months. However, after IPTW adjustment, the adjusted relative risk of treatment discontinuation was 1.23 (95% CI; 1.05, 1.44) in patients originally receiving branded medicines than in patients receiving generic medicines. In the landmark analysis including only patients who survived and were observed for at least six months and who were still using the originally prescribed medication, the aHR for mortality was very similar to the main analysis, aHR = 1.17 (95% CI; 1.03, 1.34), and not significantly different from the aHR computed in patients who discontinued treatment within six months from index prescription, aHR = 1.12 (95% CI; 1.04, 1.21) (p for interaction = 0.558). In the landmark analysis that included treatment discontinuation up to six months as a covariate, the aHR was virtually unchanged, aHR = 1.16 (95% CI; 1.04, 1.26).

11.3.4 Lipid-Lowering Drugs: Primary Time-to-Event Outcomes

Patients using branded or generic lipid-lowering drugs experienced 24.4 (95% CI; 23.0, 25.9) and 16.0 (95% CI; 15.0, 17.2) deaths per 1000 patient-years, respectively. The IPTW-

adjusted incidence rates per 1000 patient-years were 20.8 (95% CI; 19.9, 21.8) for branded drugs and 17.8 (95% CI; 16.9, 18.6) for generics. Cumulative five year survival rates were 86.6% (95% CI; 82.8%, 90.6%) and 91.1% (95% CI; 89.9%, 92.4%) in these two groups, and corresponded to adjusted survival rates of 89.0% (95% CI; 87.4%, 90.6%) and 90.6% (95% CI; 89.5%, 91.8%), respectively (Fig. 11.3).

The unadjusted pooled hazard ratio for mortality was 1.69 (95% CI; 1.09, 2.63), which was no longer significant after IPTW weighting, 1.13 (95% CI; 0.86, 1.47). For both individual substances, results suggested a lower hazard for generic medicines, however, results did not reach statistical significance for simvastatin (C10AA01). There was no significant effect modification by history of CVDD (p = 0.07) or time period (p = 0.35).

Branded and generic lipid-lowering drug users exhibited incidence rates for MACCE of 59.7 (95% CI; 57.4, 62.1) and 40.9 (95% CI; 39.1, 42.7) events per 1000 patient-years, respectively, which changed to 53.1 (95% CI; 51.6, 54.1) and 44.4 (95% CI; 43.0, 45.3) after IPTW adjustment. The pooled IPTW-adjusted hazard ratio was 1.20 (95% CI; 1.05, 1.38), and was more pronounced and significant for fluvastatin (C10AA04), while being smaller and non-significant in simvastatin (C10AA01). In subgroup analyses, we again observed a larger treatment effect estimate in patients with no previous history of CVDD, pooled aHR = 1.60 (95% CI; 1.19, 2.14), compared to patients with CVDD history, pooled aHR = 1.16 (95% CI; 1.03, 1.32), but interaction analyses failed to reach statistical significance (p = 0.052, Table 11.5). Similarly, there was no clear evidence for a time-dependent treatment effect (aHR for first six months, 1.37; aHR after six months, 1.10; interaction p = 0.151).

11.3.5 Lipid-Lowering Drugs: Treatment Discontinuation

Treatment discontinuation rates were significantly higher in branded medicines with simvastatin (C10AA01), 43.4% vs. 27.6%, adjusted relative risk 1.80 (95% CI; 1.63, 1.99). However, discontinuation rates were virtually equal for fluvastatin (C10AA04), 31.5% vs. 32.2%, adjusted relative risk 1.03 (95% CI; 0.99, 1.08). A pooled relative risk estimate of 1.36 (95% CI; 0.79, 2.36) resulted for lipid-lowering drugs. Pooled adjusted hazard ratios for

mortality from landmark analyses conditional on treatment continuation or discontinuation at six months were similar (interaction p-value = 0.75). Furthermore, including treatment discontinuation status at six month as covariate in a landmark analysis did not lead to a significant change in the overall result, aHR = 1.08 (95% CI; 0.80, 1.45).

11.3.6 Hypoglycemic Drugs: Primary Time-to-Event Outcomes

Incidence rates of mortality in patients using branded or generic hypoglycemic drugs were 55.5 (95% CI; 54.5, 56.6) and 29.8 (95% CI; 29.0, 30.6) events per 1000 patient-years, respectively. The corresponding IPTW-adjusted numbers were 45.9 (95% CI; 45.5, 46.9) and 40.3 (95% CI; 39.6, 40.9). Cumulative five-year survival rates were 75.1% (95% CI; 74.1%, 76.2%) and 85.0% (84.1%, 85.9%), respectively, and corresponded to IPTW adjusted survival rates of 77.4% (95% CI; 76.6%, 78.3%) and 81.9% (95% CI; 81.2%, 82.6%) (Fig. 11.3). The crude pooled hazard ratio for mortality of 1.43 (95% CI; 1.37, 1.49) reduced after IPTW adjustment to 1.09 (95% CI; 0.93, 1.28) (Table 11.2). A significantly lower mortality hazard for generics was observed for metformin (A10BA02) only, with aHR 1.21 (95% CI; 1.15, 1.27) (Table 11.3). Interaction tests did not reveal any significant differences between subgroups, nor between time periods (Table 11.6).

The incidence rates of MACCE for branded and generic hypoglycemic drug users were 88.5 (95% CI; 87.2, 89.9) and 54.2 (95% CI; 53.1, 55.3) events per 1000 patient-years, respectively. After IPTW adjustment, the corresponding MACCE incidence rate were 76.0 (95% CI; 75.1, 76.7) and 66.8 (95% CI; 66.0, 67.4). The IPTW-adjusted hazard ratio of MACCE confirmed a small but significant difference in favor of generics, aHR = 1.11 (95% CI; 1.03, 1.20) (Table 11.2), which was also seen in separate analyses of metformin (A10BA02) and gliclazide (A10BB09), while no difference was found for repaglinide (A10BX02) (Table 11.3). IPTW-adjusted hazard ratios in subgroups were very similar, and no differences in treatment effect could be confirmed by interaction tests. The aHR of MACCE did not change over time (Table 11.6).

11.3.7 Hypoglycemic Drugs: Treatment Discontinuation

Hypoglycemic treatment discontinuation rates at six months were 26.2% in branded users and 30.0% in generics users, and after IPTW-adjustments, there was no difference in the risk for discontinuation, with adjusted relative risk 1.02 (95% CI; 0.97, 1.07). In landmark analyses including only patients on treatment at six months, there was a clear benefit for metformin (A10BA02) generics users with respect to mortality, aHR = 1.27 (95% CI; 1.19, 1.35), but overall, the results pointed towards equivalence but with a wide confidence interval, pooled aHR = 1.01 (95% CI; 0.77, 1.34). In patients discontinuing their initial treatment, there was evidence for a small overall difference favoring generics, pooled aHR = 1.13 (95% CI; 1.04, 1.22). If the landmark analyses included treatment discontinuation as covariate, the pooled aHR was unchanged compared to the landmark analysis without further adjustment for treatment discontinuation, pooled aHR = 1.03 (95% CI; 0.83, 1.29).

11.4 Discussion

We compared death and the incidence of MACCE for 17 branded versus generic versions of several medications commonly prescribed for chronic metabolic illnesses (hypertension or heart failure, hyperlipidemia, diabetes mellitus) within a national dataset representing nearly all insured persons in Austria from 2007 to 2012. Drawing from national hospitalization and pharmaceutical prescription fill data, we found a small but clear advantage for generic drugs over their branded counterparts for most of the studied substances. This generic advantage was robust across various levels of covariate adjustment, in landmark analyses considering drug discontinuation, and among sub-analyses based on age, sex, and previous disease status.

Among the studied patients, users of branded drugs generally appeared sicker than generic drug users. As shown in Table 11.1, branded drug patients had higher rates of recent hospitalizations, longer hospitalizations (and hospitals were more likely to initiate therapy with originator products), higher medication use, higher rates of previous MACCE events, and higher rates of copayment waivers suggesting lower socioeconomic status. As a result, unad-

justed rates of mortality and MACCE favored generic drugs across all three drug categories. However, despite overall good covariate balancing achieved by IPTW weighting, the associations favoring generics were considerably attenuated but not eliminated after inverse probability of treatment weighting. Thus, residual confounding by unmeasured characteristics remains a possibility. Our sensitivity analyses supplied E-values between 1.4 and 1.69 for the point estimates. For comparison, the ratios of unadjusted and fully adjusted hazard ratios could be interpreted as the amount of bias removed by the considered covariates and ranged from 1.22 to 1.52. We find it unlikely that there is additional unmeasured confounding as strong as or even stronger than all measured covariates considered simultaneously.

Nevertheless, we believe disease severity is a possible source of unmeasured confounding. Because our data only included hospital discharge diagnoses instead of comprehensive medical records data with diagnoses from outpatient health care encounters, we were unable to identify patients with prior disease with high sensitivity. This would explain the large difference in adjusted hazard ratios among patients with and without prior CVDD for hypertension and hyperlipidemia drugs. The subgroups with prior CVDD more accurately reflect a sicker patient pool and have hazard ratios closer to 1, whereas the subgroups without prior CVDD may include patients with relatively more severe disease who were given branded medications. This subgroup difference based on CVDD disappeared when considering the diabetes drugs, and instead it is prior diabetes status that yielded a small but significant difference in subgroup analysis.

Prescribing doctor characteristics have been shown to affect medication prescription preferences, including generic substitution [257]. Physician skepticism about generic medication has been associated with age [258], and pharmaceutical marketing [259], and these trends may extend to Austrian physicians. In Austria, where generic substitution at pharmacies is generally prohibited by law, generic prescription lies entirely with the doctor. Unfortunately, we did not have detailed information on prescribing physicians other than specialty, and therefore could not extensively adjust for physician characteristics.

Two studied drugs had a maximum covariate standardized mean difference greater than 0.1 after IPTW weighting. These were bisoprolol (C07AB07) and repaglinide (A10BX02).

Bisoprolol was associated with a higher copayment waive rate, 0.314 vs 0.265 among generic users, and perhaps lower socioeconomic status among generic users of bisoprolol contributes its being one of only two antihypertensive substances with significant branded drug advantage for mortality and MACCE outcomes. Repaglinide was associated with higher rates of dilated cardiomyopathy (as indicated by past discharge diagnoses) among generic users (0.011 vs 0.002). It was also associated with lower rates of mortality and MACCE among branded users, although this difference is not significant given the small sample size of repaglinide. These observations indicate that the observed outcomes are appreciably affected by imbalance among important covariates. For the other drugs that have maximum SMD smaller than 0.1, there may still be enough residual covariate imbalance after IPTW weighting to impact results. Perhaps other propensity score estimation methods than the high-dimensional propensity score may produce better covariate balance, such as including all covariates via lasso [19] or machine learning algorithms [42].

Medication adherence has previously been identified as a potential cause for the differences between generic and branded drug users, with the observation that the more expensive branded drugs engender lower drug adherence and therefore worse clinical outcomes [246, 247, 260]. However, other than for simvastatin (C10AA01), for which we observed higher branded discontinuation rates, we did not find significant differences in discontinuation among the studied drugs. As branded simvastatin was predominantly reimbursed in the 20mg strength at the time of the study, some patients may have been switched to the generic 40mg form for convenience, as this can be split, providing a longer duration of therapy per pack (and copayment). Landmark analyses using 6-month drug discontinuation also did not produce any significant differences. In Austria, copayment increased from 4.70€ in 2007 to 5.15€ in 2012. Since medication copayments are not different between branded and generic drugs and are generally low [234], we did not expect drug discontinuation to be a prominent concern, as opposed to populations in systems with higher copayments or those with greater copayments for branded medications.

Our studied drugs are not representative of the narrow therapeutic index drugs that pose particular problems regarding generic substitution. Instead, we have studied common drugs

for chronic diseases that could generate the greatest economic savings upon switching to generic formulations. Other studies focusing on similar therapeutic targets find nonsignificant differences between generic and branded drugs. A meta-analysis by Manzoli et al. [243] of small randomized studies for cardiovascular drugs found no difference between generic and branded drugs for both soft and hard clinical outcomes. Randomized studies for statins also did not identify any differences in blood cholesterol levels between generic and branded users [261, 262]. Ahrens et al. [263] studied metoprolol in an observational study and found higher unadjusted cardiovascular event rates among generic users that disappeared upon confounder adjustment. Corrao et al. [245] studied simvastatin in an observational study and found similar discontinuation rates and CV outcomes between generic and branded patients. By contrast, our results favored generic drugs both before and after adjustment. As discussed, this may be due to unmeasured confounding by indication with Austrian health providers, especially doctors in hospitals and specialists, perhaps preferring branded formulations for sicker patients. Although there is no biologically plausible rationale for this strategy, the economic incentives for choosing a particular brand or generic in the hospital setting in Austria can be different from those in the outpatient setting, because the systems of drug acquisition differ markedly in the two sectors [264].

While the limitations of our study include its observational nature, which may lead to residual confounding from unobserved characteristics, there are also several strengths. These include a large study population that is nationally representative and conducted in a country with widely available access to healthcare, thus minimizing adherence differences and the impact of socioeconomic status as confounders. We used state-of-the-art propensity score methods to achieve good balance in observed covariates between compared groups, and we provide robust results with multiple subgroup and E-value sensitivity analyses. Future research would benefit from more detailed outpatient data to better characterize the patients' health status, and more detailed prescriber characteristics, which could be achieved by linking additional data sources.

We conclude from this comprehensive study of almost all insured individuals in Austria that use of generic medications associated with similar or even slightly lower rates of mortality

or nonfatal cardiovascular events. While there remains the potential for residual confounding by indication, our findings support the safety of policies towards greater use of generic substitute medications relative to their branded, and usually more expensive, versions.

Variable	Branded anti-hypertensive(N= 427,641)	Generic anti-hypertensive(N= 558,508)	Branded lipid-lowering (N= 21,665)	Generic lipid-lowering (N= 25,694)	Branded hypo-glycemic (N= 101,045)	Generic hypo-glycemic (N= 99,993)
Age (years), mean (SD)	64.5 (15.4)	63.3 (14.5)	63.0 (12.8)	62.4 (12.5)	65.0 (13.9)	62.7 (13.5)
Sex: female	54.3%	54.5%	50.3%	54.3%	49.5%	50.2%
Copayment waiver	34.2%	29.9%	30.6%	26.3%	41.1%	38.3%
Hospitalization (ending in last 180 days)	30.0%	19.9%	23.0%	18.1%	26.2%	18.7%
Hospitalization > 14 days (ending in last 14 days)	7.3%	2.5%	3.7%	2.0%	5.7%	2.0%
Hospitalization > 14 days (ending in last 180 days)	10.8%	5.4%	6.9%	4.6%	9.1%	4.9%
Index year:						
2007	0.7%	0.6%	0.1%	0.4%	1.0%	0.7%
2008	10.1%	9.1%	1.4%	4.5%	11.2%	8.2%
2009	22.4%	22.0%	33.3%	16.9%	22.4%	18.4%
2010	27.0%	24.9%	28.1%	35.5%	25.6%	25.7%
2011	20.6%	23.3%	22.1%	25.4%	22.2%	25.4%
2012	19.2%	20.1%	15.0%	17.3%	17.6%	21.5%
Specialty of prescriber:						
General practitioner	67.7%	78.5%	68.9%	75.4%	77.5%	81.7%
Internal medicine specialist	11.6%	13.3%	12.8%	15.5%	9.7%	10.4%
Hospital	9.8%	3.7%	10.2%	3.7%	5.7%	2.8%
Other	10.8%	4.6%	8.1%	5.4%	7.1%	5.2%
Recent myo-cardial infarction	2.5%	0.6%	3.8%	1.1%	0.9%	0.3%
Recent cerebro-vascular event	2.3%	1.1%	2.3%	1.6%	1.6%	0.6%
Any diagnosis in group of endocrine, nutritional or metabolic diseases or in group of diseases of circulatory system	33.0%	16.6%	25.3%	15.5%	28.0%	15.5%
Previous use of antihypertensive, lipid-lowering or hypoglycemic medicines	67.5%	64.1%	73.0%	69.0%	81.3%	76.4%
Previous use of injectable insulins	1.9%	1.4%	2.0%	1.4%	3.2%	2.5%
Previous use of oral hypoglycemic drugs	12.7%	11.5%	14.6%	13.0%	39.2%	21.6%

Table 11.1: Characteristics of patients at first index prescription for antihypertensive, lipid-lowering or hypoglycemic treatment.

Indication	Adjustment variables	Mortality HR (95%CI) for branded vs. generic	MACCE HR (95%CI) for branded vs. generic
Antihypertensive drugs	No adjustment	1.75 (1.56, 1.98)	1.62 (1.47, 1.77)
	Age, sex, copayment waiver	1.52 (1.37, 1.69)	1.44 (1.33, 1.56)
	Extended set of covariates*	1.23 (1.13, 1.34)	1.18 (1.12, 1.25)
	IPTW via high-dimensional propensity scores	**1.15 (1.06, 1.25)**	**1.13 (1.07, 1.20)**
	E-value (lower 95% confidence limit)**	1.57 (1.31)	1.51 (1.34)
Lipid-lowering drugs	No adjustment	1.69 (1.09, 2.63)	1.49 (1.04, 2.12)
	Age, sex, copayment waiver	1.40 (0.84, 2.32)	1.32 (0.89, 1.96)
	Extended set of covariates*	1.33 (1.07, 1.64)	1.26 (1.17, 1.35)
	IPTW via high-dimensional propensity scores	**1.13 (0.86, 1.47)**	**1.20 (1.05, 1.38)**
	E-value (lower 95% confidence limit)**	1.51 (1)	1.69 (1.28)
Hypoglycemic drugs	No adjustment	1.43 (1.37, 1.49)	1.35 (1.31, 1.39)
	Age, sex, copayment waiver	1.32 (1.24, 1.40)	1.29 (1.26, 1.33)
	Extended set of covariates*	1.11 (1.01, 1.23)	1.10 (1.01, 1.18)
	IPTW via high-dimensional propensity scores	**1.09 (0.93, 1.28)**	**1.11 (1.03, 1.20)**
	E-value (lower 95% confidence limit)**	1.4 (1)	1.45 (1.21)

Table 11.2: Pooled IPTW-adjusted hazard ratios (HR) for all-cause mortality and major cardiac or cerebrovascular events (MACCE) comparing branded vs. generic medicines applying different levels of adjustment. *Age, sex, copayment waiver status, calendar year, specialty of prescriber, previous hospitalizations, recent MI or cerebrovascular events, any diagnosis in group of endocrine, nutritional or metabolic diseases or in group of diseases of circulatory system, any previous use of antihypertensive, lipid-lowering or hypoglycemic drugs. **E-values [1] for IPTW adjusted point estimate and lower confidence limit.

Indication Substance	ATC code	N branded	N generics	Mortality: HR (95%CI) for branded vs. generic	MACCE: HR (95%CI) for branded vs. generic
Antihypertensive drugs					
Metropolol	C07AB02	9,185	8,202	1.13 (0.96, 1.32)	1.15 (1.01, 1.30)
Bisoprolol	C07AB07	135,208	29,442	0.84 (0.76, 0.92)	0.91 (0.85, 0.98)
Nebivolol	C07AB12	91,283	18,561	0.81 (0.68, 0.97)	0.98 (0.86, 1.11)
Carvedilol	C07AG02	20,837	37,181	1.19 (1.10, 1.28)	1.17 (1.10, 1.25)
Amlodipine	C08CA01	28,118	84,988	1.41 (1.35, 1.48)	1.28 (1.23, 1.33)
Enalapril	C09AA02	17,053	71,065	1.08 (1.02, 1.15)	1.06 (1.01, 1.11)
Lisinopril	C09AA03	51,443	99,145	1.15 (1.10, 1.20)	1.17 (1.13, 1.21)
Ramipril	C09AA05	27,388	79,301	1.32 (1.24, 1.41)	1.26 (1.20, 1.33)
Enalapril and diuretics	C09BA02	4,492	16,024	1.08 (0.96, 1.22)	1.01 (0.91, 1.12)
Lisinopril and diuretics	C09BA03	27,967	66,727	1.24 (1.16, 1.32)	1.19 (1.13, 1.25)
Ramipril and diuretics	C09BA05	10,643	39,711	1.25 (1.15, 1.37)	1.23 (1.15, 1.32)
Losartan and diuretics	C09DA01	4,024	8,161	1.64 (1.23, 2.20)	1.20 (0.98, 1.46)
Lipid-lowering drugs					
Simvastatin	C10AA01	1,862	10,079	0.97 (0.76, 1.42)	1.09 (0.87, 1.33)
Fluvastatin	C10AA04	19,803	15,615	1.28 (1.05, 1.56)	1.26 (1.14, 1.40)
Hypoglycemic drugs					
Metformin	A10BA02	41,889	87,929	1.21 (1.15, 1.26)	1.16 (1.11, 1.20)
Gliclazide	A10BB09	50,520	11,504	1.02 (0.94, 1.11)	1.08 (1.01, 1.16)
Repaglinide	A10BX02	2,636	560 0.91	(0.57, 1.45)	0.83 (0.57, 1.22)

Table 11.3: IPTW-adjusted hazard ratios (HR) and 95% confidence intervals (CI) of all-cause mortality and major cardiac or cerebrovascular events (MACCE) for individual substances.

Subgroup	N branded	N generics	Mortality HR (95%CI)	p-value for interaction*	MACCE HR (95%CI)	p-value for interaction*
Females	232,229	304,174	1.17 (1.06, 1.28)	0.3075	1.14 (1.07, 1.22)	0.6376
Males	195,412	254,334	1.14 (1.04, 1.24)		1.13 (1.07, 1.19)	
Age $\leq$ 70 years	263,733	371,355	1.24 (1.11, 1.38)	0.0019	1.20 (1.13, 1.28)	< 0.0001
Age $>$ 70 years	163,908	187,153	1.13 (1.04, 1.22)		1.10 (1.04, 1.17)	
No history of CVDD	110,087	183,383	1.47 (1.31, 1.64)	< 0.0001	1.34 (1.22, 1.47)	< 0.0001
History of CVDD	317,554	375,125	1.10 (1.01, 1.20)		1.09 (1.03, 1.16)	
No diabetes	373,288	494,452	1.17 (1.07, 1.28)	0.0006	1.15 (1.08, 1.23)	0.0003
Oral DM therapy but no insulin use	46,134	56,091	1.10 (0.99, 1.22)		1.08 (1.01, 1.15)	
Insulin use	8,219	7,965	1.06 (0.96, 1.18)		1.08 (0.99, 1.18)	
Time-dependent effect: $\leq$ 6 months	427,641	558,508	1.12 (1.01, 1.24)	0.0876	1.28 (1.03, 1.58)	0.0111
$>$ 6 months	352,719	479,187	1.16 (1.06, 1.27)		1.13 (1.07, 1.20)	

Table 11.4: Antihypertensive drugs: pooled IPTW-adjusted hazard ratios from subgroup analyses. *p-value for interaction of a variable with treatment, i.e., for testing the null hypothesis that HR is equal in the subgroups.

Subgroup	N branded	N generics	Mortality HR (95%CI)	p-value for interaction*	MACCE HR (95%CI)	p-value for interaction*
Females	10,900	13,945	1.09 (0.79, 1.50)	0.3072	1.19 (1.04, 1.366)	0.3533
Males	10,765	11,749	1.18 (0.95, 1.47)		1.24 (1.07, 1.44)	
Age $\leq$ 70 years	15,381	18,658	1.07 (0.68, 1.69)	0.3721	1.18 (0.92, 1.51)	0.5031
Age $>$ 70 years	6,284	7,036	1.18 (0.97, 1.43)		1.23 (1.08, 1.40)	
No history of CVDD	4,798	7,237	1.64 (1.14, 2.37)	< 0.0001	1.60 (1.19, 2.14)	< 0.0001
History of CVDD	16,867	18,457	1.08 (0.81, 1.43)		1.16 (1.03, 1.32)	
No diabetes	18,512	22,351	1.21 (1.01, 1.44)	0.2248	1.26 (1.14, 1.40)	0.0046
Oral DM therapy but no insulin use	2,727	2,977	0.85 (0.40, 1.81)		1.08 (0.86, 1.34)	
Insulin use	426	366	0.97 (0.36, 2.56)		0.76 (0.19, 3.06)	
Time-dependent effect:	21,665	25,694	1.32 (0.99, 1.77)	0.0294	1.37 (1.16, 1.62)	0.0009
$>$ 6 months	19,055	22,857	1.08 (0.80, 1.45)		1.10 (0.86 1.41)	

Table 11.5: Lipid-lowering drugs: pooled IPTW-adjusted hazard ratios from subgroup analyses. *p-value for interaction of a variable with treatment, i.e., for testing the null hypothesis that HR is equal in the subgroups.

Subgroup	N branded	N generics	Mortality HR (95%CI)	p-value for interaction*	MACCE HR (95%CI)	p-value for interaction*
Females	50,026	50,147	1.06 (0.85, 1,31)	0.3072	1.14 (1.03, 1.25)	0.2597
Males	51,019	49,846	1.12 (1.00, 1.25)		1.11 (1.06, 1.16)	
Age $\leq$ 70 years	63,719	70,352	1.19 (1.11,1.29)	0.0075	1.16 (1.11, 1.22)	0.0008
Age $>$ 70 years	37,326	29,641	1.04 (0.84, 1.29)		1.07 (0.97, 1.18)	
No history of CVDD	15,754	21,629	1.06 (0.78, 1.43)	0.5352	1.04 (0.77, 1.40)	0.2720
History of CVDD	85,291	78,364	1.11 (0.96, 1.28)		1.12 (1.06, 1.19)	
No diabetes	61,401	78,420	1.14 (1.00, 1.30)	0.0475	1.15 (1.11, 1.19)	0.0027
Oral DM therapy but no insulin use	36,445	19,061	1.07 (0.93, 1.23)		1.09 (1.02, 1.16)	
Insulin use	3,199	2,512	0.97 (0.69, 1.38)		1.09 (0.84, 1.40)	
Time-dependent effect:	101,045	99,993	1.16 (1.06, 1.26)	0.0269	1.08 (1.02, 1.15)	0.3153
$>$ 6 months	85,409	84,899	1.03 (0.82, 1.29)		1.11 (1.00, 1.24)	

Table 11.6: Hypoglycemic drugs: pooled IPTW-adjusted hazard ratios from subgroup analyses. *p-value for interaction of a variable with treatment, i.e., for testing the null hypothesis that HR is equal in the subgroups.

CHAPTER 12

Conclusion

12.1 Vertical Integration

Observational clinical research spans multiple fields including computer science, statistics, epidemiology, and medicine. Progress in one area propel advances in other fields. The main computational bottleneck in performing observational studies is the time required to construct large-scale propensity scores. Chapter 7 describes advances in GPU programming for logistic regression that decrease the runtime of constructing a PS with over 10,000 covariates by almost four times compared to multi-threaded CPU and more than ten times compared to single-threaded CPU. This allowed the research of Chapter 5 to be computed on a single personal TITAN V GPU card, whereas the earlier research of Chapter 4 required use of the shared Hoffman2 computing cluster at UCLA. I had to queue for days to acquire computing resources on Hoffman2 and utilized up to approximately 100 CPUs simultaneously. If our methodology were applied on a larger scale requiring the use of cloud computing resources, the improved efficiency of our statistical regressions could represent significant savings. In such a case, even a 2× improvement in runtime, possibly unimpressive to some, would result in halving the cost of computing.

Even without GPU programming, the Cyclops R package, published on CRAN, facilitates novel large-scale clinical research approaches. Using Cyclops, the research presented in Chapter 10 was conducted in "five weeks on a computer with 32 processing cores and 168GB of memory" [229]. That is a seemingly long but previously unattainable process to conduct an all-by-all comparison of treatments within an entire clinical domain (depression). These computational statistics advances of Suchard et al. [28] have introduced a

new paradigm in exploring up to thousands of hypotheses under a consistent study design. The antihypertensive medication study of [210] uses the same new observational analytic framework.

The software suite developed in the OHDSI community, described at the end of Chapter 3, has provided tools for start-to-finish execution of a modern observational study. From the moment of clinical question conception, the investigator can interface with ATLAS to create cohort definitions and study specifications. Epidemiological considerations such as PS estimation, PS adjustment, and outcome model decisions are incorporated into the study design, and implemented through CohortMethod. A full stable of other software support all aspects of the study execution and analysis. As an open community, OHDSI invites all interested participants of observational clinical analytics, from medicine, academia, industry, and government. The aforementioned software tools are open source, with the goal of advancing observational research to improve clinical practice and benefit the most patients.

12.2 GPU All the Things

Developed by Marc Suchard, the currently published version of Cyclops provides efficient cyclic coordinate descent optimization of common generalized linear models using the CPU. I have continued the work of developing GPU code, started in [28] for the self-controlled case series, in my research in Chapter 7 on logistic regression. Immediate future work includes developing this GPU code to production quality for other researchers' use, particularly for fitting large-scale propensity scores. Subsequent research includes developing GPU code for the other regressions serviced by Cyclops, including Cox proportional hazards, tied Cox models, conditional logistic regression, and Poisson regression.

In Chapter 7, by interleaving vectors to achieve coalesced memory access, I develop GPU code that scales well with the number of cross-validation replicates, of which I tested up to 1,000. This GPU code can be repurposed and further developed for other applications in which multiple regressions share the same underlying data but have different weights. The most immediate application is to perform cross-validation on a grid of λ values, an approach

favored by many researchers. Another application is to apply this principle to bootstrapping, in which each sample is a weighted resampling of the original population.

As mentioned in [28], a full Bayesian analysis of our regularization hyperparameters is yet computationally intractable. The maximum likelihood estimation methods we use are frequentist stand-in until Bayesian methods become available. Future research should focus on developing a Bayesian alternative to cross-validation, ideally with GPU acceleration. Machine learning is also a burgeoning field of computer science that is able to construct predictive models such as the propensity score. While there have been some machine-learning extensions to existing PS estimation algorithms [89], there remains a need to comprehensively compare large-scale regularized regression [19] to machine learning alternatives.

12.3 The Network Study

Dubbed the "Save Our Sisyphus Challenge," the alendronate vs raloxifene study of Chapter 8 was the first OHDSI collaboration demonstrating the network capability of the community. Individual OHDSI collaborators who are data holders volunteered participation in the study, and UCLA operated as a central study coordinator for data analysis (Figure 12.1). Each participating center was sent a R package that fully executed the study, and sent back to the study coordinator a zip file including study results devoid of individual patient-level information. By presenting meta-analyses of results from 9 different databases, the study provides greater evidence than what can be provided by a single data source.

Chapters 8-10 all execute studies across multiple databases in the OHDSI network. By using a consistent study design, differences among the databases are attributable to population differences and residual bias, and not from study implementation discrepancies. The data reported in Chapter 11 are displayed graphically online at `https://data.ohdsi.org/SystematicEvidence/`, and highlight the reality that different databases can provide different answers to the same question. Network studies increase the scope of a studies by providing evidence from multiple data sources, allowing clinical questions to be answered using the maximum study populations available.

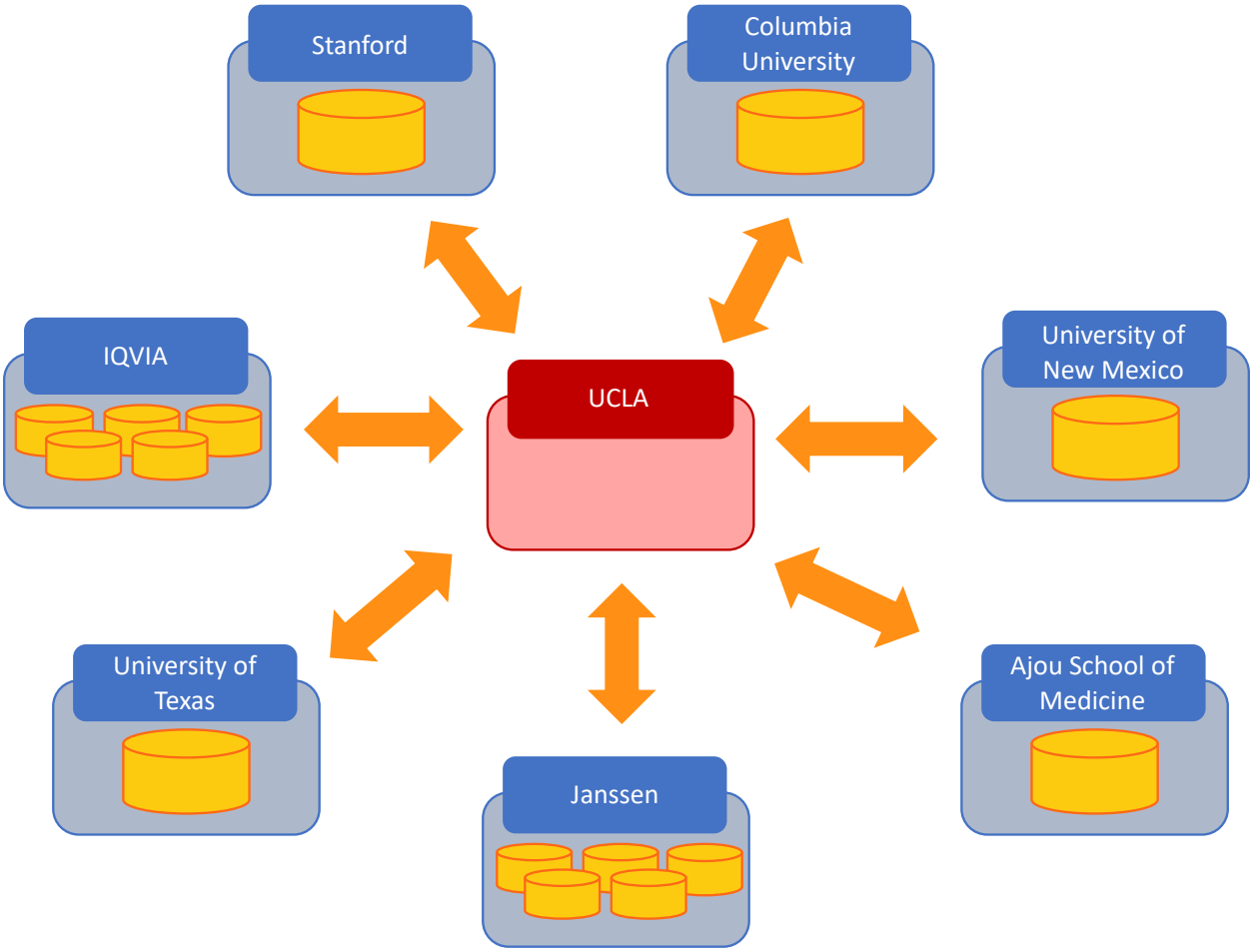


Figure 12.1: Collaborators in alendronate-raloxifene study

Unfortunately, several US databases overlap in their patient populations. We don't know which patients overlap, but we know that the total combined sizes of available databases exceeds the population of the country. The ideal network study combines nonoverlapping databases. For example, chapter 11 is a study encompassing nearly all residents of Austria, through a national database that would be ideal for a network study with other national databases. The United States has no national database providing cradle-to-grave longitudinal health data as some European countries do, complicating observational research.

12.4 Automated Drug Surveillance

I am interested in applying observational analytical tools in building systems for utilizing health data. The United States lacks automated drug surveillance systems to detect drug safety issues, with the emphasis on "automated." While the Food and Drug Administration does have the ability to conduct drug safety studies in a distributed network of data sources [265], it still relies on postmarket reviews or spontaneous reporting to identify signals worth investigating. I envision a system in which drugs are constantly and automatically evaluated for signals among outcomes of interest, so that adverse events can be detected at the earliest possible time. Perhaps the all-by-all study paradigm used in Chapter 10 and recent OHDSI studies [229, 210] for depression and hypertension can be run regularly for all major clinical domains. This process would require large computational resources to query databases and compute propensity scores, and sharp attention to detail in storing, updating, and presenting results. As new longitudinal data become available, warm starts in fitting PS and outcome models can reduce the computational burden of repeated model fitting. With up-to-date and fully statistically adjusted hazard ratios comparing relevant treatments across many clinical domains, we can truly provide pharmacoviligance for patients regarding the medical products they utilize.

Another application of the observational tools used in the OHDSI community is to provide observational studies that parallel ongoing randomized clinical trials. Clinical trials for products already on the market can benefit from a concurrent observational study researching the same clinical question in a larger, real-world population. The observational study may produce interesting safety signals that the clinical trial is underpowered to detect, or provide ongoing results before clinical study endpoints are measured. Discrepancies between observational and randomized trial results can be of interest to investigators. One close collaborator told me that it takes as much faith to generalize randomized and highly controlled trial results to broader, real-world study populations as it does to believe the unverifiable assumptions that underlie observational analysis. Both observational and randomized data have flaws and benefits, and contribute to the net body of clinical evidence.

Bibliography

[1] VanderWeele TJ and Ding P (2017). Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine*, 167(4):268–274.

[2] Overhage JM, Ryan PB, Reich CG, Hartzema AG, and Stang PE (2011). Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc*, 19(1):54–60.

[3] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. (2015). Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*, 216:574.

[4] of the National Coordinator for Health Information Technology O (2016a). Office-based physician electronic health record adoption. *Health IT Quick-Stat #50*.

[5] of the National Coordinator for Health Information Technology O (2016b). Non-federal acute care hospital electronic health record adoption.

[6] Menachemi N and Collum TH (2011). Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, 4:47.

[7] Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, et al. (2012). Design considerations, architecture, and use of the mini-sentinel distributed data system. *Pharmacoepidemiology and drug safety*, 21:23–31.

[8] Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

[9] Imbens GW and Rubin DB (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[10] Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, and Madigan D (2014). Interpreting observational studies: why empirical calibration is needed to correct p-values. *Stat Med*, 33(2):209–218.

[11] Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, and Suchard MA (2018). Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences*, 115(11):2571–2577.

[12] Cardwell CR, Abnet CC, Cantwell MM, and Murray LJ (2010). Exposure to oral bisphosphonates and risk of esophageal cancer. *JAMA*, 304(6):657–663.

[13] Green J, Czanner G, Reeves G, Watson J, Wise L, and Beral V (2010). Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort. *BMJ*, 341:c4444.

[14] Lipsitch M, Tchetgen ET, and Cohen T (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388.

[15] Arnold BF and Ercumen A (2016). Negative control outcomes: a tool to detect bias in randomized trials. *JAMA*, 316(24):2597–2598.

[16] Rodríguez G (2007). Lecture notes on generalized linear models. *URL: http://data. princeton. edu/wws509/notes/c4. pdf*.

[17] Wu TT, Lange K, et al. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244.

[18] Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, and Brookhart MA (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512 – 522.

[19] Tian Y, Schuemie MJ, and Suchard MA (2018). Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *International Journal of Epidemiology*, 47(6):2005–2014.

[20] Tibshirani R (1996). Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*, 57(1):267–288.

[21] Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[22] Rubin DB (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36.

[23] Rubin DB (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, pages 808–840.

[24] Rubin DB (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28(9):1420–1423.

[25] Schuemie MJ, Cepeda MS, Suchard MA, Yang J, Schuler Y. TA, Ryan PB, et al. (2020). How confident are we about observational findings in health care: A benchmark study. *Harvard Data Science Review*, 2(1).

[26] Austin PC (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*, 46(3):399–424.

[27] Schuemie M, Suchard M, and Ryan P (2019). *CohortMethod: New-user cohort method with large scale propensity and outcome models*. https://ohdsi.github.io/CohortMethod, https://github.com/OHDSI/CohortMethod.

[28] Suchard MA, Simpson SE, Zorych I, Ryan P, and Madigan D (2013). Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Trans Model Comput Simul*, 23(1):10.

[29] Simpson SE, Madigan D, Zorych I, Schuemie MJ, Ryan PB, and Suchard MA (2013). Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics*, 69(4):893–902.

[30] Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. (2016). Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A*, 113(27):7329–7336.

[31] Brookhart MA, Stürmer T, Glynn RJ, Rassen J, and Schneeweiss S (2010). Confounding control in healthcare database research: challenges and potential approaches. *Med Care*, 48(6 Suppl):S114–S120.

[32] Ryan PB, Madigan D, Stang PE, Marc Overhage J, Racoosin JA, and Hartzema AG (2012). Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat Med*, 31(30):4401–4415.

[33] Rubin DB (1997). Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*, 127(8 Pt 2):757–763.

[34] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, and Stürmer T (2006). Variable selection for propensity score models. *Am J Epidemiol*, 163(12):1149–1156.

[35] King G and Nielsen R (2016). Why propensity scores should not be used for matching. *Copy at http://j.mp/1sexgVw Download Citation BibTex Tagged XML Download Paper*, 378.

[36] Rassen JA, Glynn RJ, Brookhart MA, and Schneeweiss S (2011). Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *Am J Epidemiol*, 173(12):1404–1413.

[37] Imai K and Ratkovic M (2014). Covariate balancing propensity score. *J R Stat Soc Series B Stat Methodol*, 76(1):243–263.

[38] Ryan PB, Schuemie MJ, Gruber S, Zorych I, and Madigan D (2013). Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. *Drug Saf*, 36(1):59–72.

[39] Greenland S (2008). Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol*, 167(5):523–529.

[40] Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen J, and Franklin JM (2017). Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. *Epidemiology*, 28(2):237–248.

[41] Rubin DB and Thomas N (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, pages 249–264.

[42] Franklin JM, Eddings W, Austin PC, Stuart EA, and Schneeweiss S (2017). Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. *Stat Med*, 36(12):1946–1963.

[43] Franklin JM, Eddings W, Glynn RJ, and Schneeweiss S (2015). Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *Am J Epidemiol*, 182(7):651–659.

[44] Franklin JM, Schneeweiss S, Polinski JM, and Rassen JA (2014). Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*, 72:219–226.

[45] Vaughan LK, Divers J, Padilla MA, Redden DT, Tiwari HK, Pomp D, et al. (2009). The use of plasmodes as a supplement to simulations: A simple example evaluating individual admixture estimation methodologies. *Comput Stat Data Anal*, 53(5):1755–1766.

[46] Lawless J (1998). Parametric models in survival analysis. *Encyclopedia of Biostatistics*.

[47] Whittemore AS and Keller JB (1986). Survival estimation using splines. *Biometrics*, pages 495–506.

[48] Efron B (1977). The efficiency of Cox's likelihood function for censored data. *J Am Stat Assoc*, 72(359):557–565.

[49] Breslow N (1972). Discussion of the paper by D.R. Cox. *J R Stat Soc Series B Stat Methodol*, 34:216–217.

[50] Graham DJ, Reichman ME, Wernecke M, Zhang R, Southworth MR, Levenson M, et al. (2015). Cardiovascular, bleeding, and mortality risks in elderly Medicare patients treated with dabigatran or warfarin for non-valvular atrial fibrillation. *Circulation*, 131:157–164.

[51] Garbe E, Kloss S, Suling M, Pigeot I, and Schneeweiss S (2013). High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications. *Eur J Clin Pharmacol*, 69(3):549–557.

[52] Mittal S, Madigan D, Burd RS, and Suchard MA (2013). High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis. *Biostatistics*, 15(2):207–221.

[53] Walker AM (1991). Confounding. In *Observation and Inference*, chapter 9, pages 119–128. Epidemiology Resources Incorporated.

[54] Connolly JG, Maro JC, Wang SV, Toh S, Fuller CC, Panozzo CA, et al. (2016). Development, applications, and methodological challenges to the use of propensity score matching approaches in fda's sentinel program. In *Pharmacoepidemiology and Drug Safety*, volume 25, pages 402–403. Wiley-Blackwell 111 River St. Hoboken 07030-5774, NJ USA.

[55] Zhou M, Wang SV, Leonard CE, Gagne JJ, Fuller C, Hampp C, et al. (2017). Sentinel modular program for propensity-score matched cohort analyses: Application to glyburide, glipizide, and serious hypoglycemia. *Epidemiology*.

[56] Schuemie MJ, Suchard MA, and Ryan PB (2016). *CohortMethod: New-user cohort method with large scale propensity and outcome models*. R package version 2.1.0.

[57] King G and Nielsen R (2015). Why propensity scores should not be used for matching. *Working Paper*.

[58] Austin PC and Stuart EA (2015). Optimal full matching for survival outcomes: a method that merits more widespread use. *Stat Med*, 34(30):3949–3967.

[59] Austin PC (2008). Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. *Pharmacoepidemiol Drug Saf*, 17(12):1218–1225.

[60] Rassen JA, Shelat AA, Myers J, Glynn RJ, Rothman KJ, and Schneeweiss S (2012). One-to-many propensity score matching in cohort studies. *Pharmacoepidemiol Drug Saf*, 21(S2):69–80.

[61] Voss EA, Boyce RD, Ryan PB, van der Lei J, Rijnbeek PR, and Schuemie MJ (2017). Accuracy of an automated knowledge base for identifying drug adverse reactions. *J Biomed Inform*, 66:72–81.

[62] Owen AB (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

[63] Franklin JM, Rassen JA, Ackermann D, Bartels DB, and Schneeweiss S (2014). Metrics for covariate balance in cohort studies of causal effects. *Stat Med*, 33(10):1685–1699.

[64] Walker AM, Patrick AR, Lauer MS, Hornbrook MC, Marin MG, Platt R, et al. (2013). A tool for assessing the feasibility of comparative effectiveness research. *Comparative Effectiveness Research*, 2013(1):11–20.

[65] Shortreed SM and Ertefaie A (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122.

[66] Kumamaru H, Gagne JJ, Glynn RJ, Setoguchi S, and Schneeweiss S (2016). Comparison of high-dimensional confounder summary scores in comparative studies of newly marketed medications. *J Clin Epidemiol*, 76:200–208.

[67] Ding P, Vanderweele T, and Robins J (2017). Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*, 104(2):291–302.

[68] Greenland S (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729.

[69] Angrist J (1991). Instrumental variables estimation of average treatment effects in econometrics and epidemiology.

[70] Angrist JD, Imbens GW, and Rubin DB (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

[71] Martens EP, Pestman WR, de Boer A, Belitser SV, and Klungel OH (2006). Instrumental variables: application and limitations. *Epidemiology*, 17(3):260–267.

[72] Brooks JM and Ohsfeldt RL (2013). Squeezing the balloon: propensity scores and unmeasured covariate balance. *Health Serv Res*, 48(4):1487–1507.

[73] Pearl J (2011). Invited commentary: understanding bias amplification. *American Journal of Epidemiology*, 174(11):1223–1227.

[74] Wooldridge JM (2016). Should instrumental variables be used as matching variables? *Research in Economics*, 70(2):232–237.

[75] Bhattacharya J and Vogt WB (2007). Do instrumental variables belong in propensity scores?

[76] Lefebvre G, Delaney JA, and Platt RW (2008). Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics in Medicine*, 27(18):3629–3642.

[77] Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, et al. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, 174(11):1213–1222.

[78] Caruana E, Chevret S, Resche-Rigon M, and Pirracchio R (2015). A new weighted balance measure helped to select the variables to be included in a propensity score model. *J Clin Epidemiol*, 68(12):1415–1422.

[79] Hernán MA and Robins JM (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4):360–372.

[80] D'Agostino RB (2007). Estimating treatment effects using observational data. *JAMA*, 297(3):314–316.

[81] Garabedian LF, Chu P, Toh S, Zaslavsky AM, and Soumerai SB (2014). Potential bias of instrumental variable analyses for observational comparative effectiveness research. *Annals of Internal Medicine*, 161(2):131–138.

[82] Patrick AR, Schneeweiss S, Brookhart MA, Glynn RJ, Rothman KJ, Avorn J, et al. (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidemiology and Drug Safety*, 20(6):551–559.

[83] Bross ID (1966). Spurious effects from an extraneous variable. *Journal of Chronic Diseases*, 19(6):637–647.

[84] Cain LE, Cole SR, Greenland S, Brown TT, Chmiel JS, Kingsley L, et al. (2009). Effect of highly active antiretroviral therapy on incident aids using calendar period as an instrumental variable. *American Journal of Epidemiology*, 169(9):1124–1132.

[85] Mack CD, Brookhart MA, Glynn RJ, Meyer AM, Carpenter WR, Sandler RS, et al. (2015). Comparative effectiveness of oxaliplatin vs. 5-flourouricil in older adults: an instrumental variable analysis. *Epidemiology (Cambridge, Mass.)*, 26(5):690.

[86] Ray WA (2003). Evaluating medication effects outside of clinical trials: new-user designs. *American Journal of Epidemiology*, 158(9):915–920.

[87] Schuemie MJ, Suchard MA, and Ryan PB (2017). *CohortMethod: New-user cohort method with large scale propensity and outcome models*. R package version 2.4.4.

[88] VanderWeele TJ (2019). Principles of confounder selection. *European Journal of Epidemiology*, 34(3):211–219.

[89] Schneeweiss S (2018). Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clinical Epidemiology*, 10:771.

[90] Austin PC and Mamdani MM (2006). A comparison of propensity score methods: a case-study estimating the effectiveness of post-ami statin use. *Statistics in Medicine*, 25(12):2084–2106.

[91] Austin PC (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine*, 32(16):2837–2849.

[92] Austin PC and Schuster T (2016). The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study. *Statistical Methods in Medical Research*, 25(5):2214–2237.

[93] Zhang G and Little R (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65(3):911–918.

[94] Lechner M (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies*, pages 43–58. Springer.

[95] McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, and Burgette LF (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med*, 32(19):3388–3414.

[96] Yoshida K, Hernández-Díaz S, Solomon DH, Jackson JW, Gagne JJ, Glynn RJ, et al. (2017). Matching weights to simultaneously compare three treatment groups: comparison to three-way matching. *Epidemiology (Cambridge, Mass.)*, 28(3):387.

[97] Rosenbaum PR (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

[98] Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, and Smith D (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health*, 13(2):273–277.

[99] Austin PC and Stuart EA (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.

[100] Greenland S (1991). Estimating standardized parameters from generalized linear models. *Statistics in Medicine*, 10(7):1069–1074.

[101] Sylvain D and Richard S (1989). Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561.

[102] Stone CJ (1986). [generalized additive models]: Comment. *Statistical Science*, 1(3):312–314.

[103] Wood SN (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

[104] Cole SR and Hernán MA (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664.

[105] Lee BK, Lessler J, and Stuart EA (2011). Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174.

[106] T. S, R. W, J. GR, and A. BM (2014). Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. *Journal of Internal Medicine*, 275(6):570–580.

[107] Gutman R and Rubin DB (2013). Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. *Statistics in Medicine*, 32(11):1795–1814.

[108] Petersen ML, Porter KE, Gruber S, Wang Y, and van der Laan MJ (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54.

[109] Myers JA and Louis TA (2012). Comparing treatments via the propensity score: stratification or modeling? *Health Services and Outcomes Research Methodology*, 12(1):29–43.

[110] Rassen JA, Shelat AA, Franklin JM, Glynn RJ, Solomon DH, and Schneeweiss S (2013). Matching by propensity score in cohort studies with three treatment groups. *Epidemiology*, 24(3):401–409.

[111] L L and T G (2015). A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics*, 9(2):215–234.

[112] Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. (2010). Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Annals of Internal Medicine*, 153(9):600–606.

[113] Madigan D, Ryan P, Simpson S, and Zorych I (2010). Bayesian methods in pharmacovigilance. *Bayesian Statistics*, 9:421–438.

[114] Kyung M, Gill J, Ghosh M, Casella G, et al. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal*, 5(2):369–411.

[115] Rossini AJ, Tierney L, and Li N (2007). Simple parallel statistical computing in r. *Journal of Computational and Graphical Statistics*, 16(2):399–420.

[116] Raina R, Madhavan A, and Ng AY (2009). Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, pages 873–880. ACM.

[117] Cano A (2018). A survey on graphic processing unit computing for large-scale data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(1):e1232.

[118] Athanasopoulos A, Dimou A, Mezaris V, and Kompatsiaris I (2011). Gpu acceleration for support vector machines. In *Procs. 12th Inter. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2011), Delft, Netherlands*, pages 17–55.

[119] Li Q, Salman R, Test E, Strack R, and Kecman V (2013). Parallel multitask cross validation for support vector machine using gpu. *Journal of Parallel and Distributed Computing*, 73(3):293–302.

[120] Wu Z, Wang Q, Plaza A, Li J, Sun L, and Wei Z (2015). Real-time implementation of the sparse multinomial logistic regression for hyperspectral image classification on gpus. *IEEE Geoscience and Remote Sensing Letters*, 12(7):1456–1460.

[121] Suchard MA, Wang Q, Chan C, Frelinger J, Cron A, and West M (2010). Understanding gpu programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of Computational and Graphical Statistics*, 19(2):419–438.

[122] Zhou H, Lange K, and Suchard MA (2010). Graphics processing units and high-dimensional optimization. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, 25(3):311.

[123] Genkin A, Lewis DD, and Madigan D (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304.

[124] Stone JE, Gohara D, and Shi G (2010). Opencl: A parallel programming standard for heterogeneous computing systems. *Computing in Science & Engineering*, 12(3):66.

[125] Kirk D et al. (2007). Nvidia cuda software and gpu parallel computing architecture. In *ISMM*, volume 7, pages 103–104.

[126] Tarjan D, Skadron K, and Micikevicius P (2009). The art of performance tuning for cuda and manycore architectures. *Birds-of-a-feather Session at SC*, 9.

[127] Micikevicius P (2009). 3d finite difference computation on gpus using cuda. In *Proceedings of 2nd Workshop on General Purpose Processing on Graphics Processing Units*, pages 79–84. ACM.

[128] Fujii Y, Azumi T, Nishio N, Kato S, and Edahiro M (2013). Data transfer matters for gpu computing. In *2013 International Conference on Parallel and Distributed Systems*, pages 275–282. IEEE.

[129] Park T and Casella G (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

[130] Minka TP (2003). A comparison of numerical optimizers for logistic regression. *Unpublished draft*.

[131] Krstajic D, Buturovic LJ, Leahy DE, and Thomas S (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6(1):10.

[132] Friedman J, Hastie T, and Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

[133] Hunter DR and Lange K (2004). A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37.

[134] Bone HG, Hosking D, Devogelaer JP, Tucci JR, Emkey RD, Tonino RP, et al. (2004). Ten years' experience with alendronate for osteoporosis in postmenopausal women. *N Engl J Med*, 350(12):1189–1199.

[135] Kanis J, Johnell O, Odén A, Johansson H, and McCloskey E (2008). Frax^{TM} and the assessment of fracture probability in men and women from the uk. *Osteoporos Int*, 19(4):385–397.

[136] Hernlund E, Svedbom A, Ivergård M, Compston J, Cooper C, Stenmark J, et al. (2013). Osteoporosis in the European Union: medical management, epidemiology and economic burden. *Arch Osteoporos*, 8(1-2):136.

[137] Foster SA, Shi N, Curkendall S, Stock J, Chu B.-C, Burge R, et al. (2013). Fractures in women treated with raloxifene or alendronate: a retrospective database analysis. *BMC Women Health*, 13(1):15.

[138] Park EJ, Joo IW, Jang MJ, Kim YT, Oh K, and Oh HJ (2014). Prevalence of osteoporosis in the korean population based on Korea National Health and Nutrition Examination Survey (KNHANES), 2008-2011. *Yonsei Med J*, 55(4):1049–1057.

[139] Lin T, Yan SG, Cai XZ, Ying ZM, Yuan FZ, and Zuo X (2014). Alendronate versus raloxifene for postmenopausal women: a meta-analysis of seven head-to-head randomized controlled trials. *Int J Endocrinol*, 2014.

[140] Miller P and Derman R (2010). What is the best balance of benefits and risks among anti-resorptive therapies for postmenopausal osteoporosis? *Osteoporos Int*, 21(11):1793–1802.

[141] Black DM, Cummings SR, Karpf DB, Cauley JA, Thompson DE, Nevitt MC, et al. (1996). Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. *Lancet*, 348(9041):1535–1541.

[142] Ettinger B, Black DM, Mitlak BH, Knickerbocker RK, Nickelsen T, Genant HK, et al. (1999). Reduction of vertebral fracture risk in postmenopausal women with osteoporosis treated with raloxifene: results from a 3-year randomized clinical trial. *JAMA*, 282(7):637–645.

[143] Cadarette SM, Katz JN, Brookhart MA, Stürmer T, Stedman MR, and Solomon DH (2008). Relative effectiveness of osteoporosis drugs for preventing nonvertebral fracture. *Ann Intern Med*, 148(9):637–646.

[144] Luckey M, Kagan R, Greenspan S, Bone H, Kiel R. DP, Simon J, et al. (2004). Once-weekly alendronate 70 mg and raloxifene 60 mg daily in the treatment of postmenopausal osteoporosis. *Menopause*, 11(4):405–415.

[145] Sambrook P, Geusens P, Ribot C, Solimano J, Ferrer-Barriendos J, Gaines K, et al. (2004). Alendronate produces greater effects than raloxifene on bone density and bone turnover in postmenopausal women with low bone density: results of EFFECT (Efficacy of FOSAMAX® versus EVISTA® Comparison Trial) International. *J Intern Med*, 255(4):503–511.

[146] Tanaka S, Yamamoto T, Oda E, Nakamura M, and Fujiwara S (2018). Real-world evidence of raloxifene versus alendronate in preventing non-vertebral fractures in Japanese women with osteoporosis: retrospective analysis of a hospital claims database. *J Bone Miner Metab*, 36(1):87–94.

[147] Ryan P (2013). Statistical challenges in systematic evidence generation through analysis of observational healthcare data networks. *Stat Methods Med Res*, 22(1):3–6.

[148] Jackson D, Veroniki AA, Law M, Tricco AC, and Baker R (2017). Paule-Mandel estimators for network meta-analysis with random inconsistency effects. *Res Synth Methods*, 8(4):416–434.

[149] Foster SA, Foley KA, Meadows ES, Johnston JA, Wang S, Pohl GM, et al. (2008). Characteristics of patients initiating raloxifene compared to those initiating bisphosphonates. *BMC Womens Health*, 8(1):24.

[150] Murad MH, Drake MT, Mullan RJ, Mauck KF, Stuart LM, Lane MA, et al. (2012). Comparative effectiveness of drug treatments to prevent fragility fractures: a systematic review and network meta-analysis. *J Clin Endocrinol Metab*, 97(6):1871–1880.

[151] MacLean C, Alexander A, Carter J, Chen S, Desai SB, Grossman J, et al. (2007). Comparative effectiveness of treatments to prevent fractures in men and women with low bone density or osteoporosis. *Europe PMC*.

[152] Ensrud KE, Stock JL, Barrett-Connor E, Grady D, Mosca L, Khaw K.-T, et al. (2008). Effects of raloxifene on fracture risk in postmenopausal women: the Raloxifene Use for the Heart Trial. *J Bone Miner Res*, 23(1):112–120.

[153] Wells GA, Cranney A, Peterson J, Boucher M, Shea B, Welch V, et al. (2008). Alendronate for the primary and secondary prevention of osteoporotic fractures in postmenopausal women. *Cochrane Libr*.

[154] Khosla S (2009). Increasing options for the treatment of osteoporosis. *N Engl J Med*, 361(8):818.

[155] Wysowski DK and Greene P (2013). Trends in osteoporosis treatment with oral and intravenous bisphosphonates in the United States, 2002–2012. *Bone*, 57(2):423–428.

[156] Black DM, Kelly MP, Genant HK, Palermo L, Eastell R, Bucci-Rechtweg C, et al. (2010). Bisphosphonates and fractures of the subtrochanteric or diaphyseal femur. *N Engl J Med*, 362(19):1761–1771.

[157] Kim SY, Schneeweiss S, Katz JN, Levin R, and Solomon DH (2011). Oral bisphosphonates and risk of subtrochanteric or diaphyseal femur fractures in a population-based cohort. *J Bone Miner Res*, 26(5):993–1001.

[158] Schilcher J, Michaëlsson K, and Aspenberg P (2011). Bisphosphonate use and atypical fractures of the femoral shaft. *N Engl J Med*, 364(18):1728–1737.

[159] Gedmintas L, Solomon DH, and Kim SC (2013). Bisphosphonates and risk of subtrochanteric, femoral shaft, and atypical femur fracture: a systematic review and meta-analysis. *J Bone Miner Res*, 28(8):1729–1737.

[160] Shane E, Burr D, Ebeling PR, Abrahamsen B, Adler RA, Brown TD, et al. (2010). Atypical subtrochanteric and diaphyseal femoral fractures: report of a task force of the American Society for Bone and Mineral Research. *J Bone Miner Res*, 25(11):2267–2294.

[161] Abdelmalek M and Douglas DD (1996). Alendronate-induced ulcerative esophagitis. *Am J Gastroenterol*, 91(6):1282–1283.

[162] Castell DO (1996). "Pill esophagitis" — the case of alendronate.

[163] De Groen PC, Lubbe DF, Hirsch LJ, Daifotis A, Stephenson W, Freedholm D, et al. (1996). Esophagitis associated with the use of alendronate. *N Engl J Med*, 335(14):1016–1021.

[164] Liberman UA and Hirsch LJ (1996). Esophagitis and alendronate. *N Engl J Med*, 335(14):1069–1070.

[165] Wysowski DK (2009). Reports of esophageal cancer with oral bisphosphonate use. *N Engl J Med*, 360(1):89–90.

[166] Sun K, Liu J, Sun H, Lu N, and Ning G (2013). Bisphosphonate treatment and risk of esophageal cancer: a meta-analysis of observational studies. *Osteoporos Int*, 24(1):279–286.

[167] Chen LX, Ning GZ, Zhou ZR, Li YL, Zhang D, Wu QL, et al. (2015). The carcinogenicity of alendronate in patients with osteoporosis: evidence from cohort studies. *PLoS One*, 10(4):e0123080.

[168] Seo GH and Choi HJ (2015). Oral bisphosphonate and risk of esophageal cancer: a nationwide claim study. *J Bone Metab*, 22(2):77–81.

[169] Madigan D, Ryan PB, Schuemie M, Stang PE, Overhage JM, Hartzema AG, et al. (2013). Evaluating the impact of database heterogeneity on observational study results. *Am J Epidemiol*, 178(4):645–651.

[170] Qaseem A, Forciea MA, McLean RM, and Denberg TD (2017). Treatment of low bone density or osteoporosis to prevent fractures in men and women: a clinical practice guideline update from the American College of Physicians. *Ann Intern Med*, 166(11):818–839.

[171] Martin BI, Deyo RA, Mirza SK, Turner JA, Comstock BA, Hollingworth W, et al. (2008). Expenditures and health status among adults with back and neck problems. *JAMA*, 299(6):656–664.

[172] Deyo RA, Gray DT, Kreuter W, Mirza S, and Martin BI (2005). United states trends in lumbar fusion surgery for degenerative conditions. *Spine*, 30(12):1441–1445.

[173] Wang EA, Rosen V, D'Alessandro JS, Bauduy M, Cordes P, Harada T, et al. (1990). Recombinant human bone morphogenetic protein induces bone formation. *Proceedings of the National Academy of Sciences*, 87(6):2220–2224.

[174] Axelrad TW and Einhorn TA (2009). Bone morphogenetic proteins in orthopaedic surgery. *Cytokine & Growth Factor Reviews*, 20(5-6):481–488.

[175] Cahill KS, Chi JH, Day A, and Claus EB (2009). Prevalence, complications, and hospital charges associated with use of bone-morphogenetic proteins in spinal fusion procedures. *JAMA*, 302(1):58–66.

[176] Deyo RA, Ching A, Matsen L, Martin BI, Kreuter W, Jarvik JG, et al. (2012). Use of bone morphogenetic proteins in spinal fusion surgery for older adults with lumbar stenosis: trends, complications, repeat surgery, and charges. *Spine*, 37(3):222.

[177] Ong KL, Villarraga ML, Lau E, Carreon LY, Kurtz SM, and Glassman SD (2010). Off-label use of bone morphogenetic proteins in the united states using administrative data. *Spine*, 35(19):1794–1800.

[178] Smucker JD, Rhee JM, Singh K, Yoon ST, and Heller JG (2006). Increased swelling complications associated with off-label usage of rhbmp-2 in the anterior cervical spine. *Spine*, 31(24):2813–2819.

[179] Cole T, Veeravagu A, Jiang B, and Ratliff JK (2014). Usage of recombinant human bone morphogenetic protein in cervical spine procedures: analysis of the marketscan longitudinal database. *JBJS*, 96(17):1409–1416.

[180] Vaidya R, Carp J, Sethi A, Bartol S, Craig J, and Les CM (2007). Complications of anterior cervical discectomy and fusion using recombinant human bone morphogenetic protein-2. *European Spine Journal*, 16(8):1257–1265.

[181] Carragee EJ, Hurwitz EL, and Weiner BK (2011). A critical review of recombinant human bone morphogenetic protein-2 trials in spinal surgery: emerging safety concerns and lessons learned. *The Spine Journal*, 11(6):471–491.

[182] Fu R, Selph S, McDonagh M, Peterson K, Tiwari A, Chou R, et al. (2013). Effectiveness and harms of recombinant human bone morphogenetic protein-2 in spine fusion: a systematic review and meta-analysis. *Annals of Internal Medicine*, 158(12):890–902.

[183] Epstein NE (2013). Complications due to the use of bmp/infuse in spine surgery: the evidence continues to mount. *Surgical Neurology International*, 4(Suppl 5):S343.

[184] Rodgers MA, Brown JV, Heirs MK, Higgins JP, Mannion RJ, Simmonds MC, et al. (2013). Reporting of industry funded study outcome data: comparison of confidential and published data on the safety and effectiveness of rhbmp-2 for spinal fusion. *BMJ*, 346:f3981.

[185] Lao L, Cohen JR, Lord EL, Buser Z, and Wang JC (2016). Trends analysis of rhbmp

utilization in single-level posterior lumbar fusion (plf) in the united states. *European Spine Journal*, 25(3):783–788.

[186] Lord EL, Cohen JR, Buser Z, Meisel H.-J, Brodke DS, Yoon ST, et al. (2017). Trends, costs, and complications of anterior cervical discectomy and fusion with and without bone morphogenetic protein in the united states medicare population. *Global Spine Journal*, 7(7):603–608.

[187] Martin BI, Lurie JD, Tosteson AN, Deyo RA, Farrokhi FR, and Mirza SK (2015). Use of bone morphogenetic protein among patients undergoing fusion for degenerative diagnoses in the united states, 2002 to 2012. *The Spine Journal*, 15(4):692–699.

[188] Faundez A, Tournier C, Garcia M, Aunoble S, and Le Huec J.-C (2016). Bone morphogenetic protein use in spine surgery—complications and outcomes: a systematic review. *International Orthopaedics*, 40(6):1309–1319.

[189] Guppy KH, Harris J, Chen J, Paxton EW, Alvarez J, and Bernbeck J (2016). Reoperation rates for symptomatic nonunions in posterior cervical (subaxial) fusions with and without bone morphogenetic protein in a cohort of 1158 patients. *Journal of Neurosurgery: Spine*, 24(4):556–564.

[190] Haid RW, Branch CL, Alexander JT, and Burkus JK (2004). Posterior lumbar interbody fusion using recombinant human bone morphogenetic protein type 2 with cylindrical interbody cages. *The Spine Journal*, 4(5):527–538.

[191] Laurie AL, Chen Y, Chou R, and Fu R (2016). Meta-analysis of the impact of patient characteristics on estimates of effectiveness and harms of recombinant human bone morphogenetic protein-2 in lumbar spinal fusion. *Spine*, 41(18):E1115–E1123.

[192] Paul JC, Lonner BS, Vira S, Kaye ID, and Errico TJ (2016). Use of recombinant bone morphogenetic protein is associated with reduced risk of reoperation after spine fusion for adult spinal deformity. *Spine*, 41(1):E15–E21.

[193] Simmonds MC, Brown JV, Heirs MK, Higgins JP, Mannion RJ, Rodgers MA, et al. (2013). Safety and effectiveness of recombinant human bone morphogenetic protein-2 for spinal fusion: a meta-analysis of individual-participant data. *Annals of Internal Medicine*, 158(12):877–889.

[194] Woo EJ (2012). Recombinant human bone morphogenetic protein-2: adverse events reported to the manufacturer and user facility device experience database. *The Spine Journal*, 12(10):894–899.

[195] Pugely AJ, Martin CT, Harwood J, Ong KL, Bozic KJ, and Callaghan JJ (2015). Database and registry research in orthopaedic surgery: part i: claims-based data. *JBJS*, 97(15):1278–1287.

[196] Williams BJ, Smith JS, Fu K.-MG, Hamilton DK, Polly Jr DW, Ames CP, et al. (2011). Does bone morphogenetic protein increasethe incidence of perioperative complicationsin spinal fusion?: A comparison of 55,862 cases of spinal fusion with and without bone morphogenetic protein. *Spine*, 36(20):1685–1691.

[197] Veeravagu A, Cole TS, Jiang B, Ratliff JK, and Gidwani RA (2014). The use of bone morphogenetic protein in thoracolumbar spine procedures: analysis of the marketscan longitudinal database. *The Spine Journal*, 14(12):2929–2937.

[198] Hindoyan K, Tilan J, Buser Z, Cohen JR, Brodke DS, Youssef JA, et al. (2017). A retrospective analysis of complications associated with bone morphogenetic protein 2 in anterior lumbar interbody fusion. *Global Spine Journal*, 7(2):148–153.

[199] Savage JW, Kelly MP, Ellison SA, and Anderson PA (2015). A population-based review of bone morphogenetic protein: associated complication and reoperation rates after lumbar spinal fusion. *Neurosurgical Focus*, 39(4):E13.

[200] Cahill KS, McCormick PC, and Levi AD (2015). A comprehensive assessment of the risk of bone morphogenetic protein use in spinal fusion surgery and postoperative cancer diagnosis. *Journal of Neurosurgery: Spine*, 23(1):86–93.

[201] Carragee EJ, Chu G, Rohatgi R, Hurwitz EL, Weiner BK, Yoon ST, et al. (2013). Cancer risk after use of recombinant bone morphogenetic protein-2 for spinal arthrodesis. *JBJS*, 95(17):1537–1545.

[202] Malham GM, Giles GG, Milne RL, Blecher CM, and Brazenor GA (2015). Bone morphogenetic proteins in spinal surgery: what is the fusion rate and do they cause cancer? *Spine*, 40(22):1737–1742.

[203] Sayama C, Willsey M, Chintagumpala M, Brayton A, Briceño V, Ryan SL, et al. (2015). Routine use of recombinant human bone morphogenetic protein–2 in posterior fusions of the pediatric spine and incidence of cancer. *Journal of Neurosurgery: Pediatrics*, 16(1):4–13.

[204] Mesfin A, Buchowski JM, Zebala LP, Bakhsh WR, Aronson AB, Fogelson JL, et al. (2013). High-dose rhbmp-2 for adults: major and minor complications: a study of 502 spine cases. *JBJS*, 95(17):1546–1553.

[205] Baldus C, Kelly MP, Yanik EL, Drake BF, Ahmad A, Mesfin A, et al. (2017). Incidence of cancer in spinal deformity patients receiving high-dose ($\geq$ 40 mg) bone morphogenetic protein (rhbmp-2). *Spine*, 42(23):1785–1791.

[206] Cooper GS and Kou TD (2013). Risk of cancer after lumbar fusion surgery with recombinant human bone morphogenic protein-2 (rh-bmp-2). *Spine*, 38(21):1862.

[207] Lad SP, Bagley JH, Karikari IO, Babu R, Ugiliweneza B, Kong M, et al. (2013). Cancer after spinal fusion: the role of bone morphogenetic protein. *Neurosurgery*, 73(3):440–449.

[208] Kelly MP, Savage JW, Bentzen SM, Hsu WK, Ellison SA, and Anderson PA (2014). Cancer risk from bone morphogenetic protein exposure in spinal arthrodesis. *The Journal of Bone and Joint Surgery. American Volume*, 96(17):1417.

[209] DerSimonian R and Laird N (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188.

[210] Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. (2019). Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *The Lancet*, 394(10211):1816–1826.

[211] Joseph V and Rampersaud YR (2007). Heterotopic bone formation with the use of rhbmp2 in posterior minimal access interbody fusion: a ct analysis. *Spine*, 32(25):2885–2890.

[212] Niu S, Anastasio AT, Faraj RR, and Rhee JM (2019). Evaluation of heterotopic ossification after using recombinant human bone morphogenetic protein–2 in transforaminal lumbar interbody fusion: A computed tomography review of 996 disc levels. *Global Spine Journal*, page 2192568219846074.

[213] Sebastian AS, Wanderman NR, Currier BL, Pichelmann MA, Treder VM, Fogelson JL, et al. (2019). Prospective evaluation of radiculitis following bone morphogenetic protein-2 use for transforaminal interbody arthrodesis in spine surgery. *Asian Spine Journal*.

[214] Organization WH et al. (2017). Depression and other common mental disorders: global health estimates. Technical report, World Health Organization.

[215] Cuijpers P, van Straten A, Warmerdam L, and Andersson G (2009). Psychotherapy versus the combination of psychotherapy and pharmacotherapy in the treatment of depression: a meta-analysis. *Depression and Anxiety*, 26(3):279–288.

[216] Hawton K, Witt KG, Salisbury T. LT, Arensman E, Gunnell D, Hazell P, et al. (2015). Pharmacological interventions for self-harm in adults. *Cochrane Database of Systematic Reviews*, 7.

[217] Cougnard A, Verdoux H, Grolleau A, Moride Y, Begaud B, and Tournier M (2009). Impact of antidepressants on the risk of suicide in patients with depression in real-life conditions: a decision analysis model. *Psychological Medicine*, 39(8):1307–1315.

[218] Mulder RT, Joyce P, Frampton C, and Luty SE (2008). Antidepressant treatment is associated with a reduction in suicidal ideation and suicide attempts. *Acta Psychiatrica Scandinavica*, 118(2):116–122.

[219] Hannan EL (2008). Randomized clinical trials and observational studies: guidelines for assessing respective strengths and limitations. *JACC: Cardiovascular Interventions*, 1(3):211–217.

[220] Hammad TA, Laughren T, and Racoosin J (2006). Suicidality in pediatric patients treated with antidepressant drugs. *Archives of General Psychiatry*, 63(3):332–339.

[221] Jick H, Kaye JA, and Jick SS (2004). Antidepressants and the risk of suicidal behaviors. *JAMA*, 292(3):338–343.

[222] Miller M, Pate V, Swanson SA, Azrael D, White A, and Stürmer T (2014). Antidepressant class, age, and the risk of deliberate self-harm: a propensity score matched cohort study of ssri and snri users in the usa. *CNS Drugs*, 28(1):79–88.

[223] Rubino A, Roskell N, Tennis P, Mines D, Weich S, and Andrews E (2007). Risk of suicide during treatment with venlafaxine, citalopram, fluoxetine, and dothiepin: retrospective cohort study. *BMJ*, 334(7587):242.

[224] Su K.-P, Lu N, Tang C.-H, Chiu W.-C, Chang H.-C, and Huang K.-C (2019). Comparisons of the risk of medication noncompliance and suicidal behavior among patients with depressive disorders using different monotherapy antidepressants in taiwan: A nationwide population-based retrospective cohort study. *Journal of Affective Disorders*, 250:170–177.

[225] Pozzi M, Radice S, Clementi E, Molteni M, and Nobile M (2016). Antidepressants and, suicide and self-injury: Causal or casual association? *International Journal of Psychiatry in Clinical Practice*, 20(1):47–51.

[226] Bridge JA, Iyengar S, Salary CB, Barbe RP, Birmaher B, Pincus HA, et al. (2007). Clinical response and risk for reported suicidal ideation and suicide attempts in pedi-

atric antidepressant treatment: a meta-analysis of randomized controlled trials. *JAMA*, 297(15):1683–1696.

[227] Zalsman G, Hawton K, Wasserman D, van Heeringen K, Arensman E, Sarchiapone M, et al. (2016). Suicide prevention strategies revisited: 10-year systematic review. *The Lancet Psychiatry*, 3(7):646–659.

[228] Mann JJ and Kapur S (1991). The emergence of suicidal ideation and behavior during antidepressant pharmacotherapy. *Archives of General Psychiatry*, 48(11):1027–1033.

[229] Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, and Suchard MA (2018). Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170356.

[230] Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, and Suchard MA (2016). Robust empirical calibration of p-values using observational data. *Statistics in Medicine*, 35(22):3883–3888.

[231] Lisanby SH (2007). Electroconvulsive therapy for depression. *New England Journal of Medicine*, 357(19):1939–1945.

[232] Duerden MG and Hughes DA (2010). Generic and therapeutic substitutions in the UK: Are they a good thing? *British Journal of Clinical Pharmacology*, 70(3):335–341.

[233] Haas JS, Phillips KA, Gerstenberger EP, and Seger AC (2005). Potential savings from substituting generic drugs for brand-name drugs: medical expenditure panel survey, 1997-2000. *Annals of Internal Medicine*, 142(11):891–897.

[234] Heinze G, Hronsky M, Reichardt B, Baumgärtel C, Müllner M, Bucsics A, et al. (2015). Potential savings in prescription drug costs for hypertension, hyperlipidemia, and diabetes mellitus by equivalent drug substitution in austria: a nationwide cohort study. *Applied Health Economics and Health Policy*, 13(2):193–205.

[235] Banahan BF and Kolassa E (1997). A physician survey on generic drugs and substitution of critical dose medications. *Archives of Internal Medicine*, 157(18):2080–2088.

[236] Crawford P, Feely M, Guberman A, and Kramer G (2006). Are there potential problems with generic substitution of antiepileptic drugs?: A review of issues. *Seizure*, 15(3):165–176.

[237] Kjoenniksen I, Lindbaek M, and Granas AG (2006). Patients' attitudes towards and experiences of generic drug substitution in norway. *Pharmacy World and Science*, 28(5):284–289.

[238] Riner B, Bussy A, Hélène-Pelage J, Moueza N, Lamy S, and Carrère P (2017). "no generics, doctor!" the perspective of general practitioners in two french regions. *BMC Health Services Research*, 17(1):707.

[239] Wilner AN (2004). Therapeutic equivalency of generic antiepileptic drugs: Results of a survey. *Epilepsy and Behavior*, 5(6):995–998.

[240] Dunne S, Shannon B, Dunne C, and Cullen W (2013). A review of the differences and similarities between generic drugs and their originator counterparts, including economic benefits associated with usage of generic medicines, using Ireland as a case study. *BMC Pharmacology and Toxicology*, 14.

[241] Kesselheim AS, Misono AS, Lee JL, Stedman MR, Brookhart MA, Choudhry NK, et al. (2008). Clinical equivalence of generic and brand-name drugs used in cardiovascular disease: A systematic review and meta-analysis. *JAMA - Journal of the American Medical Association*, 300(21):2514–2526.

[242] Kesselheim AS, Stedman MR, Bubrick EJ, Gagne JJ, Misono AS, Lee JL, et al. (2010). Seizure outcomes following the use of generic versus brand-name antiepileptic drugs: A systematic review and meta-analysis. *Drugs*, 70(5):605–621.

[243] Manzoli L, Flacco ME, Boccia S, D'Andrea E, Panic N, Marzuillo C, et al. (2016).

Generic versus brand-name drugs used in cardiovascular diseases. *European Journal of Epidemiology*, 31(4):351–368.

[244] Strom BL (1987). Generic Drug Substitution Revisited. *New England Journal of Medicine*, 316(23):1456–1462.

[245] Corrao G, Soranna D, Arfè A, Casula M, Tragni E, Merlino L, et al. (2014). Are generic and brand-name statins clinically equivalent? evidence from a real data-base. *European Journal of Internal Medicine*, 25(8):745–750.

[246] Gagne JJ, Choudhry NK, Kesselheim AS, Polinski JM, Hutchins D, Matlin OS, et al. (2014). Comparative effectiveness of generic and brand-name statins on patient outcomes: A cohort study. *Annals of Internal Medicine*, 161(6):400–407.

[247] Gagne JJ, Kesselheim AS, Choudhry NK, Polinski JM, Hutchins D, Matlin OS, et al. (2015). Comparative effectiveness of generic versus brand-name antiepileptic medications. *Epilepsy and Behavior*, 52:14–18.

[248] Hansen RN, Campbell JD, and Sullivan SD (2009). Association between antiepileptic drug switching and epilepsy-related events. *Epilepsy and Behavior*, 15(4):481–485.

[249] Hartung DM, Middleton L, Svoboda L, and McGregor JC (2012). Generic substitution of lamotrigine among medicaid patients with diverse indications: A cohort-crossover study. *CNS Drugs*, 26(8):707–716.

[250] Rascati KL, Richards KM, Johnsrud MT, and Mann TA (2009). Effects of antiepileptic drug substitutions on epileptic events requiring acute care. *Pharmacotherapy*, 29(7):769–774.

[251] Gothe H, Schall I, Saverno K, Mitrovic M, Luzak A, Brixner D, et al. (2015). The Impact of Generic Substitution on Health and Economic Outcomes: A Systematic Review. *Applied Health Economics and Health Policy*, 13:21–33.

[252] Heinze G, Jandeck LM, Hronsky M, Reichardt B, Baumgärtel C, Bucsics A, et al. (2016). Prevalence and determinants of unintended double medication of antihypertensive,

lipid-lowering, and hypoglycemic drugs in austria: a nationwide cohort study. *Pharmacoepidemiology and Drug Safety*, 25(1):90–99.

[253] Organization WH et al. (2006). Who collaborating centre for drug statistics methodology: Atc classification index with ddds and guidelines for atc classification and ddd assignment. *Oslo, Norway: Norwegian Institute of Public Health*.

[254] Group T et al. (2016). Postgresql: The world's most advanced open source database.

[255] Team RC et al. (2013). R: A language and environment for statistical computing. *Vienna, Austria*.

[256] Epstein M (2005). Guidelines for good pharmacoepidemiology practices (gpp). *Pharmacoepidemiology and Drug Safety*, 14(8):589–595.

[257] Winkelmayer WC, Stedman MR, Pogantsch M, Wieninger P, Bucsics A, Asslaber M, et al. (2011). Guideline-conformity of initiation with oral hypoglycemic treatment for patients with newly therapy-dependent type 2 diabetes mellitus in austria. *Pharmacoepidemiology and Drug Safety*, 20(1):57–65.

[258] Shrank WH, Liberman JN, Fischer MA, Girdish C, Brennan TA, and Choudhry NK (2011). Physician perceptions about generic drugs. *Annals of Pharmacotherapy*, 45(1):31–38.

[259] Kesselheim AS, Gagne JJ, Eddings W, Franklin JM, Ross KM, Fulchino LA, et al. (2016). Prevalence and predictors of generic drug skepticism among physicians: results of a national survey. *JAMA Internal Medicine*, 176(6):845–847.

[260] Wijk B. LV, Klungel OH, Heerdink ER, and Boer Ad (2006). Generic substitution of antihypertensive drugs: does it affect adherence? *Annals of Pharmacotherapy*, 40(1):15–20.

[261] Boh M, Opolski G, Poredos P, Ceska R, and Jezovnik M (2011). Therapeutic equivalence of the generic and the reference atorvastatin in patients with increased coronary risk.

International Angiology: a Journal of the International Union of Angiology, 30(4):366–374.

[262] Kim S.-H, Park K, Hong S.-J, Cho Y.-S, Sung J.-D, Moon G.-W, et al. (2010). Efficacy and tolerability of a generic and a branded formulation of atorvastatin 20 mg/d in hypercholesterolemic korean adults at high risk for cardiovascular disease: a multicenter, prospective, randomized, double-blind, double-dummy clinical trial. *Clinical Therapeutics*, 32(11):1896–1905.

[263] Ahrens W, Hagemeier C, Mühlbauer B, Pigeot I, Püntmann I, Reineke A, et al. (2007). Hospitalization rates of generic metoprolol compared with the original beta-blocker in an epidemiological database study. *Pharmacoepidemiology and Drug Safety*, 16(12):1298–1307.

[264] Pruckner GJ and Schober T (2018). Hospitals and the generic versus brand-name prescription decision in the outpatient sector. *Health Economics*, 27(8):1264–1283.

[265] Robb MA, Racoosin JA, Sherman RE, Gross TP, Ball R, Reichman ME, et al. (2012). The us food and drug administration's sentinel initiative: expanding the horizons of medical product safety. *Pharmacoepidemiology and Drug Safety*, 21:9–11.