New Solutions for Real-World Causal Inference



Xi Lin

St Cross College

University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

June 2025

To those who couldn't be here to see this,
I carry your memory with me.

Acknowledgements

I would like to thank my supervisor, Prof. Robin Evans, for his guidance and support over the past four years, and for always being there when I needed him. He is simply the best supervisor that I could have ever hoped for.

I am thankful to my collaborator and friend, Jens Magelund Tarp, for our great catchups, his brilliant ideas, and for always cheering me up when things were tough.

Thanks also to my colleagues in the Department of Statistics, at Novo Nordisk, and in the Joint Initiative of Causal Inference for many thoughtful discussions and helpful comments. I am thankful to my friends in Oxford for sharing these exciting and memorable years with me.

I am forever grateful to my parents, Hongyu Jia and Pengjiu Lin. I would not be who I am without them. Thank you for always believing in me, supporting all my aspirations, and loving me unconditionally.

To the love of my life, Kane—thank you. Whenever I feel low, you are there, lifting the clouds and filling the day with light. You are my sunshine.

Abstract

Causality can be pragmatically defined as *implementable knowledge*—knowledge that informs decisions about actions. From this perspective, causal inference is the task of learning the effects of interventions from data. As causal methods are increasingly deployed in complex, real-world settings, new challenges emerge. This thesis presents new solutions to real-world causal inference, addressing three such problems: limited trial data, method selection for data fusion, and simulation of longitudinal data under time-varying treatment and confounding.

The first contribution develops a Bayesian power-likelihood approach to augment underpowered trials with real-world data. By calibrating the influence of external data using posterior predictive performance, the method improves statistical efficiency while managing confounding bias.

The second contribution systematically evaluates contemporary data fusion methods. Using simulations under controlled violations of key assumptions, we compare their empirical performance, revealing the trade-offs between bias and variance in finite samples. We further implement these methods in a real-world case study augmenting the PIONEER 6 cardiovascular outcomes trial with US health insurance claims data. This provides a comprehensive demonstration of data fusion in practice and guidance for method selection.

The third contribution addresses the challenge of simulating longitudinal data from marginal structural models. We propose an algorithm that enables exact simulation from a specified marginal model. Unlike existing approaches, our method directly targets the causal margin, avoids restrictive assumptions, and scales efficiently to realistic longitudinal settings involving time-varying treatments and survival outcomes.

Collectively, these contributions extend the methodological and practical reach of causal inference. While developed in collaboration with clinical researchers, the proposed methods are broadly applicable to other

domains where robust causal conclusions are essential for guiding high-stakes decisions.

Contents

1 Introduction **1**

1.1 Statistical frameworks for causal inference 2

1.2 Motivations and contributions 4

1.2.1 Enhancing efficiency through data fusion 5

1.2.2 Evaluating and selecting data fusion methods 9

1.2.3 Simulating longitudinal data from causal models 11

1.3 Preliminaries 14

1.3.1 Notation 14

1.3.2 Causal graphical models 14

1.3.3 Potential outcomes 18

1.3.4 Marginal structural models 22

1.3.5 Frugal parameterization 28

1.4 Summary and roadmap 30

2 Combining experimental and observational data through a power likelihood **31**

2.1 Introduction 32

2.2 Related literature 33

2.3 Setup, notation and assumptions 35

2.4 A power-likelihood approach 38

2.4.1 Frugal parameterization 38

2.4.2 Choosing the optimal influence factor η 40

2.4.3 Normal approximations of the posterior distribution 43

2.4.4 Asymptotics 44

2.5 Simulation 45

2.5.1 Moderating the bias-variance trade-off 45

2.5.2 Inference 47

2.6 PIONEER 6 application 47

2.6.1 Target trial emulation 48

2.6.2 Design 1: Incorporating control arm only 49
2.6.3 Design 2: Incorporating both arms 49
2.7 Conclusion and future work 51

3 Evaluating data fusion methods 54
3.1 Introduction 55
3.2 Existing methods 57
3.2.1 Test-then-pool 57
3.2.2 Weighted combination 59
3.2.3 Shrinkage 61
3.2.4 Bias correction 63
3.2.5 Prognostic adjustment 64
3.2.6 A summary table for application 65
3.3 Simulation studies 68
3.3.1 Simulation setup 68
3.3.2 Augmenting both arms of the RWD 69
3.3.3 Augmenting only the control arm of the RWD 78
3.4 Real-world application: PIONEER 6 Trial 82
3.4.1 Target trial emulation 83
3.4.2 Initial analysis: PIONEER 6 alone 85
3.4.3 Design 1: Incorporating control arm only 87
3.4.4 Design 2: Incorporating both arms 87
3.5 Summary 88

4 Exact simulation of longitudinal data from marginal structural models 91
4.1 Introduction 92
4.2 Related literature 94
4.3 Copulas 95
4.3.1 Pair-copula construction 96
4.3.2 Simulating from pair-copula constructions 97
4.4 Simulation algorithm 98
4.4.1 The fully-conditional approach 99
4.4.2 Frugal parameterization for outcome simulation 99
4.4.3 Simulation algorithm 101
4.4.4 Illustrative example 102
4.5 Simulating survival outcomes 103
4.5.1 Simulation algorithm 106

4.5.2 Illustration: Concomitant treatment . . . 111
4.5.3 Competing risks . . . 115
4.6 Discussion . . . 118

5 Conclusion **119**
5.1 Summary . . . 119
5.2 Future work . . . 120
5.3 A many-data power-likelihood approach . . . 120
5.4 Simulating longitudinal survival data with competing events from MSMs targeting the cause-specific hazard . . . 121
5.5 Realistic emulation of trial data for power and sample size calculations 122

A Philosophical foundations of causality **124**

B Appendix for Chapter 2: Combining experimental and observational data through a power likelihood **126**
B.1 Proof of Lemma 2.1 . . . 126
B.2 Proof of Theorem 2.1 . . . 127
B.3 Proof of Proposition 2.2 . . . 128

C Appendix for Chapter 3: Data fusion for efficient average treatment effect estimation **131**
C.1 Implementation details of methods in simulation studies . . . 131
C.1.1 Baseline IPTW estimators . . . 131
C.1.2 Shrinkage estimators . . . 132
C.1.3 Bias-correction methods . . . 133
C.1.4 Power-likelihood method . . . 134
C.1.5 Test-then-pool methods . . . 135
C.1.6 Weighted combination methods . . . 135
C.1.7 Prognostic adjustment . . . 136
C.2 Details of real-world application . . . 136
C.2.1 Target trial emulation . . . 136
C.2.2 Implementation of methods . . . 148

D Appendix for Chapter 4: Exact simulation of longitudinal data from marginal structural models **154**
D.1 Extension to include multiple time-varying variables . . . 154

D.2 Proof of Lemma 4.1 156
D.3 Proof of Proposition 4.1 157
D.4 Additional details of the illustration in Section 4.5.2 158
D.5 Seaman and Keogh (2024) simulations 160

Bibliography **164**

Chapter 1

Introduction

Causal inference seeks to answer questions about the effects of actions based on observed data. In contrast to statistical association, which describes patterns in data, causal inference aims to uncover structures—offering deeper insight that can inform scientific discoveries, shape policies, and support individual and institutional actions. It asks not merely what is associated with what, but rather, *what would happen if we did something differently?*

The study of causality has occupied philosophers for millennia. Yet only in recent decades have formal statistical frameworks been developed that make causal inference tractable and operational in empirical research. Reflecting its impact, Vehtari et al. (2020) ranked causal inference first among the eight most important statistical ideas of the past 50 years. These advances have greatly expanded our ability to extract meaningful conclusions from data, and to design studies that yield robust, actionable answers to causal questions.

Real-world applications continue to pose new challenges: As causal inference becomes more widely adopted and applied in complex settings, new methodological demands emerge. This thesis addresses two such problems: (i) how to combine experimental and real-world data to improve the efficiency of causal effect estimates; and (ii) how to simulate longitudinal data from marginal structural models in a way that preserves causal structure.

The work presented in this thesis is driven by a persistent habit of inquiry. We begin by asking the most fundamental questions: What is causality? Section 1.1 introduces the key statistical frameworks that have emerged from, and been shaped by, these philosophical perspectives. Section 1.2 discusses the real-world challenges that motivate this thesis and examines whether these are problems of practical relevance or merely theoretical curiosities. Section 1.3 presents the notation and core concepts that underpin the methods and results developed in the chapters that follow.

Contents

1.1 Statistical frameworks for causal inference 2
1.2 Motivations and contributions 4
1.2.1 Enhancing efficiency through data fusion 5
1.2.2 Evaluating and selecting data fusion methods 9
1.2.3 Simulating longitudinal data from causal models 11
1.3 Preliminaries 14
1.3.1 Notation 14
1.3.2 Causal graphical models 14
1.3.3 Potential outcomes 18
1.3.4 Marginal structural models 22
1.3.5 Frugal parameterization 28
1.4 Summary and roadmap 30

Causality has long been a subject of philosophical debate, shaping how we understand knowledge, explanation, and intervention. In *Physics*, Aristotle presented a pluralistic theory of causation, distinguishing between material, formal, efficient, and final causes (Aristotle, trans. Hardie and Gaye, 1930). Centuries later, David Hume challenged the notion of causality as an objective feature of the world. In *A Treatise of Human Nature* (1739) and *An Enquiry Concerning Human Understanding* Hume (1748), he argued that causal belief arises from psychological habit formed through the repeated observation of *constant conjunctions*. Immanuel Kant, in Critique of Pure Reason Kant (1781), responded by asserting that causal reasoning is a necessary *a priori* condition for making sense of empirical experience. In the twentieth century, David Lewis advanced a counterfactual theory of causation, formalizing the idea of causality as difference-making: a cause is something without which the effect would not have occurred (Lewis, 1973). These philosophical accounts continue to inform contemporary approaches to causal inference, including the potential outcomes framework and structural causal models. A more detailed discussion of these foundations is provided in Appendix A.

1.1 Statistical frameworks for causal inference

While Hume's critique of *constant conjunction* warns against equating observed associations with causation, modern causal methodologies reject radical skepticism by formalizing the conditions under which causal claims can be justified. The standard

problem for causal inference in statistics is concerned with analyzing observational data in order to infer, and often quantify, causal relations. Rigorous causal inference requires explicit assumptions that can be formally stated, thoroughly understood, and properly addressed in study designs and analysis.

Since the 20th century, multiple conceptual frameworks or *languages* for causal inference have emerged, particularly within the fields of statistics, computer science, and economics. The two most prominent paradigms are:

- The **Neyman-Rubin Potential Outcomes (PO) framework** (Neyman, 1923; Rubin, 1974), defines causal effects as contrasts between counterfactual states under hypothetical interventions. This philosophically aligns with Lewis's counterfactual theory of causality. It provides a conceptual basis for defining causal estimands such as the average treatment effect (ATE), and motivates identification strategies like exchangeability and consistency.

- **Causal graphical models** (Spirtes et al., 2000) and the ***do*-approach** represent causal relationships via directed acyclic graphs (DAGs), where variables are shown as nodes, and arrows between them represent assumptions about causal direction. While the structure of a DAG can capture conditional independences among variables, it does not imply causality on its own. To move from mere association to causal interpretation, we must distinguish between *seeing* and *doing*: observing a variable versus actively intervening the system. Pearl (2009) formalized this distinction through the introduction of the $do(\cdot)$-notation, which provides a way to reason about interventions and their effects within a graphical model.

These frameworks are formally equivalent: theorems in one domain translate to the other. A recent advance, Single World Intervention Graphs (SWIGs) (Richardson and Robins, 2013), further unifies these approaches by embedding interventions and counterfactual outcomes within graphical models through node-splitting. Other frameworks include structural equation models (Goldberger, 1972), which specify causal relationships through algebraic equations, and the decision-theoretic approach (Dawid, 2021), which avoid counterfactuals altogether and makes as few assumptions as possible.

This thesis adopts a pragmatic view on causal inference, treating causality as *implementable knowledge*—knowledge that can inform real-world *decisions* about *actions*. Under this perspective, a causal effect is defined as the difference—whether in specific outcomes or their distributions—resulting from hypothetical interventions

that could, at least in principle, be implemented. Examples of such causal questions include:

- Is an influenza vaccine effective in preventing hospitalization? (e.g. Tenforde et al., 2021)
- Does reducing class size or increasing the teacher-to-student ratio improve academic performance? (e.g. Krueger and Whitmore, 2001)
- How do different get-out-to-vote messages influence voter turnout? (e.g. Gerber et al., 2008)

In these cases, the interventions—administering a medication, altering class sizes, or varying voter mobilization strategies—are well-defined and actionable. However, some variables pose challenges in defining meaningful interventions. For instance, while gender or age is sometimes treated as a causal variable, such attributes are generally not manipulable in a straightforward way, unless the focus is on causal fairness—an important but distinct topic outside the scope of this thesis. Similarly, body mass index (BMI) presents an ambiguous intervention: does one intend to modify BMI through exercise, dietary restriction, or pharmacological treatment? Each pathway carries distinct implications for both causal inference and policy decisions. Clarifying the nature of interventions is thus essential in causal analysis for real-world problems.

1.2 Motivations and contributions

Establishing a cause-effect relationship requires distinguishing true causal effects from spurious associations. While a formal definition of *identifiability* and its assumptions will be provided later, we begin with a simple motivating example: *What is the effect of influenza vaccination on the risk of hospitalization?*. Suppose we attempt to answer this question using electronic health record (eHR) data.

Figure 1.1 illustrates the relevant causal structure as a DAG. The association between vaccination and hospitalization is *confounded* by underlying health behaviours—such as regular exercise, good hygiene, and a healthy diet—which are typically unmeasured in administrative records. Individuals who engage in such behaviours are both more likely to receive the influenza vaccine and less likely to experience severe illness requiring hospitalization. As a result, vaccinated and unvaccinated groups differ systematically in ways not captured in the data. A naïve

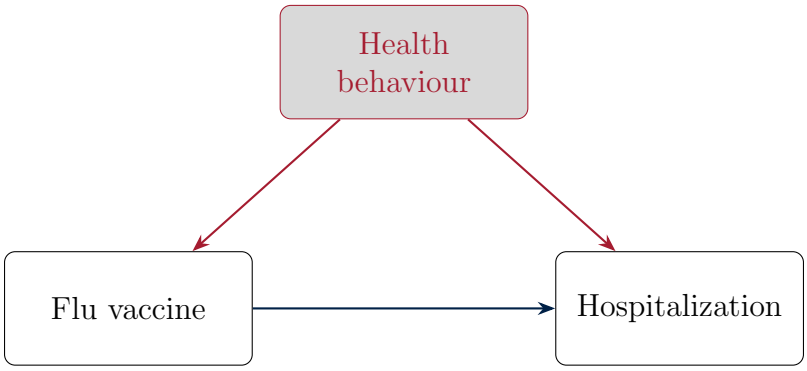


Figure 1.1: Causal structure showing confounding of the effect of flu vaccine on hospitalization outcome by health-seeking behaviour, which may differ between those who do and do not receive vaccination.

comparison of hospitalization outcomes between these groups would therefore be biased, likely overstating the vaccine's protective effect.

From a study design perspective, if we could break the correlation between vaccination and heath behaviours, we could eliminate this spurious association between vaccination and hospitalization stemming from their common cause. Theoretically, one way to achieve this is through a randomized experiment in which participants are assigned to receive the vaccine or not, independent of their preferences or characteristics. With randomization, vaccination status is no longer correlated with health behaviours. Loosely speaking, the vaccinated and unvaccinated groups become comparable, and any differences in outcomes can be attributed to the causal effect of the vaccine itself. This ability to eliminate confounding bias is what makes RCTs the gold standard for causal inference.

While randomization addresses confounding, it is not always feasible. In the influenza vaccine example, conducting a randomized trial may be impractical or ethically contentious, especially where vaccination is already standard care. More broadly, real-world decision-making often occurs in a context where trials are unavailable, underpowered, or delayed. In such cases, the ability to supplement or extend trial findings with observational data can be crucial—not just for feasibility, but for statistical efficiency: improving precision, reducing variance, and drawing stronger conclusions from limited experimental evidence.

1.2.1 Enhancing efficiency through data fusion

Experimental and observational data represent fundamentally different modes of data collection. Experimental data, such as those generated by RCTs, are collected

under controlled conditions with interventions on treatment assignment. As illustrated in the influenza vaccine example, randomization isolates the causal effect from unwanted and potentially unobserved confounding factors, with the experimental protocol ensuring—in expectation—balanced control and treatment groups. This design supports relatively simple and well-established estimators—such as outcome regression, inverse propensity weighting (IPW) (Rosenbaum and Rubin, 1983; Robins et al., 1994; Bang and Robins, 2005; Cao et al., 2009) and matching (Rubin, 1973; Hirano et al., 2003; Abadie and Imbens, 2016)—to consistently estimate causal effects. RCTs are widely regarded as the gold standard for causal inference in medicine, economics, and public policy. Regulatory agencies, such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA), rely primarily on RCT evidence to assess safety and efficacy during drug approval processes.

In practice, however, we often cannot rely solely on trials. High costs and logistical complexity can lead to limited sample sizes, restricted follow-up durations, and delayed evidence generation. Many are statistically underpowered, particularly when aiming to detect modest effects or evaluate subgroup-specific outcomes. In rare diseases, such as male breast cancer, recruiting adequate participants for adequately powered trials may be infeasible (e.g. Wedam et al., 2020). Ethical barriers may also preclude randomization. For example, in the 1990s, placebo-controlled zidovudine (AZT) trials conducted in sub-Saharan Africa to prevent mother-to-child HIV transmission in sub-Saharan withheld a known effective treatment from participants in the control arm. These trials were widely criticised and prompted major revisions to international research ethics guidelines (Lurie and Wolfe, 1997; Angell, 1997). In other high-risk settings—such as advanced or rapidly progressing disease—randomization may be ethically unacceptable when no standard-of-care comparator exists, resulting instead in single-arm designs. Moreover, RCTs frequently fail to capture long-term outcomes. For example, online A/B testing in the technology sector typically runs only for a few weeks, an insufficient window to observe durable behavioural changes. Similarly, social interventions such as job training programs may yield meaningful outcomes only after several years, far beyond the horizon of a typical trial. Moreover, restrictive eligibility criteria in trials may limit the external validity of their findings, raising concerns about their generalizability to real-world populations.

By contrast, observational data, also referred to as real-world data (RWD), are collected passively in the course of routine clinical, administrative, or behavioural processes, without researcher-imposed intervention. Examples include eHRs, bank

transactions, and digital trace data from online platforms. Today, these sources are increasingly available at scale and offer extended follow-up, broader population coverage, and rich covariate information. However, they lack the protection of randomization and are susceptible to unmeasured confounding, as in the earlier example of vaccination and hospitalization. This bias does not diminish with sample size and may persist regardless of the amount of available data (Pearl, 2009). Due to such limitations, observational data have historically been viewed as inferior to RCTs for causal inference.

This motivates *data fusion* (Bareinboim and Pearl, 2016), the principled integration of experimental and observational data sources to leverage their respective strengths. From a methodological perspective, the appeal of data fusion lies in its potential to improve precision, extend follow-up, enable subgroup analysis, and increase generalizability. These methods have garnered significant attention in recent methodological literature—and increasingly, from regulators and policymakers.

Regulatory support for real-world data integration is growing. The U.S. Food and Drug Administration (FDA) has outlined a comprehensive framework for real-world evidence (U.S. Food and Drug Administration, 2018), and issued guidance on using RWD for regulatory decision-making (U.S. Food and Drug Administration, 2021; U.S. Food and Drug Administration, 2023). The European Medicines Agency (EMA) has launched the DARWIN EU® initiative to coordinate the use of RWD across Europe and published guidance on its application (European Medicines Agency, 2021). Most recently, the International Council for Harmonisation (ICH) endorsed a 2024 reflection paper highlighting opportunities to harmonise the use of RWD to generate real-world evidence (RWE), with a particular focus on evaluating treatment effectiveness (ICH, 2024). These developments reflect a broader shift: from evaluating interventions solely under idealized conditions toward supporting decisions with evidence that reflects real-world populations and practice.

This thesis focuses primarily on **data fusion for efficiency gain**. When trial data alone are insufficient to support confident decisions, incorporating RWD can improve precision and stability of effect estimates. For example, during the COVID-19 pandemic, observational studies using administrative data played a complementary role in assessing vaccine safety and effectiveness when trial evidence was limited or slow to accumulate (Hansen et al., 2021; Pottegård et al., 2020). In such cases, observational data are not a substitute for trials, but a statistical supplement that helps extract more value from existing randomized evidence.

However, this integration is not without risk. Naïve pooling of data sources can introduce serious bias if observational data are confounded or systematically

different from trial populations. The key challenge is to develop robust methods that leverage the size and richness of observational data to improve precision, while protecting against bias through careful modelling and calibration.

1.2.1.1 A power-likelihood approach

This thesis contributes to data fusion for efficiency gain by developing a method to rigorously augment randomized trial data with real-world evidence. Incorporating real-world data (RWD) can substantially improve the statistical power of treatment effect estimation, particularly when trial data alone are insufficient, either for the overall population or for under-represented subgroups.

A motivating example is the PIONEER 6 trial, a randomized study designed to assess the cardiovascular safety of semaglutide. The trial enrolled 3,183 participants and was sufficiently powered to prove safety (demonstrate non-inferiority). However, it was underpowered to detect superiority (Husain et al., 2019). To address this, a subsequent trial, SOUL, enrolled nearly 10,000 participants and extended follow-up to over four years, ultimately concluding a significant reduction in major adverse cardiovascular events (MACE). While SOUL provided definitive evidence, it reported findings more than five years after the conclusion of PIONEER 6. In the interim, large-scale administrative datasets, such as health insurance claims, could have been used to augment PIONEER 6, offering earlier indications of benefit and informing clinical or policy decisions, even if not regulatory ones.

This example motivated us to develop a methodology that rigorously augments a trial with evidence from RWD. The Bayesian paradigm offers a natural mechanism to incorporate external information. A rich literature on *Bayesian dynamic borrowing* exists, particular in the context of historical control data. Fundamental to this class of methods is the *power prior* (Ibrahim and Chen, 2000) which has since been extended through the *normalized power prior* (Duan et al., 2006; Neuenschwander et al., 2009), *commensurate power prior* (Hobbs et al., 2011, 2012), and the *meta-analytic-predictive prior* (Schmidli et al., 2014).

Despite these advances, existing methods face practical limitations. Most require pre-specifying hyperparameters, such as power parameters or *commensurability* functions, which can be difficult to justify in advance. In addition, to date, many applications have been limited to either aggregate-level data or overly simplified data-generating models (e.g. Rietbergen et al., 2011), limiting their relevance for modern, individual-level causal inference problems.

To address these gaps, Chapter 2 introduces a *data-adaptive power likelihood* method. This method enhances the traditional power prior framework along several

dimensions. First, it determines the extent of borrowing from external data empirically through a *quasi-empirical Bayes* procedure, eliminating the need to pre-specify a power parameter. Second, the method accommodates more flexible causal models based on individual participant-level data, enabling application in more realistic settings. Third, it incorporates approximation techniques to ensure computational scalability, allowing the method to be applied efficiently to large observational datasets. The method offers two further practical advantages. It allows flexible integration of either the external control arm alone or both treatment and control arms of the RWD—an important feature when, for example, a drug which is under investigation is not yet available on the market. Additionally, as a Bayesian approach, it provides coherent uncertainty quantification via posterior credible intervals.

1.2.2 Evaluating and selecting data fusion methods

The power-likelihood approach developed in Chapter 2 contributes to the broader objective of improving statistical efficiency through data fusion. However, it is one of a growing number of methods proposed in the literature. The past decade has seen a rapid expansion of techniques aimed at integrating randomized and observational data for causal inference, each offering different trade-offs between robustness, efficiency, and complexity. This raises a fundamental practical question: *how should researchers choose among these methods in a real-world setting?*

While the goal of leveraging RWD to improve efficiency is clear, the central challenge lies in managing the confounding bias inherent in non-randomized data. To reduce bias, some approaches restrict attention to the control arm of the RWD, treating it as an external or historical comparator. Common strategies in this class include test-then-pool designs (Viele et al., 2014), Bayesian dynamic borrowing methods (Ibrahim and Chen, 2000; Hobbs et al., 2012; Schmidli et al., 2014), and prognostic score adjustment (Schuler et al., 2021). These methods seek to harness information from RWD while preserving internal validity by avoiding direct incorporation of potentially confounded treatment outcomes.

Other methods aim to incorporate both arms of the RWD to achieve greater efficiency. In doing so, they confront the core trade-off: incorporating more data reduces variance, but increases the risk of bias due to confounding. Various methodologies seek to optimize this trade-off from different angles, including James-Stein style shrinkage estimators (Rosenman et al., 2023a; Green et al., 2005), estimation under the general semi-parametric framework (Yang et al., 2020; Li and Luedtke, 2023; Parikh et al., 2023), weighted average of treatment effect estimates from the RCT and the RWD (Oberst et al., 2023; Chen et al., 2021) and bias-correction

(Kallus et al., 2018; Yang and Ding, 2020; Yang et al., 2020; Cheng and Cai, 2021; Wu and Yang, 2022; van der Laan et al., 2024).

Although we now have a diverse and sophisticated toolbox of data fusion methods, the practical question of which method to use remains challenging. It is often unclear how these methods compare under realistic sample sizes and data structures, and most theoretical analyses focus on asymptotic properties. While such theory is indispensable, the whole motivation for data fusion is that we do not have access to infinite randomized samples. What we need is insight into how these methods perform with finite data, in settings that resemble real-world applications.

Several excellent surveys of the literature exist (Colnet et al., 2024; Shi et al., 2023; Brantner et al., 2023), but they tend to focus on abstract classification—e.g., whether the method is parametric or nonparametric, which types of data it accommodates, or what identification assumptions it requires. These reviews are valuable, but they typically do not evaluate methods from the standpoint of empirical performance under finite samples, nor do they emphasize efficiency as the primary objective. In methodological papers, simulation studies often appear only as demonstrations and are designed to showcase the strengths of the proposed method—this is understandable. However, from a user's perspective, particularly in regulatory or policy settings where analysis plans must be pre-specified, it is essential to understand:

- What is the behaviour of the methods with finite samples?
- How do the methods manage the trade-off between efficiency gains and introduced bias?
- How do the methods compare on a level playing field—that is, under shared data-generating models? What are their respective strengths and weaknesses?

In Chapter 3, we address these questions through a systematic evaluation of data fusion methods aimed at efficiency gains. We go beyond theoretical comparison by conducting an extensive simulation study, placing all methods on a "level playing field". That is, we apply each method to the same set of simulated data to reveal their empirical properties—such as bias, variance, and coverage—under realistic conditions.

One might object that these methods operate under different assumptions on the data-generating process, so direct comparison may be unfair. This is a valid point. However, it is precisely this kind of controlled testing that helps us understand *which assumptions are critical*, *how robust each method is*, and *what happens when key*

assumptions are violated, especially when most of these assumptions are untestable. When simulating data, we have the advantage of knowing the ground truth, which allows us to benchmark how well each method recovers the true treatment effect, under which circumstances it succeeds or fails, and how far off it gets when it fails.

We also apply the reviewed methods to a real-world case study: the PIONEER 6 cardiovascular safety trial, which originally motivated our work on data fusion. This application is carried out end-to-end. We begin by constructing an emulated RWD sample, apply various methods, and interpret the results in context. This exercise illustrates how the methods can be implemented in practice and provides insight into their comparative performance in an applied setting.

Beyond the specific findings, this chapter highlights the central role of simulation in methodological research. Simulation provides a controlled environment for examining finite-sample behaviour, evaluating robustness to assumption violations, and comparing competing methods under transparent and reproducible conditions. To simulate realistic data, however, we must generate samples that conform to the causal model of interest. While this may sound straightforward, it is not always easy—particularly when the target model is specified in terms of marginal causal quantities, such as those in marginal structural models (MSMs) (Evans and Didelez, 2024). As we show in Chapter 4, simulating data from MSMs presents technical challenges that are not well handled by standard tools. This challenge motivates our final contribution: a new method for simulating longitudinal data from causal models that directly targets the marginal structure.

1.2.3 Simulating longitudinal data from causal models

Longitudinal data play a central role in causal inference when treatments or exposures—and their effects—evolve over time. Such data are essential in many fields, including medicine, public health, and the social sciences, where the timing, duration, and sequencing of interventions matter. Unlike cross-sectional studies, longitudinal designs capture the dynamic nature of treatment regimens, which often vary over time due to individual behaviours, institutional policies, environmental conditions, or other time-dependent factors. Even in clinical trials with strict adherence protocols, intercurrent events—such as treatment discontinuation or supplementary medication use—are common. For instance, participants assigned to the treatment arm may stop taking the intervention due to side effects, while those in the control arm may initiate treatment if their condition deteriorates. In observational settings, treatment initiation or cessation can occur at any time, contingent on changing

circumstances. The ICH E9[1] emphasizes that intercurrent events must be explicitly considered when defining the clinical question of interest (European Medicines Agency, 2020). While the intention-to-treat (ITT) principle simplifies this to some extent by reducing time-varying treatment to a point-treatment problem, it may obscure the effects of intercurrent events, and thus may not yield the most clinically meaningful or policy-relevant estimand. By accounting for these treatment dynamics, longitudinal data enable the estimation of meaningful causal quantities, such as the *on-treatment* estimand, which quantifies the treatment effect assuming treatment is administered as intended.

A central challenge in analyzing longitudinal data is addressing time-varying confounding, where a time-dependent covariate predicts both the outcome and future treatments, while itself being influenced by past treatments. For example, in the case study by Robins et al. (2000) on the effect of zidovudine (AZT) treatment on HIV infection, the biomarker CD4 lymphocyte count is both affected by prior AZT treatment and is predictive of future therapy initiation and HIV progression. The presence of such time-varying confounding renders conventional adjustment methods, like regression and stratification, generally biased.

Marginal structural models, introduced by Robins (1997), have emerged as a powerful tool for estimating causal effects under time-varying confounding. Model parameters can be consistently estimated using inverse probability of treatment weighting (Robins, 1999, 2000). Among alternative methods, including structural nested models estimated through g-estimation (Robins, 1989, 1992; Robins and Tsiatis, 1991), marginal structural models have gained prominence due to their interpretability and ease of implementation. A systematic review by Clare et al. (2019) highlights their widespread use in applied research.

As causal inference methods for longitudinal data are continually proposed to address complex real-world settings, simulation studies are essential to evaluate their effectiveness for practical use. These studies assess important properties such as bias, confidence interval coverage, relative efficiency, and robustness to violations of the assumptions used, guiding both methodological development and application.

Sample size calculations also benefit from simulations. A motivating example comes from the SOUL trial, a placebo-controlled study evaluating the cardiovascular efficacy of semaglutide (McGuire et al., 2023, 2025). As in many modern trials, a key design challenge of SOUL arises from the permitted use of concomitant glucose-lowering medications where necessary, such as SGLT2 inhibitors (SGLT2-i), which

[1]International Council on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use [ICH] Statistical Principles for Clinical Trials [E9]

were more frequently taken by participants in the placebo arm. The uneven uptake of the *drop-in* treatment between the semaglutide and placebo arms may dilute the observed intention-to-treat effect, because both groups end up receiving some active treatment, blurring the distinction between them, and potentially undermining statistical power if unaccounted for in sample size calculation. Although standard formulas may suffice for simple, time-fixed treatments, more complex scenarios with time-varying treatments make deriving analytical formulas impractical or impossible. In such cases, simulation provides a flexible and practical solution, allowing researchers to estimate sample sizes based on realistic data-generating processes and causal structures.

A fundamental requirement for both method evaluation and sample size determination is the ability to simulate data that conform to a specified causal model. However, simulating longitudinal data from a pre-specified marginal structural model is not a straightforward task. While generating data sequentially from structural equations—creating covariates, treatment status, and then the outcome conditioned on past information—may seem intuitive, these models rely on conditional rather than marginal specifications. This makes it difficult to sequentially simulate from conditional distributions that align with a pre-specified marginal model; in particular, it is extremely challenging to sample from a null hypothesis—the so-called *g-null paradox*.

Several methods have been proposed for simulating data from MSMs (Xiao et al., 2010; Young and Tchetgen Tchetgen, 2014; Sofrygin et al., 2017; Keogh et al., 2021; Seaman and Keogh, 2024). However, existing approaches are limited in important ways: some do not simulate directly from the MSM but rather from underlying structural models that only approximate the marginal model under restrictive assumptions; others rely on computationally intensive Monte Carlo approximations, making it hard to scale, or are difficult to configure and reverse-engineer for specific target parameters.

These limitations motivate the development of a new simulation approach, introduced in Chapter 4. Our proposed method allows for exact simulation from a specified MSM, supports multiple time-varying covariates, treatments, outcomes, and censoring processes, and offers precise control over the distributions of these variables. Importantly, it avoids restrictive assumptions on the data-generating process, scales well with the number of time points and variables, and is designed for practical application in realistic longitudinal settings.

1.3 Preliminaries

1.3.1 Notation

We start by clarifying the general notation used throughout this thesis. We use A to represent treatment, and for illustration of ideas, we often default A to be binary with $A = 1$ representing treatment, although many results, specially in Chapter 3 and Chapter 4 are not limited to binary treatments. Y represents outcome. Depending on the specific context, we use Z ,C ,L and U, to denote covariates. In longitudinal settings, mainly Chapter 4, we use subscript to indicate time. For example, L_k denotes the value of L observed at time point k. To indicate the history of a variable up to time k, we use the bar notation. For instance, $\bar{A}_k = \{A_0, \ldots, A_k\}$ and $\bar{L}_k = \{L_0, \ldots, L_k\}$, with the convention $\bar{A}_{-1} = \emptyset$ and $\bar{L}_{-1} = \emptyset$.

Random variables are represented by uppercase letters (e.g. $\boldsymbol{X}$, Y) and their instantiations by the corresponding lowercase letters (e.g. $\boldsymbol{x}$, y). Greek letters (e.g. $\boldsymbol{\theta}, \phi$) typically denote model parameters, and bold letters (e.g. $\boldsymbol{Z}$, $\boldsymbol{\theta}$) indicate vectors, as opposed to scalars. We loosely refer to p as distribution or probability, but let it also stand for either the probability density or probability mass function of a distribution.

Additional notation may be introduced in later chapters as needed.

1.3.2 Causal graphical models

This section introduces causal graphical models as a framework for representing assumptions about causal structure and for reasoning about interventions. It is organized into three parts: first, we define directed acyclic graphs (DAGs) and their factorization properties; second, we discuss how DAGs acquire causal interpretation via the do-operator; and finally, we present graphical rules for identifying causal effects, including the *back-door* and *front-door* criteria. The notations and definitions used in this section are informed by the exposition of Didelez (2018).

1.3.2.1 Directed acyclic graphs

A *directed graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a finite set of *vertices* $\mathcal{V}$, and a collection of ordered pairs of distinct vertices, $\mathcal{E}$, these ordered pairs are called *edges* or *arrows*. We write $v \to w$ to denote an arrow from vertex v to vertex w in $\mathcal{V}$, and say that v is a parent of w, or dually, that w is a child of v. The set of all parents of a node $w \in \mathcal{V}$ is denoted by $\mathrm{pa}_{\mathcal{G}}(w) := \{v \in \mathcal{V} : v \to w \in \mathcal{G}\}$.

Two vertices v and w are adjacent if either $v \to w$ or $w \to v$. A path in $\mathcal{G}$ is a sequence of distinct vertices such that each adjacent pair in the sequence is adjacent

in $\mathcal{G}$. A path is *directed* if all arrows point in the same direction, away from the starting point.

A directed graph contains a *directed cycle* if there exists a directed path from v to w, along with an edge $w \to v$. A directed graph is *acyclic* if it contains no directed cycles. We refer to such graphs as *directed acyclic graphs* (DAGs), for example, Figure 1.2.

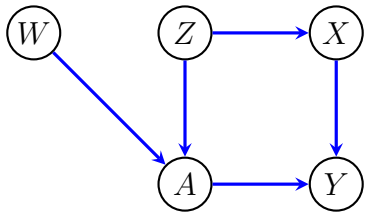


Figure 1.2: An example of a DAG over vertices A, Z, W, X, Y.

Definition 1.1. (Factorization) Let $\boldsymbol{X} = (X_1, \ldots, X_K)$ be a random vector. We say that the distribution $p(\boldsymbol{x})$ *factorizes* according to a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if:

$$p(\boldsymbol{x}) = \prod_{k \in \mathcal{V}} p\left(x_k \,|\, \boldsymbol{x}_{\mathrm{pa}(k)}\right).$$

This factorization implies certain conditional independencies: for each variable X_k, we have $X_k \perp\!\!\!\perp \boldsymbol{X}_{\mathrm{nd}(k)\setminus\mathrm{pa}(k)} \,|\, \boldsymbol{X}_{\mathrm{pa}(k)}$, where $\mathrm{nd}(k)$ denotes the set of non-descendants of node k.

For example, the joint distribution $p(w, z, a, x, y)$ factorizes according to the DAG in Figure 1.2 as:

$$p(w, z, a, x, y) = p(w)p(z)p(x \,|\, z)p(a \,|\, w, z)p(y \,|\, a, x),$$

and a conditional independence that can be read off this DAG is $Y \perp\!\!\!\perp W \,|\, A, X$.

1.3.2.2 Do-approach and causal DAGs

However, while a DAG can encode statistical dependencies and conditional independencies, it does not, on its own, imply any causal interpretation. The factorization $p(\boldsymbol{x}) = \prod_{k \in \mathcal{V}} p\left(x_k \,|\, \boldsymbol{x}_{\mathrm{pa}(k)}\right)$ implies that each variable is independent of its non-descendants given its parents, but it does not tell us what would happen under an intervention. To reason about causality, we must go beyond statistical associations and consider how the system responds to external interventions. This leads to the concept of causal DAGs (Spirtes et al., 2000), which formalize how the system behaves under manipulations.

To distinguish passive observation from active intervention, Pearl introduced the do-operator (Pearl, 2009). The expression $p(y \mid do(x))$ denotes the distribution of Y under an intervention that sets $X = x$, breaking the natural mechanisms that would otherwise determine X.

Definition 1.2. (Interventional distribution) Let $\boldsymbol{X} = (X_1, \ldots, X_K)$ be a random vector indexed by $\mathcal{V}$ of a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Suppose we intervene to set the variables $\boldsymbol{X}_A$ to fixed values $\boldsymbol{x}_A$. The interventional distribution $p(\boldsymbol{x} \mid do(\boldsymbol{x}_A))$ is defined as:

$$p(\boldsymbol{x} \mid do(\boldsymbol{x}_A)) = \prod_{k \in \mathcal{V} \setminus A} p\left(x_k \mid \boldsymbol{x}_{\mathrm{pa}(k)}\right) \prod_{j \in A} \mathbf{1}\{X_j = x_j\}.$$

This equation is also known as the *g-formula* (Robins, 1986).

This definition captures two key modifications: (i) graphically, the removal of incoming edges into $\boldsymbol{X}_A$, and (ii) the replacement of their conditional distributions with point masses at the fixed intervention values. Recall our flu vaccination example, $do(vaccine)$ means that in Figure 1.1, the edge from health behaviours $\rightarrow$ vaccine is removed and this subgraph corresponds to an RCT where vaccination is randomly assigned and independently of any predecessors, i.e. health behaviour, while all other relations remain the same.

Definition 1.3. (Causal DAG) Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a DAG and let p be a distribution over the $\boldsymbol{X} = (X_1, \ldots, X_K)$. Then $\mathcal{G}$ is said to be a causal DAG for p if:

(i) p factorizes according to $\mathcal{G}$, and

(ii) for any $A \subset \mathcal{V}$,

$$p(\boldsymbol{x} \mid do(\boldsymbol{x}_A)) = \prod_{k \in \mathcal{V} \setminus A} p\left(x_k \mid \boldsymbol{x}_{\mathrm{pa}(k)}\right) \prod_{j \in A} \mathbf{1}\{X_j = x_j\}.$$

Intuitively, a causal DAG can be thought of as representing a causal system in which each directed edge encodes a causal relationship. For example, in Figure 1.2, the edge $A \rightarrow Y$ implies that intervening on A can change the distribution of Y; Conversely, intervening on Y does not affect A—a manifestation of the asymmetry inherent in causal relationships.

1.3.2.3 Graphical rules for identification

A key question in causal inference is whether a causal quantity, such as $p(\boldsymbol{x}_D \mid do(\boldsymbol{x}_A))$ is identifiable from observed data. In the graphical framework, this question is addressed by examining the structure of a causal DAG to determine whether interventional distributions can be expressed purely in terms of observational quantities. The following definition formalizes the notion of identifiability.

Definition 1.1. (Identifiability of Causal Effects) Consider $\boldsymbol{X} = \boldsymbol{X}_V$ and let (O, U) be a partition of $\mathcal{V}$. Let $\mathcal{P}$ be a class of distributions for $\boldsymbol{X}$, including the relevant observational and interventional distributions, for instance, but not necessarily, as induced by a causal DAG. Finally, let $A \subset O$ and $D \subset O \setminus A$.

We say that the causal effect of $\boldsymbol{X}_A$ on $\boldsymbol{X}_D$ is *identified by* X_O under $\mathcal{P}$ if for any two distributions, $p', p'' \in \mathcal{P}$:

$$p'(\boldsymbol{x}_O) = p''(\boldsymbol{x}_O) \qquad \Rightarrow p'(\boldsymbol{x}_D \,|\, do(\boldsymbol{x}_A)) = p''(\boldsymbol{x}_D \,|\, do(\boldsymbol{x}_A)).$$

That is, the interventional distribution is uniquely determined by the observational distribution.

In the context of DAGs, graphical criteria such as the back-door and front-door criteria (Pearl, 1993, 1995, 2009) provide simple, sufficient conditions for identification of interventional distributions. These conditions rely on the notion of d-separation, a graphical criterion for reading conditional independencies from a DAG.

Definition 1.2. (d-separation) A path p is said to be d-separated (or blocked) by a set of nodes C, if and only if,

(i) p contains a chain $i \to w \to j$ or a fork $i \leftarrow w \to j$ such that the middle node w is in C, or

(ii) p contains an inverted fork (or collider) $i \to w \leftarrow j$ such that the middle node w is not in C and such that no descendant of w is in C.

A set C is said to d-separate X from Y if and only if C blocks every path from a node in X to a node in Y.

We now state two identification results for DAGs.

Theorem 1.1. *(Back-Door Criterion (Pearl, 1993, 2009)) Let $\mathcal{G}$ be a causal DAG for $\boldsymbol{X}$. The causal effect of X_j on X_k is identified by $\boldsymbol{X}_{\{j,k\} \cup C}$, $C \subset V \setminus \{j, k\}$ if*

(i) $C \cap \mathrm{de}(j) = \emptyset$,

(ii) C blocks every path between j and k that contains an arrow into j in $\mathcal{G}$.

The intervention distribution is then identified as

$$p(x_k \,|\, do(x_j)) = \int p(x_k \,|\, x_j, \boldsymbol{x}_C) p(\boldsymbol{x}_C) d\boldsymbol{x}_C$$

Theorem 1.2. *(Front-Door Criterion (Pearl, 1995, 2009)) Let $\mathcal{G}$ be a causal DAG for $\boldsymbol{X}$. The causal effect of X_j on X_k is identified by a variable set $C \subset V \setminus \{j, k\}$ if:*

(i) C intercepts all directed paths from X_j to X_k,

(ii) all back-door paths from X_j to C are blocked, and

(iii) all back-door paths from C to X_k are blocked by X_j.

Then the intervention distribution is identified as

$$p(x_k \,|\, do(x_j)) = \int p(\boldsymbol{x}_C \mid x_j) \int p(x_k \,|\, \boldsymbol{x}_C, x'_j) p(x'_j)\, dx'_j\, d\boldsymbol{x}_C.$$

These graphical rules are widely used in practice for their interpretability and tractability. This is precisely an advantage of DAGs for causal modelling, the ability to infer identifiability given a complex model. However, an important limitation is that they rely on the availability of a correctly specified DAG. In many real-world problems, the causal structure may be only partially known, and relevant unmeasured variables may not be explicitly included. This motivates the use of alternative frameworks for defining and reasoning about causal effects.

In the next section, we turn to the potential outcomes framework, which formalizes causal effects as comparisons between counterfactual outcomes and provides a complementary perspective to graphical models.

1.3.3 Potential outcomes

While DAGs and do-calculus emphasize structural relationships and graphical criteria for identification, the potential outcomes framework (Rubin, 1974, 1978) centres on defining causal effects as contrasts between hypothetical treatment assignments—often referred to as counterfactuals. These two frameworks differ in language and emphasis but are mathematically compatible and often used in tandem.

This section introduces the concepts of potential outcomes, estimands, identification, and estimation. Each is presented using potential outcomes notation, while noting connections to the graphical approach.

1.3.3.1 Potential Outcomes

Let $Y(a)$ denote the potential outcome that would be observed if treatment A were set to value $a \in \mathcal{A}$. For dichotomous treatments, the individual-level causal effect is

defined as $Y(1)-Y(0)$, though only one of these potential outcomes can be observed for each unit.

In order for the potential outcomes $Y(a)$ to be well-defined, the framework relies on the *Stable Unit Treatment Value Assumption (SUTVA)*:

Assumption 1.1. (SUTVA)

(i) No interference: The treatment assignment of one unit does not affect the outcomes of another.

(ii) No multiple versions of treatment: Each treatment level is well-defined and consistently implemented.

This framework provides a formal language to define causal effects even when a DAG is not fully specified. However, the connection between the two frameworks is direct: the potential outcome $Y(a)$ corresponds to the interventional distribution $p(Y \mid do(A = a))$ in a causal DAG. Note that potential outcomes are formulated at the level of variables instead of distributions.

In this thesis, we are pragmatic about the use of notations: we tend to use the do-notations for interventional distributions and use the potential outcomes notation to talk about estimands, assumptions and estimators.

1.3.3.2 Causal estimands

An estimand is a mathematical quantity you aim to learn about. It is defined independently of the data observed or the statistical method used for estimation. Because individual-level causal effects $\tau_i = Y(1) - Y(0)$ are unobservable, causal inference typically focuses on *population-level* estimands that average or condition these effects across units.

Assuming binary treatment $A \in \{0, 1\}$, common estimands include:

Definition 1.3. (Average Treatment Effect (ATE))

$$\tau = \mathbb{E}[Y(1) - Y(0)]$$

In the do-notation, it corresponds to $\mathbb{E}[Y \mid do(A = 1)] - \mathbb{E}[Y \mid do(A = 0)]$

Definition 1.4. (Average Treatment Effect on the Treated (ATT) and on the Controls (ATC))

$$\tau_T = \mathbb{E}[Y(1) - Y(0) \mid A = 1]$$
$$\tau_C = \mathbb{E}[Y(1) - Y(0) \mid A = 0]$$

Definition 1.5. (Conditional Average Treatment Effect on the Treated (CATE))

$$\tau(C) = \mathbb{E}[Y(1) - Y(0) \,|\, C]$$

If the treatment effect varies by values of C, then we say that the treatment is modified by C, or there is treatment effect heterogeneity (HTE) across values of C. As a function of C, some papers refer to $\tau(c)$ as the HTE function.

The above definitions calculate the treatment effect as the average difference in potential outcomes. In fact, the concept of treatment is more flexible than this description suggests. For binary or time-to-event outcomes, other estimands like risk ratios, odds ratios, or hazard ratios may be more appropriate. The estimand should align with the scientific question, decision context, and properties of the data.

1.3.3.3 Identification

A causal estimand is identifiable if it can be uniquely expressed as a function of the observed data distribution under a set of assumptions.

Definition 1.6. (Identification) A parameter θ is identifiable if it can be written as a function of the distribution of the observed data under certain model assumptions. A parameter θ is nonparametrically identifiable if this expression does not rely on parametric models.

In both the potential outcomes and graphical frameworks, identifying assumptions play a crucial role. For the ATE, the following standard assumptions are typically made:

Assumption 1.2. (Consistency) $A = a \Rightarrow Y = Y(a)$.

For the vast majority of methods, it is necessary to assume that the observed (or measured) covariates are sufficient to explain any non-causal correlation between the treatment and the outcome. This is sometimes called conditional *exchangeability*, *conditional ignorability*, or *causal sufficiency*.

Assumption 1.3. (Exchangeability) $Y(a) \perp\!\!\!\perp A \,|\, X$.

Assumption 1.4. (Positivity or overlap) $0 < \mathbb{P}(A = 1 \mid X = x) < 1 \quad \text{for all } x \in \mathcal{X}$

A well-conducted RCT usually satisfies these assumptions.

Under these assumptions, the ATE is identifiable from observed data via the following expression:

$$\tau = \mathbb{E}\left[\mathbb{E}[Y \mid A = 1, X] - \mathbb{E}[Y \mid A = 0, X]\right].$$

This representation translates the causal estimand into a statistical functional of the observed data distribution.

1.3.3.4 Estimation

Given identification assumptions, several methods can be used to estimate causal effects. Common approaches include:

Outcome regression: If we assume a linear outcome model of the form: $\mu_A(X) = \mathbb{E}[Y \mid A, X] = \beta_0 + \beta_a A + \beta_X^T \boldsymbol{X} + \beta_{AX}^T \boldsymbol{X} A$, then the CATE is:

$$\tau(X) = \mu_1(X) - \mu_0(X) = \beta_a + \beta_{AX}^T \boldsymbol{X},$$

which leads to the ATE $\tau = \mathbb{E}[\tau(X)] = \beta_a + \beta_X^T \mathbb{E}(\boldsymbol{X})$. A plug-in estimator for τ is given by $\hat{\beta}_a + \hat{\beta}_X^T \bar{\boldsymbol{X}}$, where $\bar{\boldsymbol{X}}$ denotes the sample mean of $\boldsymbol{X}$.

More generally, one can estimate $\mu_1(X)$ and $\mu_0(X)$ using more flexible models—such as regression trees, random forests (e.g. Wager and Athey (2018)), or other machine learning algorithms—rather than relying on parametric forms. This yields a general outcome regression estimator:

$$\hat{\tau}_{\text{reg}} = \frac{1}{n} \sum_{i=1}^{n} \left[\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i) \right].$$

A known limitation of this approach is its sensitivity to model specification. Parametric models (e.g., linear or logistic regression) risk bias if misspecified. Nonparametric or machine learning models offer greater flexibility but may suffer from high variance or overfitting, especially in small samples.

Inverse probability weighting (IPW): Inverse propensity weighting creates a *pseudo-population* in which the arrow from the covariates $\boldsymbol{X}$ to the treatments is removed. The pseudo-population has two properties: $\boldsymbol{X}$ and A are statistically independent and the weighted means follow $\mathbb{E}_{ps}[Y \mid A = a] = \sum_{\boldsymbol{x}} \mathbb{E}[Y \mid A = a, \boldsymbol{X} = \boldsymbol{x}] \mathbb{P}[\boldsymbol{X} = \boldsymbol{x}]$ in the actual population. These properties are true even if conditional exchangeability (Assumption 1.3) does not hold.

If exchangeability (Assumption 1.3) holds, we can estimate τ by first estimating the propensity score $e(X) = \mathbb{P}(A = 1 \mid X)$ and the computing

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{A_i Y_i}{\widehat{e}(X_i)} - \frac{(1 - A_i) Y_i}{1 - \widehat{e}(X_i)} \right].$$

When $\boldsymbol{X}$ only consists of discrete variables and is of low dimensions, $e(X)$ can be non-parametrically estimated in each stratum of $\boldsymbol{X}$. Otherwise, we need a parametric model to estimate $e(X)$, a common choice is a logistic regression model. If the model is correctly specified, then $\mathbb{E}[\hat{\tau}] = \tau$, otherwise, the estimates are likely biased.

Doubly robust estimation: The outcome regression and IPW estimators are consistent only if the outcome model or the propensity score model are correctly specified, respectively. However, with observational data one can never be sure that either is correct. In a causal inference model, an estimator is doubly robust if it remains consistent when either, but not necessarily both, is correctly specified. So doubly robust estimators give the analyst two chances, instead of only one, to make a valid inference.

One way the IPW estimator can be improved is by fully utilizing the information in the conditioning set. The conditioning set X contains information about the probability of treatment, but it also contains predictive information about the outcome variable. The AIPW estimator efficiently uses this information.

We posit a model for the conditional means of the outcome $\mu_1(X, \beta_1)$ and $\mu_0(X, \beta_0)$, indexed by the parameters β_1 and β_0, and a working model for the propensity score $e(X, \alpha)$, indexed by the parameter α.

Define

$$\tilde{\mu}_1^{\text{dr}} = \mathbb{E}\left[\frac{A\{Y - \mu_1(X, \beta_1)\}}{e(X, \alpha)} + \mu_1(X, \beta_1)\right], \tag{1.1}$$

$$\tilde{\mu}_0^{\text{dr}} = \mathbb{E}\left[\frac{(1 - A)\{Y - \mu_0(X, \beta_0)\}}{1 - e(X, \alpha)} + \mu_0(X, \beta_0)\right], \tag{1.2}$$

Then if either the propensity score model, or the outcome regression model is correctly specified, the doubly robust estimator $\tilde{\mu}_1^{\text{dr}} - \tilde{\mu}_0^{\text{dr}}$ is consistent for τ.

Theorem 1.3. *Assume Assumption 1.3 and Assumption 1.4, if either* $e(X, \alpha) = e(X)$ *or* $\{\mu_1(X, \beta_1) = \mu_1(X), \mu_0(X, \beta_0) = \mu_0(X)\}$*, then* $\tilde{\mu}_1^{\text{dr}} - \tilde{\mu}_0^{\text{dr}} = \tau$.

Each method has strengths and limitations in finite samples, particularly in the presence of model misspecification, limited overlap, or high-dimensional covariates. Practical implementations often involve regularization, cross-fitting, or nonparametric methods to improve robustness and efficiency.

1.3.4 Marginal structural models

1.3.4.1 Time-varying confounding

In the absence of unmeasured confounding and model misspecification, standard methods for estimating the causal effect of time-varying treatments on survival are biased when

(i) there exists a time-dependent covariate that predicts outcome as well as subsequent treatment; and

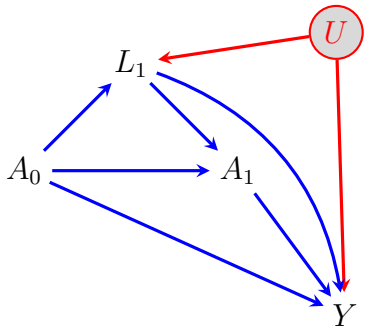


Figure 1.3: An example DAG showing temporal relationships in a longitudinal discrete setting.

(ii) past treatment history modifies the future levels of this covariate.

Covariates that satisfy condition (i) are referred to as *time-varying confounders*. The issue of exposure-affected *time-varying confounding* arises when the time-varying confounders are themselves influenced by prior treatment exposures, known as *exposure-confounder feedback*. We illustrate this issue using a simplified example to explain why traditional methods are inadequate.

For clarity, consider the causal effect at a single time point, as represented by the DAG in Figure 1.3. In this causal structure, the covariate L_1 predicts both the outcome Y and treatment A_1, making it a time-varying confounder. Additionally, prior treatment A_0 influences L_1. There also exists a latent variable U that affects both L_1 and Y,though it does not directly influence the treatment and is not a treatment-outcome confounder. In the HIV study example mentioned in the motivation Section 1.2.3, for instance, L_1 could represent the CD4 cell count, which is influenced by whether treatment has been previously initiated, predicts future treatment, and is associated with future survival. The latent variable U could represent an unobserved factor, such as the patient's overall health status.

One might consider adjustment through regression or stratification, akin to the time-fixed case, estimating the conditional distribution $P(Y \mid L_1, A_0, A_1)$. However, in this time-dependent context, adjusting for L_1 in this way introduces bias. First, the temporal structure allows for causal pathways from treatment to outcome mediated by covariates, such as the pathway $A_0 \to L_1 \to Y$. Adjusting for L_1 could block this pathway, leading to biased estimates of the total effects of earlier treatments, as the treatment's impact through CD4 cell count is not accounted for. Second, conditioning on L_1, a collider on the path $A_0 \to Y$, induces an association between its parents A_0 and U and hence between A_0 and Y. This phenomenon, known as selection bias, further distorts the analysis. Failing to adjust for L_1, however, also results in bias due to confounding. As a result, conventional adjustment methods,

such as an ordinary time-dependent Cox model, produce biased estimators of the total causal effect, even when unmeasured confounders are absent, and the model is correctly specified.

1.3.4.2 Identification and estimation

We first introduce identifying assumptions, which are extensions of the exchangeability and positivity assumption (Assumption 1.3 and 1.4) to longitudinal settings with time-varying treatments and confounders.

Definition 1.4. (Sequential exchangeability) Sequential exchangeability holds if, for all time points k and all treatment histories $\bar{a}_k$,

$$Y(\bar{a}_k) \perp\!\!\!\perp A_k \mid \bar{L}_k, \bar{A}_{k-1}.$$

This assumption states that, conditional on the treatment and covariate history up to k, the treatment assignment at time k is independent of the potential outcome $Y(\bar{a}_k)$. Sequential exchangeability requires that all time-varying and baseline confounders of the treatment–outcome relationship be measured. However, it does not require that there be no unmeasured common causes of covariates and the outcome. For example, in the causal DAG in Figure 1.3, sequential exchangeability holds.

Definition 1.7. (Sequential positivity) For all time points k, the sequential positivity assumption holds if, for all treatment and covariate histories $(\bar{a}_k, \bar{l}_k)$ such that $p_{\bar{A}_{k-1},\bar{L}_k}(\bar{a}_{k-1}, \bar{l}_k) > 0$, we have

$$p_{A_k|\bar{A}_{k-1},\bar{L}_k}(a_k \mid \bar{a}_{k-1}, \bar{l}_k) > 0.$$

Under sequential exchangeability and positivity, several methods are available to adjust for exposure-affected time-varying confounding. *Structural Nested Models* (SNMs), introduced by Robins (1989) and Robins and Tsiatis (1991), estimate the causal effect of exposure by contrasting outcomes under hypothetical exposure scenarios. It is named as such because in the econometric and social science literature, causal models, i.e. models for counterfactual variables, are often referred to as *structural*.

Definition 1.5. (Structural equation model with a single time point)

$$\mathbb{E}\big[Y(a) - Y(0) \mid A = a, L\big] = g(a, L; \beta).$$

They are called *nested* because in the case of time-varying treatments, they model causal effects in a recursive, time-indexed structure, where the effect of a treatment at a particular time is defined conditional on past treatments and covariate history, See Definition 1.6.

Definition 1.6. (Structural equation model with multiple time points)

$$\mathbb{E}\big[Y(\bar{a}_k, 0) - Y(\bar{a}_{k-1}, 0) \,|\, \bar{a}_k, \bar{L}_k\big] = g_k(\bar{a}_k, \bar{L}_k; \beta_k).$$

for $k = 1 \ldots, K$. Specifically here, we tweak the notation a bit and $(\bar{a}_k, 0)$ denotes the treatment history given by $\bar{a}_k$ through time k and zero at times $k+1, \ldots, K$.

The *nested* nature of SNMs comes from specifying one equation for each time point. The first equation models the effect of receiving treatment at time 0 and not thereafter, while the second equation models treatment at time 1, and so on for additional time points. *G-estimation* is a key approach used with SNMs, identifying parameters β_k that render the exposure variable independent of the counterfactual outcome, assuming conditional exchangeability. For time-to-event outcomes, SNMs include Structural Nested Accelerated Failure Time Models (SNAFTM) (Robins, 1992; Hernán et al., 2005), and Structural Nested Cumulative Failure Time Models (SNCFTM) (Picciotto et al., 2012).

As an alternative to SNMs, Marginal Structural Models (MSMs) have been developed to estimate causal effects of time-varying treatments (Robins, 1997; Hernán et al., 2000; Hernán et al., 2001). MSMs generalize structural modelling to functionals of the distribution of counterfactual outcomes under treatment regimes, while marginalizing over post-baseline confounders. These models define marginal (often population-level) causal parameters that may include means, survival probabilities, or quantiles.

Definition 1.8. (Marginal Structural Model)

$$T\left(F\left(Y(\bar{a}_k) \,|\, X\right)\right) = g(\bar{a}_k, X; \beta),$$

where X is a subset of pre-treatment covariates and T is a functional of the counterfactual distribution, e.g. mean, median, quantile, survival probability, cumulative incidence.

Here are examples:

Example 1.3.1. (Marginal Structural Mean Model)

$$\mathbb{E}[Y(\bar{a})] = g(\bar{a}; \gamma).$$

For survival outcomes, the marginal structure of the outcome can be modeled via hazard, such as using Cox regression (Cox, 1972) or an additive hazard model (Aalen, 1989). A Cox MSM specifies proportional hazard function:

Example 1.3.2. (Cox Marginal Structural Model)

$$\lambda_{\bar{a}_t}(t) = \lambda_0(t) \exp\left\{g(\bar{a}_t; \boldsymbol{\gamma})\right\},$$

where $\lambda_0(t)$ is the baseline (e.g. never treated) counterfactual hazard at time t, $\bar{a}_t$ denotes the treatment history up to time t. $g(\bar{a}_t; \psi)$ is a known function describing how treatment influences the hazard.

If the hazard function is modified to $\lambda_{\bar{a}_t}(t) = \lambda_0(t) + \{g(\bar{a}_t; \boldsymbol{\theta})\}$, it results in Aalen's additive hazard model (Aalen, 1989).

In Chapter 4 we propose a simulation method to simulate longitudinal data from these two examples of widely used MSMs.

MSMs are *marginal* because they model causal effects as marginal quantities by averaging over time-varying covariates, and they are *structural* because they explicitly model causal relationships, unlike associational models that capture observational patterns. MSMs can be estimated using methods such as g-formula (Robins, 1986), inverse probability of treatment weighting (IPTW), and doubly robust methods such as targeted maximum likelihood estimation (TMLE) (van der Laan and Rubin, 2006). A systematic review of the literature (Clare et al., 2019) found that IPTW-estimated MSMs are the most commonly used approach for longitudinal data with time-varying confounding due to their simplicity in both implementation and interpretation.

Definition 1.7. (Time-varying inverse propensity weights)

$$W_k = \prod_{j=0}^{k} \frac{1}{p(A_j \mid \bar{A}_{j-1}, \bar{L}_k)} \qquad SW_k = \prod_{j=0}^{k} \frac{p(A_j \mid \bar{A}_{j-1})}{p(A_j \mid \bar{A}_{j-1}, \bar{L}_k)},$$

where $p(A_j \mid \bar{A}_{j-1}, \bar{L}_k)$ is by definition the conditional probability mass function. The first is unstabilized weight and the second is stabilized weight.

The stabilizing term, $p(A_j \mid \bar{A}_{j-1})$, helps decrease the magnitude of the weights and reduces the numerical instability related to small probability. Robins (2000) proved that given sequential exchangeability and positivity, MSMs can be consistently estimated using IPTW.

1.3.4.3 Why simulating from MSMs is challenging

For both method evaluation and sample size calculation, it is essential to be able to simulate data that precisely follows a known causal structure. However, simulating longitudinal data from a pre-specified MSM is not a straightforward task. While it

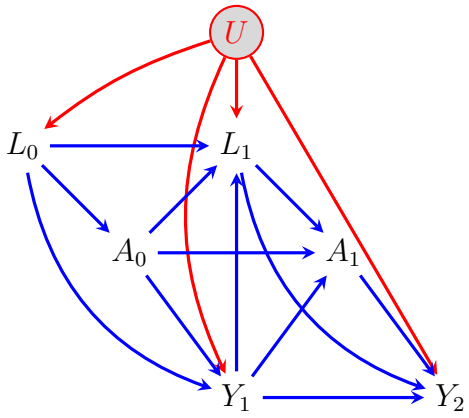


Figure 1.4: An example DAG showing temporal relationships in a longitudinal discrete setting.

seems *natural* to generate data sequentially—creating each individual's covariates, treatment status, and survival status at each time point based on past information—MSMs rely on marginal, rather than conditional models. This makes it difficult to sequentially simulate from conditional models that align with a pre-specified marginal model.

To illustrate this, we expand the causal model in the earlier example a bit and consider the causal model in the DAG shown earlier in Figure 1.4. The DAG imposes the following factorization on the joint distribution:

$$P(U, L_0, A_0, Y_1, L_1, A_1, Y_2) = P(U)P(L_0 \,|\, U)P(A_0 \,|\, L_0)P(Y_1 \,|\, U, L_0, A_0) \\ \cdot P(L_1 \,|\, U, L_0, A_0, Y_1)P(A_1 \,|\, L_1, A_0, Y_1)P(Y_2 \,|\, U, L_1, A_1, Y_1).$$

This factorization suggests a straightforward simulation process: first generate the latent variable U, then sequentially generate L_0 given U, A_0 given $\{L_0, U\}$, Y_1 given $\{U, L_0, A_0)$, followed by the same process for subsequent time points.

However, if we want this data-generating process to also follow a given MSM that parameterizes $P(Y_1 \mid do(a_0))$ and $P(Y_2 \mid do(a_0, a_1))$, we must ensure compatibility of these conditional distributions with the marginal structure. Applying g-formula yields:

$$P(Y_1 \mid do(a_0)) = \sum_{l,u} P(Y_1 \,|\, U = u, L_0 = l, A_0 = a_0)P(L_0 = l \,|\, U)P(U = u) \tag{1.3}$$

$$P(Y_2 \mid do(a_0, a_1)) = \sum_{l_0, l_1, u} P(Y_2 \,|\, U = u, L_1 = l_1, A_1 = a_1, Y_1 = 1) \\ P(L_1 = l_1 \,|\, U = u, L_0 = l_0, A_0 = a_0, Y_1 = 1)P(L_0 = l_0 \,|\, U)P(U = u) \tag{1.4}$$

For continuous variables, these sums are replaced by integrals. Simulating data from given MSMs for $P(Y_1 \mid do(a_0))$ and $P(Y_2 \mid do(a_0, a_1))$ requires compatible choices for all components in Equation (1.3) and (1.4). While this is feasible for linear models in low dimensions for a limited number of time steps, it becomes highly challenging when moving beyond linear relationships – especially with non-collapsible models like logistic regression or Cox proportional hazards models. An additional challenge arises from the *g-null paradox*(Robins and Wasserman, 1997): The g-null paradox states that, under the causal null hypothesis, no smooth parametric model for the observed data (e.g. outcome regression or propensity scores) can, in general, correctly represent the data-generating distribution when time-varying confounders are affected by prior treatment. Conditioning on such confounders induces spurious associations, making it impossible to test the null without specialized g-methods. The core difficulty lies in finding compatible conditional distributions that preserve the specified MSM—an intractable problem in non-linear settings."

1.3.5 Frugal parameterization

A key difficulty in simulating longitudinal data from MSMs lies in constructing a joint distribution that is consistent with a specified causal margin. To address this challenge, Evans and Didelez (2024) introduced the *frugal parameterization*, a flexible framework that explicitly parameterizes the causal margin and builds the remainder of the joint distribution around it.

The core idea is to decompose the joint distribution $p(\boldsymbol{l}, a, y)$ into three components:

(i) the past: $p_{\boldsymbol{L}A}(\boldsymbol{l}, a)$—the joint distribution of covariates and treatment antecedent to observing outcomes;

(ii) the causal margin: $p^*_{Y|A}(y \mid do(a))$—a marginal structural model specifying the distribution of outcomes under intervention, specified to address the causal question;

(iii) a dependency measure: $\phi^*_{Y\boldsymbol{L}|A}$—captures the dependency between the outcome Y and the covariates $\boldsymbol{L}$ under intervention.

To distinguish interventional distributions from standard conditional distributions, the notation p^* is used to indicate that treatment has been set externally at $A = a$ via intervention. This construction ensures that the marginal causal quantity is a core component of the model, rather than a derived consequence. A common

choice for the dependence measure is a copula, which allows the joint distribution to reflect desired dependence without altering the margins.

To formalize this decomposition, Evans and Didelez (2024) introduce the concept of a cognate distribution, which provides a way to embed a causal margin within a coherent joint distribution:

Definition 1.9. (Cognate distribution) A conditional distribution $p^*_{Y|A}(y \mid a)$ is said to be *cognate* to $p_{Y|A}(y \mid a)$ if it can be expressed in the form

$$p^*_{Y|A}(y \mid a) = \int_{\boldsymbol{\mathcal{L}}} p_{Y|\boldsymbol{L}A}(y \mid \boldsymbol{l}, a) \cdot w(\boldsymbol{l} \mid a)\, \mathrm{d}\boldsymbol{l}, \quad a \in \mathcal{A},\ y \in \mathcal{Y}, \tag{1.5}$$

for some kernel, i.e. conditional density, $w(\boldsymbol{l} \mid a)$.

This class includes the standard conditional as a special case when $w = p_{\boldsymbol{L}|A}$.

Definition 1.10. (Frugal parameterization) A smooth, regular parameterization of the joint distribution of (L, A, Y) is said to be *frugal* with respect to a cognate kernel $p^*_{Y|A}$ if it consists of separate parameterizations of $p_{\boldsymbol{L}A}$,$p^*_{Y|A}$ and $\phi^*_{Y\boldsymbol{L}|A}$.

The main result of Evans and Didelez (2024) ensures that such a decomposition can be made coherently:

Theorem 1.4. *Let $p_{\boldsymbol{L}AY}$ be a distribution parameterized by $\theta := (\theta_{\boldsymbol{L}A}, \theta_{Y|A}, \phi_{Y\boldsymbol{L}|A})$, with weight function $p_{\boldsymbol{L}|A}$ and a kernel w satisfying the following assumptions:*

(A1) The parameter $\phi_{Y\boldsymbol{L}|A}$ is jointly variation independent of $\theta_{\boldsymbol{L}A}$ and $\theta_{Y|A}$;

(A2) The product $p^*_{\boldsymbol{L}A} = p^*_A \cdot w$ has a smooth and regular parameterization $\eta_{\boldsymbol{L}A} := \eta_{\boldsymbol{L}X}(\theta_{\boldsymbol{L}A})$, where $\eta_{\boldsymbol{L}A}$ is a twice differentiable function with a Jacobian of constant rank;

(A3) $p_{\boldsymbol{L}A}$ is absolutely continuous with respect to $p^*_{\boldsymbol{L}A}$ at the true distribution $p_{\boldsymbol{L}AY}$.

Then θ is frugal with respect to $p_{Y|A}$ if and only if $\theta^* := (\theta_{\boldsymbol{L}A}, \theta^*_{Y|A}, \phi^*_{Y\boldsymbol{L}|A})$ is also frugal with respect to $p^*_{Y|A}$.

This result guarantees that the three components, $p_{\boldsymbol{L}A}$, $p^*_{Y|A}$, and $\phi^*_{Y\boldsymbol{L}|A}$, can be specified independently without violating the validity of the full joint distribution, offering both theoretical elegance and practical flexibility. In particular, it allows likelihood-based inference and simulation to be centred directly on the marginal causal structure of interest, without requiring a full structural specification.

The frugal parameterization thus provides a powerful tool for modeling causal quantities. It allows both likelihood-based inference and simulation to be centered directly on the causal margin—without the need to specify a full structural model. This approach forms the foundation of the methodology developed in Chapter 2 and Chapter 4 of this thesis.

1.4 Summary and roadmap

This chapter has introduced the conceptual foundations and practical motivations underlying this thesis. Two methodological challenges have emerged as central themes:

(i) how to combine randomized trials and real-world data to improve the efficiency of causal estimation; and

(ii) how to simulate longitudinal data in a way that respects the causal structure specified by MSMs, enabling robust method evaluation and principled trial design.

The remainder of the thesis is structured as follows:

- Chapter 2 develops a quasi-empirical Bayesian methodology for data fusion that augments trial data with real-world data through a power likelihood approach. The method enables adaptive borrowing of information from observational sources while balancing bias due to population differences or confounding.

- Chapter 3 undertakes a systematic evaluation of existing data fusion methods aimed at improving efficiency. A comprehensive simulation study is used to assess their empirical performance—comparing bias, variance, and coverage across scenarios with varying sample sizes and confounding structures. These methods are then applied to a real-world case study that augments the PIONEER 6 cardiovascular safety trial with administrative claims data from Optum. Demonstrating how simulation can be used to evaluate data fusion methods is itself valuable for practitioners.

- Chapter 4 introduces a simulation framework for generating longitudinal data from MSMs using frugal parameterization. The method allows for exact simulation from pre-specified marginal models, enabling benchmarking of causal inference methods and sample size calculation under known ground truth.

Together, these chapters offer new methodological tools and empirical insights for improving the practice of causal inference in complex, real-world settings.

Chapter 2

Combining experimental and observational data through a power likelihood

Abstract. *Randomized controlled trials are the gold standard for causal inference and play a pivotal role in modern evidence-based medicine. However, the sample sizes they use are often too limited to provide adequate power for drawing causal conclusions. In contrast, observational data are becoming increasingly accessible in large volumes but can be subject to bias as a result of hidden confounding. Given these complementary features, we propose a power likelihood approach to augmenting RCTs with observational data to improve the efficiency of treatment effect estimation. We provide a data-adaptive procedure for maximizing the expected log predictive density (ELPD) to select the learning rate that best regulates the information from the observational data. We validate our method through a simulation study that shows increased power while maintaining an approximate nominal coverage rate. Finally, we apply our method in a real-world data fusion study augmenting the PIONEER 6 clinical trial with a US health claims dataset, demonstrating the effectiveness of our method and providing detailed guidance on how to address practical considerations in its application.*

Contents

2.1 Introduction 32
2.2 Related literature 33
2.3 Setup, notation and assumptions 35
2.4 A power-likelihood approach 38
2.4.1 Frugal parameterization 38
2.4.2 Choosing the optimal influence factor η 40
2.4.3 Normal approximations of the posterior distribution 43
2.4.4 Asymptotics 44
2.5 Simulation 45
2.5.1 Moderating the bias-variance trade-off 45
2.5.2 Inference 47
2.6 PIONEER 6 application 47
2.6.1 Target trial emulation 48
2.6.2 Design 1: Incorporating control arm only 49
2.6.3 Design 2: Incorporating both arms 49
2.7 Conclusion and future work 51

2.1 Introduction

Combining data from randomized controlled trials (RCTs) and observational sources offers a potential solution to the persistent challenges in trial-based causal inference, including limited statistical power, delayed evidence, and constrained generalizability. As discussed in Section 1.2.1, real-world data (RWD) are increasingly used to supplement RCTs, and regulatory agencies such as the FDA, EMA, and ICH have begun to formally support such integration under appropriate conditions (U.S. Food and Drug Administration, 2018; U.S. Food and Drug Administration, 2021; European Medicines Agency, 2021; ICH, 2024).

This chapter develops a Bayesian methodology for robustly augmenting RCTs with observational data for improved estimation of treatment effects, particularly when trial data alone are underpowered for reliable inference. We apply this method to the PIONEER 6 study, a randomized, double-blind, placebo-controlled cardiovascular safety trial for semaglutide in patients with type 2 diabetes (see Section 1.2.1.1 for background). Although PIONEER 6 demonstrated non-inferiority to placebo (Husain et al., 2019), it was insufficiently powered to detect superiority. We show that by augmenting the trial with a US health insurance claims dataset,

our method leads to efficiency gains and increased statistical power. This application demonstrates the practical implementation and relevance of our approach. In Chapter 3, we revisit this setting to compare a broader class of data fusion methods.

The remainder of this chapter is organized as follows: Section 2.2 reviews Bayesian dynamic borrowing methods in recent literature. In Section 2.3 we outline the problem setup and main assumptions, before moving on to our proposed power likelihood approach in Section 2.4. In Section 2.5 we conduct a simulation study to illustrate the effectiveness of our approach. Section 2.6 features the aforementioned real-world application. We conclude with a discussion in the final section.

2.2 Related literature

Combining randomized and observational data for causal inference has attracted considerable attention among researchers. Central to the data fusion problem is the existence of unmeasured confounding in the RWD, which will bias causal effect estimation.

The Bayesian framework offers a natural mechanism to discount external information that may be in conflict with the trial data. Ibrahim and Chen (2000) introduced a general *power prior* approach that raises the likelihood of historical data to a fractional power η, which controls the discounting of external evidence:

$$\pi_{\mathrm{PP}}\left(\theta \mid D_O, \eta\right) \propto L\left(\theta \mid D_O\right)^{\eta} \pi_0\left(\theta\right),$$

where θ denotes the parameters of interest, $L(\theta \mid \cdot)$ is the likelihood function, and π_0 is the initial prior, which is often chosen to be uninformative. In their proposed approach, η is a hyperparameter and is assumed to be known, for example, set by expert opinion; obtaining a suitable value in practice can be challenging, however. The authors thus extended this formulation with a hierarchical specification where the uncertainty about η is endogenized through a prior distribution $\pi(\eta)$:

$$\pi_{\mathrm{HPP}}\left(\theta, \eta \mid D_O\right) \propto L\left(\theta \mid D_O\right)^{\eta} \pi_0\left(\theta\right) \pi\left(\eta\right).$$

However, Duan et al. (2006) and Neuenschwander et al. (2009) cautioned that this formulation violates the likelihood principle by omitting a normalizing term, which is a function of η. They showed that this inappropriate omission leads to counter-intuitive pooling results where the influence of the external data is close to zero even when the current and historical data are compatible. They therefore proposed the *normalized power prior*:

$$\pi_{\mathrm{NPP}}\left(\theta, \eta \mid D_O\right) = \frac{L\left(\theta \mid D_O\right)^{\eta} \pi(\theta)\pi(\eta)}{\int_{\Theta} L\left(\theta \mid D_O\right)^{\eta} \pi(\theta)\mathrm{d}\theta}.$$

A limitation of the power prior and its variations is that the *commensurability* of the external and current data is not explicitly parameterized. Ideally, we expect the external data are given more weight if it appears more compatible with the current data. To address this, bias-allowance models (Welton et al., 2012) that incorporate mechanisms to adjust for potential biases have been developed. Hobbs et al. (2011, 2012) introduced a location commensurate power prior (LCPP), where parameters in the current and historical data are explicitly modeled to be normally distributed:

$$\begin{aligned}
&\pi_{\text{LCPP}}\left(\theta, \eta, \omega \mid D_O\right) \\
&\propto \int \frac{\left[L\left(\theta_o \mid D_O\right)\right]^\eta}{\int \left[L\left(\theta_o \mid D_O\right)\right]^\eta \pi(\theta) d\theta_o} \cdot N\left(\theta \mid \theta_o, \frac{1}{\omega}\right) d\theta_o \\
&\quad \times \text{Beta}\left(\eta \mid g(\omega), 1\right) \times \pi(\omega).
\end{aligned}$$

Here θ_o is the parameter in the historical data and ω represents the precision of how θ is distributed around θ_o and hence parameterizes the *commensurability*. $g(\cdot)$ is a positive and increasing function of ω, so that η tends to take a large value when there is strong evidence of commensurability. Extending the commensurate prior approach, Schmidli et al. (2014) proposed a meta-analytic-predictive prior approach to borrow information from multiple historical trials, using a mixture of standard priors. This approach is applicable when data from multiple sources are available, as it includes a parameter to describe between-data heterogeneity. However, its application and effectiveness are uncertain when only two data sources are available. In such cases, a hierarchical shrinkage approach has been developed, where one estimate—typically from an RCT—is regarded as the reference in the combination (Röver and Friede, 2020).

Despite the advances in Bayesian dynamic borrowing, the challenge of specifying hyperparameters or commensurability functions persists, and applications primarily focus on aggregate data or use overly simplistic data-generating models (e.g. Rietbergen et al., 2011). Our proposed power likelihood method enhances the existing Bayesian dynamic borrowing literature in two significant ways. First, we adopt a *quasi*-empirical Bayes procedure that selects the borrowing strength η adaptively by maximizing predictive performance—eliminating the need for analysts and experts to specify a hyperparameter or commensurability function. Second, we incorporate individual-level data through the frugal parameterization of Evans and Didelez (2024), which enables estimation under flexible and causally-relevant models.

Compared to existing methods that are either limited to incorporating only external controls (e.g. Schuler et al., 2021) or require the presence of both treatment and control arms in the real-world data (e.g. Rosenman et al., 2023a; Yang et al.,

2023), our proposed approach offers the added flexibility of integrating external data either from the control arm alone or from both treatment and control arms. This flexibility is valuable because the treatments under study, such as a new drug or a social policy, may not yet be approved or implemented in the real world. Being able to leverage the real-world data, even without treatment arm observations, is important. Additionally, if a comparable treatment is already prevalent in the real world, utilizing information from both arms will make the data fusion more effective.

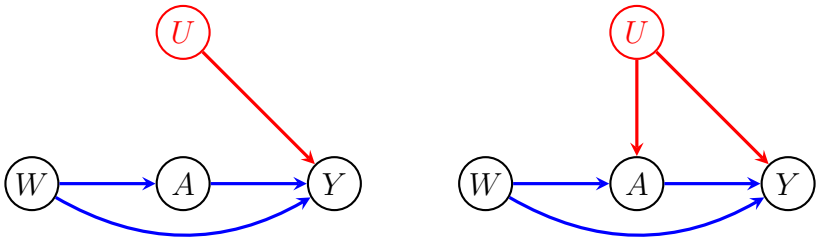
A more comprehensive survey and head-to-head comparison of data fusion methods, including non-Bayesian approaches, is provided in Chapter 3.

2.3 Setup, notation and assumptions

We consider two independent data sources, an RCT and an observational dataset. Let $S \in \{0, 1\}$ indicate whether an observation is from the RCT ($S = 1$) or the observational data ($S = 0$). The causal models for the two data sources are represented by the graphs in Figure 2.1, where $A \in \{0, 1\}$ represents a binary treatment, $\boldsymbol{W} \in \mathbb{R}^d$ is a vector of d pre-treatment covariates, $Y \in \mathbb{R}$ is the outcome of interest, and U represents unmeasured confounding. Let $\boldsymbol{X}_i = (Y_i, A_i, \boldsymbol{W}_i, S_i)$ denote the observed variables for the ith individual. The experimental data are therefore expressed as $\mathcal{D}_r = \{\boldsymbol{X}_i : \ S_i = 1\}$ with sample size n_r and the observational as $\mathcal{D}_o = \{\boldsymbol{X}_i : \ S_i = 0\}$ with size n_o. We assume the observations $\boldsymbol{X}_i$ are independently and identically distributed within each data source.

Potential outcomes (Rubin, 1974) are defined as in Section 1.3.3: we maintain the SUTVA (Assumption 1.1) and let $Y_i(a)$ denotes the outcome for unit i had it factually received treatment a, and the observed outcome $Y_i = Y_i(a)$ if $A_i = a$. The treatment effect for each individual (ITE), τ_i, is defined as a contrast of potential outcomes. For a continuous outcome, it is customary to define the ITE as the difference between potential outcomes $Y(1) - Y(0)$. For a binary outcome, candidate definitions of the treatment effect include risk difference (RD) $Y(1) - Y(0)$, relative risk $Y(1)/Y(0)$ and odds ratio (OR) $\frac{Y(1)/(1-Y(1))}{Y(0)/(1-Y(0))}$. Our proposed method is applicable to all of these definitions and—for illustration purposes—we adopt the difference in potential outcomes to define the causal effect. Averaging the ITE over a target population, we get the average treatment effect (ATE), $\tau \equiv \mathbb{E}[Y(1) - Y(0)]$.

In practice, there are scenarios where we are interested in how the treatment effect varies in subgroups. Based on the setup depicted in Figure 2.1, we can further divide the observed covariates $\boldsymbol{W}$ into two sets, $\boldsymbol{Z}$ and $\boldsymbol{C}$. The set $\boldsymbol{C} \subseteq \boldsymbol{W}$ denotes



(a) (b)

Figure 2.1: Causal models for (a) the experimental data and (b) the observational data.

the causal effect modifiers we are interested in, and $\boldsymbol{Z} = \boldsymbol{W} \setminus \boldsymbol{C}$ represents the remaining covariates that we want to marginalize over.

Our target estimand is the conditional average treatment effect (CATE) $\tau(\boldsymbol{c}) \equiv \mathbb{E}[Y(1) - Y(0) \mid \boldsymbol{C} = \boldsymbol{c}]$, which refers to the average treatment effect given $\boldsymbol{C} = \boldsymbol{c}$, but marginal over values of $\boldsymbol{Z}$ in the target population. It is worth highlighting that the choice of variables into set $\boldsymbol{C}$ depends on the specific causal question of interest; the ATE can be seen as a special case where the conditioning set $\boldsymbol{C}$ is empty.

Example 2.1. We use an example to further illustrate our setup, using distinct covariate sets $\boldsymbol{C}$ and $\boldsymbol{Z}$. In a medical context, clinicians may be interested in the efficacy of a drug on patients with certain characteristics. Suppose we have access to an RCT and an observational dataset, each of which record whether individuals took drug A, some health outcome Y, body mass index (BMI) C and creatinine level Z. We assume that both C and Z influence the outcome Y. In addition, there is a link between BMI C and creatinine level Z. There are likely unmeasured factors that serve as common confounders of A and Y, such as socio-economic status. To give guidance on the prescription of this drug, we are interested in the causal effect moderated by BMI C and want to marginalize it over the distribution of Z. Specifically, we are interested in $\tau(c) = \mathbb{E}\left[\mathbb{E}\left[Y(1) - Y(0) \,|\, C = c, Z\right]\right]$. We will revisit this example in Section 2.5, where we conduct a simulation study following this setup.

We now list the assumptions required for identification and analysis in our setting. For background on these assumptions, see Section 1.3.3.

Assumption 2.1. (RCT ignorability) (i) **Unconfoundedness** $Y(a) \perp\!\!\!\perp A \mid W, S = 1$ for $a \in \{0, 1\}$ (ii) **Positivity** $0 < \mathbb{P}(a = 1 \,|\, W, S = 1) < 1$ for all $(W, S = 1)$.

Variable	Definition	Meaning in Example 1
S	Variable indicating data source	RCT or observational data
A	Treatment	Drug under investigation
Y	Outcome	Health outcome
$\boldsymbol{C}$	PTCs: effect modifiers of interest	BMI
$\boldsymbol{Z}$	PTCs: others to marginalize over	Blood glucose level
$\boldsymbol{W}$	All PTCs: union of $\boldsymbol{C}$ and $\boldsymbol{Z}$	BMI & blood glucose level
$\boldsymbol{U}$	Unobserved confounders	Socio-economic status

Table 2.1: Description of the notations used in this chapter (PTC = pre-treatment covariate).

These conditions are typically satisfied in a well-conducted RCT and ensure that treatment assignment is conditionally independent of potential outcomes given covariates $\boldsymbol{W}$.

Assumption 2.2. (RCT representativeness) $\mathbb{E}[Y(1) - Y(0) \,|\, \boldsymbol{C}, S = 1] = \tau(\boldsymbol{c})$.

Assumption 2.2 maintains that the RCT sample is representative, allowing the CATE of interest to be transportable from the RCT to the target population. This assumption allows us to focus on gaining efficiency as the primary goal of data fusion, ignoring concerns of transporting or generalizing the inference to a different population.

Assumption 2.3. (RWD transportability)

$$\mathbb{E}[Y(1) - Y(0) \,|\, \boldsymbol{C}, S = 0] = \mathbb{E}[Y(1) - Y(0) \,|\, \boldsymbol{C}, S = 1].$$

Assumption 2.3 states that the CATE of interest in the RWD is identical to the CATE in the RCT population. It is important to highlight that Assumption 2.3 does not stipulate the mean of the potential outcomes to be the same across $\mathcal{D}_R$ and $\mathcal{D}_O$. Therefore, Assumption 2.3 is strictly weaker than assuming $\mathbb{E}[Y(a) \,|\, S = 0] = \mathbb{E}[Y(a) \,|\, S = 1]$ for $a \in \{0, 1\}$; this gives flexibility when the absolute level of the outcome differs between the RCT and RWD, for reasons such as different time windows, regions or standards of care. In practice, to improve adherence to this assumption, we can follow the target trial emulation framework, as outlined in Hernán and Robins (2016). This approach systematically emulates the design of a target RCT using observational data, thereby enhancing the comparability between the RCT and RWD, as demonstrated in the PIONEER 6 data study in Section 2.6.

2.4 A power-likelihood approach

Traditional Bayesian inference assumes that the data model, $p(x;\phi)$, is correct up to the unknown parameter value ϕ. However, Bayesian inference loses its predictive optimality when the data models are misspecified. A popular solution to robustly allow for Bayesian learning under model misspecification is to raise the likelihood to a fractional power (e.g. Grünwald and Van Ommen, 2017). That is, $\pi_\eta(\phi \mid X) \propto \pi(\phi)\prod_{i=1}^{n} p(X_i;\phi)^\eta$, indexed by a power η. This approach is described as a *power prior* in Ibrahim and Chen (2000), a *data-modified prior* in Walker and Hjort (2001), and a *power likelihood* in Holmes and Walker (2017). Adapting to our inference problem, we take the joint likelihood but only raise the likelihood of the observational data to a power η: $p_\eta(\boldsymbol{X_r}, \boldsymbol{X_o};\phi) = p_r(\boldsymbol{X_r};\phi) \times p_o\left(\boldsymbol{X_o};\phi\right)^\eta$, where $\boldsymbol{X_r}$ and $\boldsymbol{X_o}$ represent the observations from $\mathcal{D}_r$ and $\mathcal{D}_o$, and $p_r(\boldsymbol{X_r};\phi)$ and $p_o\left(\boldsymbol{X_o};\phi\right)$ are the joint densities respectively. Then by Bayes' rule, the posterior distribution of ϕ becomes:

$$\pi\left(\phi \mid \boldsymbol{X_r}, \boldsymbol{X_o}\right) \propto p_r(\boldsymbol{X_r};\phi) \cdot p_o\left(\boldsymbol{X_o};\phi\right)^\eta \cdot \pi(\phi), \tag{2.1}$$

where $\pi(\phi)$ is the prior distribution. We can see that the parameter η works like a dial moderating the influence from the observational data, relative to that from the experimental data. When $\eta = 0$, only the log-likelihood of the experimental data $\mathcal{D}_r$ is used and the influence of the observational data is cut off as though $\mathcal{D}_o$ were excluded from the analysis. When $\eta = 1$ then it is the conventional Bayesian posterior and means that we treat $\mathcal{D}_r$ and $\mathcal{D}_o$ equally for inference. Intuitively, we want to choose η to be between 0 and 1.

Two self-evidently important choices in our approach are of the likelihood itself and the power η. In the remainder of this section, we discuss the likelihood we propose to use and the robust selection of η. Note that while we advocate for the frugal parameterization below, due to the advantages we discuss, our method permits *arbitrary* models to be used and is therefore extremely flexible.

2.4.1 Frugal parameterization

To specify the joint densities p_r and p_o in (2.1), we propose to adopt the *frugal parameterization* (Evans and Didelez, 2024), which is designed specifically for causal inference applications. As set out in Section 1.3.5 of the preliminaries, we can break down a joint distribution $p(\boldsymbol{z}, a, y \mid \boldsymbol{c})$ into three variation independent pieces: (1) the quantity of inferential interest; (2) the distribution of the treatment and covariates;

(3) a dependence measure between the outcome and covariates conditional upon the treatment.

Examples of objects indicated in (3) include copulas and conditional odds ratios. The advantages of using the frugal parameterization are three-fold. Firstly, it gives a likelihood using a causally-relevant parameterization to the observational data. Secondly, it isolates the causal quantity of interest, $p^*_{Y(a)|\boldsymbol{C}}$, from the rest of the joint distribution, which means that we can directly target inference for this quantity, and treat the other parameters as a nuisance model. Finally, the frugal parameterization allows the non-causal distributions to differ between the RCT and observational data, for example, the joint distribution of covariates and treatment assignment, $p_{\boldsymbol{Z}A|\boldsymbol{C}}(\boldsymbol{z}, a \,|\, \boldsymbol{c})$. Under the frugal parameterization, we can factorize joint densities p_r:

$$p_r(y, \boldsymbol{z}, a \,|\, \boldsymbol{c}) = p_{r,\boldsymbol{Z}A|\boldsymbol{C}}(\boldsymbol{z}, a \,|\, \boldsymbol{c}) \cdot p^*_{Y|A\boldsymbol{C}}(y \mid a, \boldsymbol{c}) \cdot \varphi^*_{r,Y\boldsymbol{Z}|A\boldsymbol{C}}(y, \boldsymbol{z} \,|\, a, \boldsymbol{c}) \tag{2.2}$$

and similarly for the observational data density p_o. Let the vector $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\psi})$ denote the set of parameters. Here, θ is the set of shared causal parameters for $p^*_{Y|A\boldsymbol{C}}(y \mid a, \boldsymbol{c}; \theta)$, while γ and ψ parameterize the nuisance models for the RCT and RWD respectively. Additionally, to ensure a consistent estimation of the HTE function $\tau(\boldsymbol{c})$, we further require a correctly specified model for $p^*_{r,Y|A\boldsymbol{C}}(y \mid a, \boldsymbol{c}; \theta)$.

Assumption 2.4. The distribution of potential outcomes $P(Y(a) \mid \boldsymbol{C})$ in the randomized data is correctly specified by a parametric model $p^*_{r,Y|A\boldsymbol{C}}(y \mid a, \boldsymbol{c}; \theta)$.

A convenient choice of such specification is a linear model, although linearity is not required by our proposed method. Unlike requiring a correctly specified data model for the full joint distribution, Assumption 2.4 is strictly less stringent. Furthermore, it imposes a requirement only on the randomized data, but makes no restriction on the observational data model.

Remark 2.1. With Assumption 2.4, our inference takes place in the so-called $\mathcal{M}$-closed world. However, as famously stated by Box (1976), "all models are wrong." The assumption of a correct specification is unlikely to hold in practice. Instead, we conduct causal inference in the $\mathcal{M}$-open world where maximum likelihood estimation finds the element of the assumed model which minimizes the Kullback-Leibler (KL) divergence to the data. Additionally, in the case of misspecification, if treatment is binary, the ATE estimate constructed through marginalizing over all conditioning variables $\boldsymbol{C}$ will nonetheless be consistent.

We now apply this parameterization to specify the power posterior in (2.1).

2.4.2 Choosing the optimal influence factor η

Selection of η needs to be performed carefully: the observational data can be useful in reducing the variance; however, we do not want to contaminate our causal estimate with excessive bias by setting η too high. Ideally, η is chosen in a robust manner that is adaptive to the data, where we incorporate more information when the intrinsic bias in the estimate from the observational data, $\hat{\theta}_o$, is small and its variability is low, and vice versa.

In recent literature, several methods for selecting η have been proposed, mainly in the context of addressing model misspecification.

2.4.2.1 *SafeBayes*

Proposed by Grünwald and Van Ommen (2017), the SafeBayes algorithm is designed to address model misspecification by adaptively controlling the learning rate. The authors distinguish between *bad* and *benign* misspecification: bad misspecification arises if there is a non-negligible probability that, for a range of sample sizes, the predictive distribution remains substantially different from any of the distributions in model class $\mathcal{M}$; benign misspecification refers to the opposite case, where the model remains a reasonable approximation. *SafeBayes* is most effective under benign misspecification.

The algorithm selects η through minimizing a sequential cumulative risk measure:

$$\eta = \arg\min_{\eta} \sum_{i=1}^{n} \mathbb{E}_{\theta \sim \pi_{i-1}^{(\eta)}} [-\log p_\theta (x_i)],$$

where n is the size of the dataset, $\pi_{i-1}^{(\eta)}$ is the generalized posterior with power η given the first $i-1$ observations. *SafeBayes* can be very computationally expensive because parameter samples need to be simulated n times from different posteriors, and this sequential simulation needs to be repeated multiple times in the search for the minimizer η.

2.4.2.2 Expected information matching

Holmes and Walker (2017) proposed a method to solve for the power η by matching the expected information gain between Bayes and generalized Bayes updates:

$$\int I_\eta(x) f_0(x) \mathrm{d}x = \int I_1(x) f\left(x; \theta_0\right) \mathrm{d}x$$

where $f_0(x)$ is the true density and θ_0 is the value that minimizes the Kullback-Leibler divergence between $f_0(x)$ and the parametric family of densities. $I_\eta(x)$ denotes the expected information gain between $p_\eta(\cdot \mid x)$ and $p(\cdot)$, and the authors

suggested Fisher information and Kullback-Leibler divergence as convenient measures of information. Lyddon et al. (2019) followed a similar strategy and proposed to find η by matching the Fisher information of the η-powered generalized Bayesian posterior with that of the loss-likelihood bootstrap.

2.4.2.3 Frequentist coverage probability calibration

Another way to look at the effect of the power η is that it controls the spread: how much shall be learned from the data to narrow the spread of the posterior distribution? Syring and Martin (2019) proposed to select the η such that the η-powered generalized Bayesian posterior credible sets approximately achieve the frequentist coverage property. The essential assumption for their method is consistency, that is, the η-powered generalized Bayesian posterior centres around the true parameter θ. This assumption is unlikely to hold in our setting, where we assume that the observational data is biased due to hidden confounding.

However, these methods above primarily address misspecifications where the posterior is consistent and concentrated around the true values (Holmes and Walker, 2017; Syring and Martin, 2019; Lyddon et al., 2019), or deal with *benign* misspecifications related to heteroscedasticity (Grünwald and Van Ommen, 2017). These scenarios do not apply to our data fusion setting, where the bias caused by unobserved confounding does not diminish as the sample size increases.

2.4.2.4 Maximizing ELPD

Given our problem setting and objectives, we follow Carmona and Nicholls (2020) and select η by maximizing the expected log point-wise predictive density (ELPD):

$$\mathrm{ELPD}(\eta) = \mathbb{E}_{\tilde{\mathbb{X}}} \log p_\eta(\tilde{X} \mid \boldsymbol{x}_r, \boldsymbol{x}_o) = \int_{\tilde{\mathbb{X}}} p_t(\tilde{x}) \log p_\eta(\tilde{x} \mid \boldsymbol{x}_r, \boldsymbol{x}_o)\, d\tilde{x}, \tag{2.3}$$

where $p_\eta(\tilde{x} \mid \boldsymbol{x}_r, \boldsymbol{x}_o) = \int_\Gamma \int_\Theta p(\tilde{x} \mid \theta, \gamma) p(\theta, \gamma \mid \boldsymbol{x}_r, \boldsymbol{x}_o)\, d\theta\, d\gamma$ is the posterior predictive distribution indexed by η, and the expectation is with respect to the *true* data-generating process, p_t. As this *true* distribution is unknown, there are several ways to estimate the ELPD. A standard—though computationally very expensive—approach is to use leave-one-out cross-validation (LOO); the computational difficulty can be overcome by using an approximation based on the widely applicable information criterion (WAIC), which is shown to be asymptotically equal to LOO (Watanabe, 2010). The WAIC method simply uses the ordinary posterior to estimate the density of each observation, and subtracts the effective number of parameters, defined as

$$\widehat{\mathrm{ELPD}}(\eta) = \frac{1}{n_r} \sum_{i=1}^{n_r} \log \hat{p}_\eta(x_i \mid \boldsymbol{x}) - \hat{d}_{\mathrm{WAIC}},$$

where $\hat{p}_\eta(x_i \,|\, \boldsymbol{x})$ is estimated using the posterior samples of the parameters, and $\hat{d}_{\text{WAIC}}$ is the estimated effective number of parameters. Gelman et al. (1995) provide a mean-based and a variance-based definition for $\hat{d}_{\text{WAIC}}$, and consistent with Vehtari et al. (2017), we use the variance-based definition:

$$d_{\text{WAIC}} = \sum_{i=1}^{n_r} \text{Var}_{\text{post}} \left\{ \log p(x_i \mid \theta, \gamma) \right\},$$

where Var_{post} is the posterior variance of the log predictive density for each data point x_i in the randomized data. We propose to evaluate the ELPD on $\mathcal{D}_r$, as the randomization ensures the causal relationship in the experimental data is correct. We then apply a grid search in $[0, 1]$ to find the η^* maximizing the estimated ELPD.

Remark 2.2. It is worth noting the connection between ELPD and the KL divergence measure. We can rewrite (2.3) as

$$\begin{aligned} \text{ELPD}(\eta) &= \int_{\mathbb{X}} p_t(\tilde{x}) \log p_\eta(\tilde{x} \,|\, \boldsymbol{x}_r, \boldsymbol{x}_o)\, d\tilde{x} \\ &= \int_{\mathbb{X}} p_t(\tilde{x}) \left\{ \log \frac{p_\eta(\tilde{x} \,|\, \boldsymbol{x}_r, \boldsymbol{x}_o)}{p_t(\tilde{x})} + \log p_t(\tilde{x}) \right\} d\tilde{x} \\ &= -d_{\text{KL}} \left(p_t \,|\, p_\eta \right) - H(p_t), \end{aligned} \tag{2.4}$$

where $H(p_t) = -\int_{\mathbb{X}} p_t(\tilde{x}) \log p_t(\tilde{x})\, d\tilde{x}$ is the entropy of p_t, a constant independent of η. Essentially, choosing the η via maximizing ELPD is equivalent to selecting the posterior predictive distribution, indexed by η, that is closest to the true data distribution in KL-divergence. This confirms that ELPD is a proper scoring rule.

Our proposed method to select η is summarized in Algorithm 1.

Algorithm 1 Power likelihood η selection

```
Initialize η* ← 0 , ELPD* ← −∞
for i = 0, 1, 2, ..., N do
    Let η = i/N
    Sample (θ^(i), ψ^(i), γ^(i)) ~ π_η (θ, ψ, γ | X_r, X_o) using any appropriate sampler
    Compute ÊLPD(η) using posterior samples (θ^(i), ψ^(i), γ^(i))
    if ÊLPD(η) > ELPD* then
        ELPD* ← ÊLPD(η)
        η* ← η
    end if
end for
return η*
```

2.4.3 Normal approximations of the posterior distribution

The Bernstein-Von Mises theorem states that in a smooth, finite-dimensional model, as the sample size n approaches infinity, the posterior distribution of a parameter ϕ will converge to a normal distribution with mean $\hat{\phi}_n$ being the maximum likelihood estimator, and variance $\mathcal{I}_{\phi_0}^{-1}$, where $\mathcal{I}_{\phi_0}$ is the Fisher information matrix of the model at the true value ϕ_0 (van der Vaart, 1998). When the model is correctly specified, $\hat{\phi}_n$ converges to the true value ϕ_0 at the rate of $\sqrt{n}$. For a misspecified model, $\hat{\phi}_n$ instead converges to $\tilde{\phi}$, which minimizes the KL divergence from the model to the truth. In our context, for the combined parameter vector $\phi = (\theta, \gamma, \psi)^T \in \mathbb{R}^{m\times 1}$, where $m = p + d + k$ is the total dimension, the theorem leads to the following lemma:

Lemma 2.1. Given the MLE of ϕ in $\mathcal{D}_e$ and $\mathcal{D}_o$ as $\hat{\phi}_r$ and $\hat{\phi}_o$ respectively, both are asymptotically normal with $\sqrt{n_r}(\hat{\phi}_r - \phi^*) \to N(0, \boldsymbol{V})$ and $\sqrt{n_o}(\hat{\phi}_o - \phi^* - \delta) \to N(0, \boldsymbol{W})$, where ϕ^* is the true value of ϕ, δ represents the 'bias' and $\boldsymbol{V}, \boldsymbol{W} \in \mathbb{R}^{m\times m}$. For $\hat{\phi}_\eta$, the MLE under the η-powered likelihood, we find

$$\sqrt{n_r + \eta n_o}\left\{\hat{\phi}_\eta - \left(n_e \boldsymbol{V}^{-1} + \eta n_o \boldsymbol{W}^{-1}\right)^{-1}\left(n_r \boldsymbol{V}^{-1}\hat{\phi}_r + \eta n_o \boldsymbol{W}^{-1}\hat{\phi}_o\right)\right\}$$
$$\xrightarrow{d} \mathcal{N}\left(0, \left(\frac{n_r}{n_r + \eta n_o}\boldsymbol{V}^{-1} + \frac{\eta n_o}{n_r + \eta n_o}\boldsymbol{W}^{-1}\right)^{-1}\right).$$

Derivation of these quantities can be found in Appendix B.1. A consequence of the lemma is that we can approximate the posterior of ϕ using a normal distribution with mean $\hat{\phi}_\eta$ and variance $\mathcal{I}(\hat{\phi}_\eta)^{-1} = (n_r \boldsymbol{V}^{-1} + \eta n_o \boldsymbol{W}^{-1})^{-1}$. This result offers an intuitive interpretation: for asymptotically normal estimators, the η-exponent inflates the standard error of the observational estimate by $1/\sqrt{\eta}$, effectively down-weighting the observational sample size by a factor of η. A further corollary of this lemma is that for sufficiently large samples and a fixed bias, the optimal η scales with n_o^{-1}. This result is formalized and proved in Proposition 2.2 in the final section of this chapter. In simulations presented in Section 2.5, we use the sandwich estimators of V and W and Fisher information calculated using the R package `causl` (Evans, 2021).

Our proposed method, as described in Algorithm 1, involves iterative sampling from posterior distributions to find the optimal η. While Markov chain Monte Carlo (MCMC) is a standard approach for posterior sampling, it becomes computationally challenging for large datasets, such as we expect for $\mathcal{D}_o$. This is exacerbated when the parameter vector ϕ is high dimensional, which challenges sampling efficiency. Lemma 2.1 provides a solution for such cases. A nice property of this asymptotic

approximation is that the bigger the data sets, the better the approximation, which will clear the computational hurdle and make our method scalable for application to big data. Moreover, it allows us to work with summary statistics, i.e. the MLE and its variance, rather than requiring simultaneous access to both the complete experimental and external datasets. This is particularly valuable in contexts with privacy concerns or within a federated data framework, where full data sharing is neither feasible nor desirable.

Remark 2.3. In instances where no observational subjects receive treatment, the corresponding entries concerning treatment A in the variance-covariance matrix are undefined. In the Fisher information matrix $\mathcal{I}(\hat{\phi}_o)$, these entries are zero, indicating that it provides no information about the coefficients associated with treatment-related terms in the $p^*_{Y|AC}(y \mid a, c)$ model component.

2.4.4 Asymptotics

Intuitively, the $\hat{\eta}$ that maximizes ELPD in (2.4) depends on the magnitude of bias introduced by the observational dataset. The more biased $\mathcal{D}_o$ is, the less we should incorporate its influence. We are particularly interested in studying the asymptotics of $\hat{\eta}$ and the resulting estimator $\hat{\theta}_{\hat{\eta}}$. Similar to the discussion in Yang and Ding (2020), we assume that the bias in the observational data $\mathcal{D}_o$ due to hidden confounding is not fixed but depends on its sample size, even though this is not plausible in practice. Specifically, we consider two regimes of local alternative based on the rate of convergence of the bias to zero.

Theorem 2.1 (**Consistency**). *Let $\delta = \delta(n_o) = \delta^*/n_o^k$ denote the bias in $\mathcal{D}_o$.*

(a) If $0 \leq k < 1/2$ and $\delta^* \neq 0$, then $\hat{\eta}$ converges to 0, and $\hat{\phi}_{\hat{\eta}} = \phi^* + o_p(n^{-1/2})$.

(b) If $k \geq 1/2$ or $\delta^* = 0$ then $\hat{\eta}$ does not converge.

Theorem 2.1 suggests that when bias is asymptotically (statistically) non-negligible, the selected influence factor $\hat{\eta}$ converges to 0 and the estimator $\hat{\phi}_{\hat{\eta}}$ is $\sqrt{n}$-consistent. As a result, the MSE of the estimator $\hat{\phi}_{\hat{\eta}}$ converges to that from using the experimental data only. To provide intuition for this theorem, as n_o grows, the fraction of its influence we want to incorporate reduces and as n_r grows, the need to augment it with external data diminishes at the same time. Notably, for the special case where the bias is a constant ($k = 0$), the selected influence factor η converges to zero. However, when the observational data has asymptotically negligible bias, the selected $\hat{\eta}$ will not converge to zero. On the contrary, as bias diminishes close to zero, we will very likely infer that $\hat{\eta} = 1$ is the best value, as we restrict η to be between 0 and 1. The proof is presented in Appendix B.2.

2.5 Simulation

We followed the setup outlined in Example 2.1, observing two pre-treatment covariates: an obesity indicator C and a continuous measure of creatinine level Z; a binary treatment variable A indicating drug administration; and a continuous outcome Y which measures change in average blood glucose level. Additionally there exists an unobserved confounder U, affecting both the treatment and the outcome. The influence of U means that any causal estimate based on the assumption of unconfoundedness in the observational data will be biased. Specifically, we simulate the observational data $\mathcal{D}_o$ from the following data models:

$$U \sim \text{Bernoulli}(0.5) \qquad C \sim \text{Bernoulli}(0.5) \qquad Z \mid C \sim N(\mu_z, 1)$$
$$A \mid C, Z, U \sim \text{Bernoulli}(\mu_a) \qquad Y(a) \mid C \sim N(\mu_y, 1),$$

where, $\mu_z = 1 + C$, $\operatorname{logit} \mu_a = -3 + C + Z + C\,Z + \omega U$, and $\mu_y = 1 + C + 0.1\,A + 0.1\,C\,A + \omega U$. In the simulations, ω is varied to represent differing amounts of unobserved confounding. Additionally, let $\varphi^*_{YZ|AC}$, the dependence structure between Y and Z given A and C, be a conditionally bivariate Gaussian copula, with correlation parameter $\rho_a = 2\,\text{expit}(1 + 2.5a) - 1$.

We use the same parameterization for the experimental data $\mathcal{D}_r$ with two adjustments: (i) treatment is randomly assigned and hence replace the mean of A with $\mu_a = 0.5$, and (ii) the outcome Y is assumed independent of any hidden confounding, so we set $\omega = 0$. This simulation framework is designed to reflect a realistic data fusion scenario where, the RCT is free of confounding while there is an unmeasured factor U, such as the access to and quality of healthcare, that affects both treatment and outcome, as well as introducing extra variability in the outcome. We set the sample sizes to 1,000 and 500 for the observational and experimental data, respectively.

2.5.1 Moderating the bias-variance trade-off

Figure 2.2 presents three scenarios where unmeasured confounding has an influence of different magnitudes: no influence ($\omega = 0$), small influence ($\omega = 0.5$) and large influence ($\omega = 1$). The first column shows the average negative ELPD at each value of η. Our proposed method searches for the η that maximizes ELPD, which corresponds to the lowest point on the curve. The second and third columns show the MSE of the ATE and CATE estimates evaluated on $\mathcal{D}_e$ at different values of η. We decompose MSE into variance (in blue) and squared bias (in red) to give an intuitive illustration of the moderating effect of η on the bias-variance trade-off: as

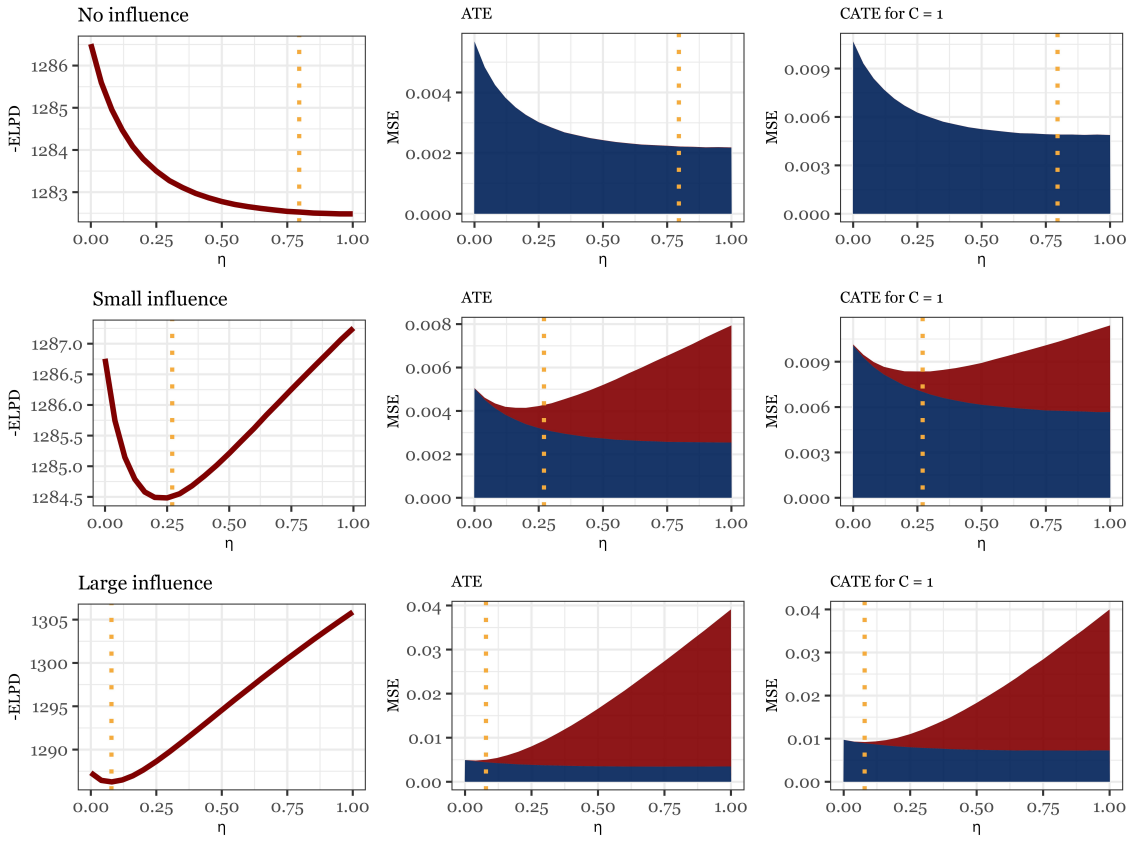


Figure 2.2: Simulation results for scenarios where U has no influence ($\omega = 0$), small influence ($\omega = 0.5$) and large influence ($\omega = 1$) on T and Y. The results are averaged across 500 sets of synthetic datasets. Left column: average negative ELPD with $\eta \in [0, 1]$. The yellow dotted lines mark the average η selected. Middle column: MSE of the ATE estimate. Right Column: MSE of the CATE estimate in the subgroup with $C = 1$. The MSE is decomposed into bias (red-shaded area) and variance (blue-shaded area).

we increase η, we include more information from $\mathcal{D}_o$ and hence the variance reduces, yet at the expense of an increase in bias.

When U has no influence on T and Y, that is, $\mathcal{D}_o$ is not biased and is just 'as good as' $\mathcal{D}_e$, the average η selected is around 0.80 (median: 0.95, 25% quantile: 0.60, 75% quantile: 1); this is reasonable, because in this context we want to include as much information in $\mathcal{D}_o$ as possible. Unsurprisingly, the corresponding MSEs of ATE and CATE estimates are both considerably lower than if we do not include $\mathcal{D}_o$ at all. When U has a small influence on T and Y, our method chooses a smaller η, averaging around 0.27 (median: 0.25, 25% quantile: 0.20, 75% quantile: 0.35). Again, the corresponding MSEs of ATE and CATE estimates are both lower than at $\eta = 0$. When U has a larger influence, that is, $\mathcal{D}_o$ is more biased, our method responds by selecting an even smaller η. Although the corresponding MSE of the ATE estimate is almost unchanged, we achieve a reduction in MSE for the CATE estimates when compared with $\eta = 0$.

2.5.2 Inference

By the Bernstein-von-Mises theorem, under the condition that the prior is continuous and positive in a neighborhood around the MLE, the Bayesian credible intervals approximately align with frequentist confidence intervals as the sample size approaches infinity. Nevertheless, it is important to recognize that the MLE derived from an η-powered likelihood of a combined dataset may not coincide with the MLE from the experimental data alone. Consequently, there is no theoretical assurance that the combined MLE's distribution will converge around the true parameters of the experimental data, and the credible interval may not offer the nominal coverage rates, particularly in the presence of hidden confounding.

Despite the lack of theoretical guarantee, the empirical coverage rates presented in Table 2.2 closely match the nominal rate of 95%, for the ATE τ, CATE $\tau(1)$, and $\tau(0)$, under varying levels of confounding, ω. Moreover, while using the experimental data alone yields test powers for τ, $\tau(1)$ and $\tau(0)$ of approximately 57%, 52%, and 18% respectively, data fusion substantially enhances these figures, delivering the efficiency gain as desired. Overall, we achieve reductions in MSEs across a range of confounding scenarios.

2.6 PIONEER 6 application

We demonstrate the potential benefit of the proposed method to evaluate the cardiovascular benefits of semaglutide, a novel glucagon-like peptide-1 (GLP-1) analogue,

ω	η	τ		CATE $\tau(C=1)$		CATE $\tau(C=0)$		Reduction in MSE		
		Coverage	Power	Coverage	Power	Coverage	Power	τ	$\tau(1)$	$\tau(0)$
0	0.80	95%	85%	95%	78%	96%	31%	54%	49%	63%
0.1	0.78	98%	87%	97%	80%	95%	36%	58%	53%	63%
0.2	0.79	94%	89%	94%	80%	95%	41%	49%	45%	57%
0.3	0.69	94%	91%	93%	85%	94%	41%	43%	38%	49%
0.4	0.47	91%	89%	95%	80%	94%	40%	26%	28%	39%
0.5	0.27	92%	83%	94%	76%	95%	32%	14%	16%	34%
0.7	0.13	94%	73%	94%	65%	94%	27%	10%	10%	15%
0.8	0.10	93%	78%	96%	67%	94%	29%	5%	17%	6%
1	0.08	93%	72%	95%	61%	94%	26%	1%	8%	9%
No fusion		94%	57%	94%	52%	94%	18%	-	-	-

Table 2.2: Simulation results for coverage probability of 95% credible interval, power and reduction in MSE for ATE τ, CATE $\tau(c)$ estimations.

in patients with type 2 diabetes, using the PIONEER 6 trial and a large US health claims dataset. PIONEER 6 is a randomized, double-blind trial that aims to assess whether oral semaglutide can reduce the risk of major adverse cardiovascular events (MACE). In this data fusion study, we analyze the occurrence of MACE within one year, targeting the risk difference as the estimand. Analyzing the PIONEER 6 data alone, we found an unadjusted risk difference of -1.32% between treatment and control groups, with a 95% confidence interval of $[-2.71\%, 0.06\%]$, which includes zero. Adjusting for key baseline covariates, ATE estimate of -1.24% with a narrowed 95% CI of $[-2.55\%, 0.06\%]$, representing a 5.25% reduction in width. Despite this improvement, as the point estimate moved slightly, the confidence interval still crosses zero. Given this result and the established non-inferiority of oral semaglutide to placebo, the traditional next step is to proceed with a superiority RCT. Instead, we explore using RWD to enhance the PIONEER 6 findings.

2.6.1 Target trial emulation

We emulate the PIONEER 6 trial using real-world data sourced from Optum's informatics® Data Mart, consistent with approaches from a recent study by Dang et al. (2023a). The dataset compiles administrative health claims for members with commercial health insurance in the US, and includes information on patient enrolment, medical and pharmacy claims, inpatient admissions, and laboratory results. We followed the framework set out in Hernán and Robins (2016) to construct a RWD sample that emulates PIONEER 6. We applied the trial's inclusion criteria to define our RWD sample and established the start of DPP4 inhibitor treatment as 'time zero' to accurately mimic trial conditions and control for biases. For covariate adjustment, we employed regularization techniques and Gradient Boosted

Machines to identify important predictors of treatment, outcome, and for managing the 15.90% missingness in 1-year MACE outcomes. This emulation process aims to improve the comparability between the RCT and the RWD, thereby increasing the chance that Assumption 3 will hold.

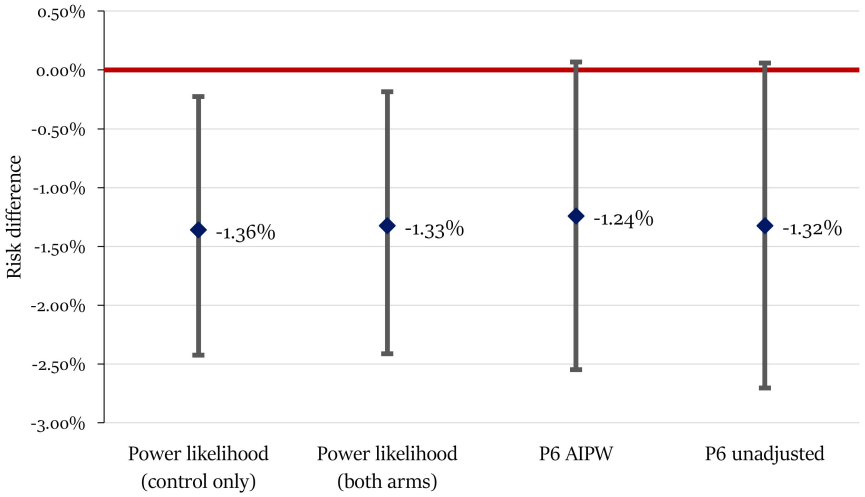
2.6.2 Design 1: Incorporating control arm only

We first focus on the design that incorporates only untreated patients from the RWD. The feature selection methods outlined in Section 3.4.1 led to the inclusion of six key variables in our model: age, serum albumin level, serum creatinine level, low-density lipoprotein (LDL), high-density lipoprotein (HDL) and history of heart failure. To address missingness in the outcome, we employed multiple imputation (Gelman et al., 1995). Across five sets of imputations, the average selected η is 0.74. Pooling the corresponding five sets of η-powered posterior samples, the ATE estimated from our proposed power likelihood method is -1.36%, with a 95% credible interval of $[-2.42\%, -0.23\%]$. This represents a 20.5% reduction in width compared to the unadjusted analysis of the PIONEER 6 data. Notably, this enhanced efficiency resulted in a 95% credible interval for the effect on 1-year MACE risk that does not exceed zero, despite the minimal shift in the point estimate.

2.6.3 Design 2: Incorporating both arms

Incorporating both arms highlighted the issue of treatment effect heterogeneity. We aim to neutralize the difference in ATE between the RCT and RWD attributable to the varied distribution of effect modifiers, so that Assumption 2.3 is more likely to hold, thereby increasing potential efficiency gains from data fusion. Applying the Causal Forest method (Wager and Athey, 2018) on the PIONEER 6 data, we identified LDL, insulin use, serum creatinine, and Hemoglobin A1c (HbA1c) as the top variables influencing treatment effect heterogeneity, although the differences were not statistically significant, with the smallest p-value being 0.17, likely due to the limited sample size. Even if the effects were significant, it is crucial to understand that these findings from Causal Forest, such as insulin use, should not be interpreted causally but rather as predictors of treatment effect variation. After including these effect modifiers in the matching process, we obtained a 1:1 matched final RWD sample.

Across five sets of imputations, the η selected is 0.6 on average, and the estimated ATE is -1.33% with a 95% credible interval of $[-2.41\%, -0.19\%]$, a 19.5% reduction in width from the confidence interval from analyzing PIONEER 6 alone and the CI no longer crosses zero.



	Risk difference (%) (95% CI)	% of P6 (unadjusted) CI length
Power likelihood (control only)	-1.36 $[-2.42, -0.23]$	79.54%
Power likelihood (both arms)	-1.33 $[-2.41, -0.19]$	80.62%
P6 AIPW	-1.24 $[-2.55, \quad 0.06]$	94.75%
P6 unadjusted	-1.32 $[-2.71, \quad 0.06]$	100.00%

Figure 2.3: Estimated difference in 1-year MACE risk and 95% confidence/credible interval. P6 unadjusted: Wald test of the MACE rate in the treatment and the placebo arms of PIONEER 6. P6 AIPW: AIPW estimate after adjusting for key covariates in PIONEER 6.

In this case study, as shown in Figure 2.3, the two data fusion designs yielded consistent results: the ATE's point estimates closely aligned with the mean difference observed in PIONEER 6, yet the confidence intervals effectively shrank by approximately 20%, demonstrating the effectiveness of our method in robustly augmenting the RCT to increase power.

Further details on how the RWD sample was constructed from the Optum data, along with the full implementation of the power likelihood method, are provided in Appendix C.2. In Chapter 3, we return to this application and apply a range of alternative data fusion methods to enable a comparative evaluation.

2.7 Conclusion and future work

Compared to many existing data fusion methods, our proposed power-likelihood approach offers several advantages: 1) adaptivity to alternative causal estimands such as odds ratios, 2) flexibility to incorporate external data with either both arms or only the controls, 3) automatic uncertainty quantification as a Bayesian method, empirically shown to yield nominal coverage rates approximately. Importantly, our proposed approach can be extended to virtually any Bayesian model for observational and experimental data, beyond the specific causal models described. The power parameter can generally be selected based on standard MCMC processes.

In Algorithm 1 we propose a simple grid search procedure to select η by maximizing the ELPD. While effective in the one-dimensional case where η is a scalar, this strategy becomes infeasible in higher dimensions. Theoretically, our method could be extended to incorporate multiple external sources, each associated with its own learning rate. This naturally leads to a multidimensional generalization of η, which complicates optimization and raises questions about how best to characterize and estimate η in such settings.

To address this, we explore whether the ELPD can be expressed as an explicit function of η and maximized analytically. In the single-source case, we have the following result:

Proposition 2.2. *Let $\hat{\phi}_r$ and $\hat{\phi}_o$ be the MLEs based on experimental and observational data, with associated variance matrices $\boldsymbol{V}$ and $\boldsymbol{W}$. Let $\hat{\phi}_\eta$ and Σ_η denote the η-powered posterior mean and covariance, and $\boldsymbol{A} = \frac{\partial^2 p_t(\tilde{x}|\phi)}{\partial \phi^2}\Big|_{\phi=\hat{\phi}_r}$, and suppose $\mathbb{E}(\boldsymbol{A})$ exists. Then the ELPD can be approximated as:*

$$\mathrm{ELPD}(\eta) \approx \mathbb{E}\big[\log p_t(\tilde{X} \mid \hat{\phi}_r)\big] + f(\eta), \tag{2.5}$$

where

$$f(\eta) = (\hat{\phi}_\eta - \hat{\phi}_r)^\top \mathbb{E}(\boldsymbol{A})(\hat{\phi}_\eta - \hat{\phi}_r) + \operatorname{tr}(\mathbb{E}(\boldsymbol{A})\Sigma_\eta). \tag{2.6}$$

The approximate optimizer of the ELPD is thus given by:

$$\eta^* = \arg\max_\eta f(\eta).$$

The proof is provided in Appendix B.3. This result shows that the ELPD can be decomposed into a constant term—reflecting the entropy of the true data-generating distribution evaluated at $\hat{\phi}_r$—and a function of η that resembles the squared-bias and variance term. Optimizing η therefore reduces to maximizing the analytic function $f(\eta)$, eliminating the need for computationally expensive grid search. This allows for efficient selection of η via gradient-based optimization.

Furthermore, by substituting the expressions for $\hat{\phi}_\eta$ and Σ_η from Lemma 2.1, we observe that all the η terms include n_o as a factor. This implies that, all else being equal, the optimal η^* scales with n_o. Intuitively, as the size of the observational data doubles, the optimal proportion of information drawn from it halves. This supports the scaling behaviour discussed in Section 2.4.3.

A natural extension of this framework is to allow η to be a vector rather than a scalar, assigning separate learning rates to different external sources or strata within a source. Equation (5.1) generalizes accordingly, and $f(\eta)$ becomes a multivariate function over $\eta \in \mathbb{R}^K$, where K denotes the number of external sources or strata.

This extension is practically useful in at least two settings:

(i) Subgroup-specific learning rates: For instance, in our application using Optum data to augment PIONEER 6, we may suspect that certain subgroups (e.g. insulin users) are more comparable to their counterparts in the PIONEER 6 population than others. Assigning different η values to each subgroup allows the model to borrow strength differentially—down-weighting less comparable segments.

(ii) Multiple external sources: When multiple datasets are available (e.g. historical trials), this extension enables simultaneous integration of heterogeneous evidence, each with its own calibrated influence.

This generalization is straightforward to implement and offers substantial practical benefits by improving both flexibility and robustness in real-world data fusion settings.

While these extensions offer promising directions for improving the flexibility and scope of our approach, several limitations remain that merit further consideration.

A common limitation of Bayesian methods is the difficulty in optimizing the inference of a particular parameter of interest. In contrast, frequentist methods such as targeted learning using TMLE (van der Laan and Rubin, 2006) are designed precisely for this purpose. In our proposal, we take a typical Bayesian route and select η based on expected prediction quality instead of directly targeting the CATE estimates. The loss of efficiency is more pronounced as covariates that are weakly correlated, or uncorrelated with the outcome are included in the model. Again, this issue is not unique to our method. As mitigation, we recommend employing variable selection techniques to retain only covariates that are predictive of the treatment and outcomes, as demonstrated in the PIONEER 6 data fusion study. Another potential issue is model misspecification, as our method relies on correctly specifying the potential outcome model $P(Y(a) \mid \boldsymbol{C})$ (Assumption 2.4). This is closely related to correctly specifying the dependence component. For example, assuming a multivariate Gaussian copula might not always be accurate, leading to misspecified potential outcome distributions and inconsistent estimates. Future research could explore incorporating more flexible parameterization of the dependence structure to enhance robustness.

While this chapter focused on a specific Bayesian framework for data fusion, alternative approaches address the problem from different methodological perspectives. In the next chapter, we review and compare a broader set of data fusion methods through simulation and applied examples.

Chapter 3

Evaluating data fusion methods

Abstract. *Integrating real-world data (RWD) with randomized controlled trials (RCTs) is increasingly recognized as a strategy for improving the efficiency and applicability of causal inference. While numerous data fusion methods have been proposed—including the power-likelihood approach developed in Chapter 2—selecting an appropriate approach for a given research problem remains challenging. This chapter provides a systematic comparison of methods aimed at leveraging RWD to enhance statistical efficiency in causal effect estimation. We examine their assumptions, robustness to bias from unmeasured confounding, and implementation complexity. Through extensive simulation studies designed to reflect practical scenarios, we assess each method's empirical performance, including its bias–variance trade-off under finite samples. Our findings reveal key patterns in how different methods manage the tension between efficiency gains and susceptibility to bias. We also revisit the PIONEER 6 case study, using the same setup as in Chapter 2, to enable direct comparison across methods. By offering critical insights into the strengths and limitations of each method, this chapter supports more informed methodological choices in the application of data fusion for causal inference.*

Contents

3.1 Introduction 55
3.2 Existing methods 57
3.2.1 Test-then-pool 57
3.2.2 Weighted combination 59
3.2.3 Shrinkage 61
3.2.4 Bias correction 63
3.2.5 Prognostic adjustment 64
3.2.6 A summary table for application 65
3.3 Simulation studies 68
3.3.1 Simulation setup 68
3.3.2 Augmenting both arms of the RWD 69
3.3.3 Augmenting only the control arm of the RWD 78
3.4 Real-world application: PIONEER 6 Trial 82
3.4.1 Target trial emulation 83
3.4.2 Initial analysis: PIONEER 6 alone 85
3.4.3 Design 1: Incorporating control arm only 87
3.4.4 Design 2: Incorporating both arms 87
3.5 Summary 88

3.1 Introduction

Randomized controlled trials (RCTs) are the gold standard for establishing causality, but they are often limited by small sample sizes, narrow eligibility criteria, short follow-up periods, and high costs. These limitations can reduce both statistical power and generalizability. As outlined in Chapter 1, real-world data (RWD)—such as electronic health records, administrative claims, and registries—are increasingly used to complement RCTs, offering broader coverage and longer-term outcomes, but at the cost of potential confounding due to the absence of randomization.

To leverage the complementary strengths of these data sources, a growing number of methods have been developed for data fusion, integrating RCTs and RWD to improve estimation of causal effects. Chapter 2 introduced one such method, the power-likelihood approach, which robustly combines information to improve efficiency while protecting against bias. In this chapter, we take a broader view, reviewing and comparing a range of existing data fusion methods, including the power-likelihood method, with a focus on statistical efficiency.

Efficiency gains are a central motivation for data fusion: augmenting RCTs with RWD can increase precision in estimating treatment effects, particularly for underpowered trials or hard-to-reach subgroups. However, the use of RWD introduces potential bias due to unmeasured confounding. To mitigate this, some methods restrict integration to the control arm of the RWD, treating it as an external or historical comparator. Approaches in this class include test-then-pool designs (Viele et al., 2014), Bayesian dynamic borrowing methods (Ibrahim and Chen, 2000; Hobbs et al., 2012; Schmidli et al., 2014), and prognostic score adjustment (Schuler et al., 2021).

When both treatment and control arms are available in the RWD, incorporating both can yield greater efficiency gains—but with heightened risk of bias. Several methods attempt to manage this bias–variance trade-off through different strategies: James–Stein shrinkage estimators (Green et al., 2005; Rosenman et al., 2023a), semiparametric estimation frameworks (Yang et al., 2020; Li and Luedtke, 2023; Parikh et al., 2023), weighted averaging of estimates from RCT and RWD (Oberst et al., 2023; Chen et al., 2021), and explicit bias-correction techniques (Kallus et al., 2018; Yang and Ding, 2020; Yang et al., 2020; Cheng and Cai, 2021; van der Laan et al., 2024).

Although several recent reviews have surveyed this landscape—such as Colnet et al. (2024), Shi et al. (2023), and Brantner et al. (2023)— these discussions predominantly focus on theoretical properties rather than empirical performance. There remains limited evidence comparing these methods head-to-head under realistic sample sizes, data structures, and assumption violations. This chapter aims to address that gap.

We conduct an extensive simulation study to evaluate the empirical performance of a wide set of data fusion methods. All methods are placed on a level playing field, on the same set of simulated data. We assess each method‘s behaviour in terms of bias, variance, mean squared error, and coverage. To complement the simulations, we revisit the PIONEER 6 application in Chapter 2, applying the reviewed methods and compare their performance.

The remainder of this chapter is organized as follows: Section 3.2 reviews the fusion methods designed to achieve efficiency gains. Section 3.3 presents the results of our simulations, offering an in-depth analysis and interpretation of each method's performance. Section 3.4 reports findings from the real-world case study of the PIONEER 6 trial. Finally, Section 3.5 concludes with a discussion and offers concrete advice for practitioners, guiding them in selecting the most suitable method for their specific purposes.

3.2 Existing methods

This section provides a detailed review of recent data fusion techniques aimed at efficiency gains. As there is no universally accepted taxonomy for these methods, we organize them by their overarching strategy—the *philosophy* they adopt to leverage complementary information across data sources. Bayesian dynamic borrowing approaches, including those based on power priors, were reviewed in Chapter 2 in the context of the power-likelihood method and are not repeated here.

Alternative grouping criteria, such as whether a method assumes a parametric model, incorporates both treatment and control arms from RWD or only the control, or targets CATE versus ATE, are also relevant but secondary to this conceptual framing. Many methods share overlapping assumptions about the data, including those discussed in Chapter 2. We state key assumptions where first introduced and refer back to them as needed.

In the latter part of this section, we present a table summarizing the various characteristics of each method.

3.2.1 Test-then-pool

Test-then-pool style approaches start with the null hypothesis of equality between the causal estimates from multiple sources and only pool if the null hypothesis is not rejected. These approaches are commonly applied in meta-analyses or multi-site studies, where results from several studies are combined. Test-then-pool is the most basic form of dynamic borrowing (Viele et al., 2014). Typically, test-then-pool approaches require the researcher to specify the size of the hypothesis test, say $\alpha = 0.025$, which offers a way to cap the Type I error inflation introduced by the external data. Essentially, test-then-pool is an *all or nothing* approach and a major issue with such methods is that when the experimental data is small, the hypothesis test for any discrepancy is under-powered and rarely gets rejected, resulting in the observational data being pooled even if it is biased (Li et al., 2020).

The *Elastic Integration* approach proposed by Yang et al. (2023) aims at a semi-parametric efficient estimation of the HTE function. The method requires the standard assumptions of RCT ignorability and representativeness, corresponding to Assumption 2.1 and 2.2. Additionally, for the RWD, they consider transportability (Assumption 2.3) and ignorability, stated as below:

Assumption 3.1. (RWD ignorability) (i) **Unconfoundedness** $Y(a) \perp\!\!\!\perp A \mid W, S = 0$ for $a \in \{0, 1\}$ (ii) **Positivity** $0 < \mathbb{P}(a = 1 \mid W, S = 0) < 1$ for all $(W, S = 0)$.

Recognizing that these two assumptions for the RWD may be idealistic, they constructed a test statistic $T = \varphi(V)^T \widehat{\Sigma}^{-1} \varphi(V)$ to simultaneously falsify Assumption 2.3 and 3.1, where

$$\varphi(V) = \frac{1}{\sqrt{n_o}} \sum_{i \in \mathcal{O}} \widehat{S}_{\mathrm{O},\widehat{\psi}_{\mathrm{E}}}(V_i)$$

and $\widehat{S}_{\mathrm{O},\widehat{\psi}_{\mathrm{E}}}(\cdot)$ is the estimated semi-parametric efficiency score (SES) for an assumed HTE function ψ. For a detailed formulation of the SES, see Section 3.2 of Yang et al. (2020). If Assumption 2.3 and 3.1 hold, then the test statistic T should be close to zero and a large T indicates a violation. The authors proposed to only use the RCT if T exceeds a threshold c_γ, otherwise, pool the two datasets together. They proposed a data-adaptive procedure to select the threshold c_γ to minimize MSE.

Dang et al. (2022) proposed an *Experiment-Selector Cross-Validated TMLE* (ESCVTMLE) algorithm which splits the pooled data into folds, and employs a cross-validation approach to select between two experiment designs: RCT only, and pooled RCT + RWD, based on potential MSE reduction. This can be broadly described as a test-then-pool approach, though it does not involve specific hypothesis testing. According to their notation, the *selector* operates as follows:

$$s_n^\star = \underset{s}{\operatorname{argmin}} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + \left(\hat{\Psi}_s^{\#}(P_n)\right)^2.$$

Intuitively, though maybe imprecisely, $s_n^\star$ is the *selector* to decide whether to pool the RWD. $\frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n}$ represents the asymptotic variance derived from the efficient influence curve, and $\left(\hat{\Psi}_s^{\#}(P_n)\right)^2$ is the estimated squared bias. Essentially, the decision to pool a certain fold of the RWD is based on minimizing the MSE.

Moreover, when a negative control outcome (NCO) is available, they innovatively proposed to include the difference in NCO into the squared-bias term to improve the estimation of bias:

$$s_n^{\star\star} = \underset{s}{\operatorname{argmin}} \frac{\hat{\sigma}^2_{D^*_{\Psi_s}}}{n} + \left(\hat{\Psi}_s^{\#}(P_n) + \hat{\Phi}_s(P_n)\right)^2.$$

Here, $\hat{\Phi}_s(P_n)$ represents the estimated difference in the NCO between the treatment and control groups in the pooled RWD + RCT dataset. In an RCT, or in RWD without confounding, if the NCO is valid, $\Phi_s(P_n)$ should be zero. Thus, any deviation of $\hat{\Phi}_s(P_n)$ serves as a proxy for hidden confounding.

For the inclusion of an NCO to function as intended, two conditions must be met.

Condition 3.1. (U-compatibility) The unmeasured factors confounding the treatment-outcome relationship are the same as those confounding the treatment-NCO relationship. In other words, the unmeasured confounders are fully represented by the NCO, with no additional confounders between the NCO and treatment.

Condition 3.2. (Same direction) The unmeasured factors that affect the treatment-outcome relationship should affect the treatment-NCO relationship in the same direction. Otherwise the two confounding biases may cancel each other out.

Scale is another important consideration. Any significant difference in the scale of the outcome and the NCO can alter the balance between the two biases. For example, in a case study where the authors applied the ES-CVTMLE method to study the effect of an anti-diabetic drug (Dang et al., 2023a), they used rate of fracture at the NCO, which was around 0.7%, compared to a primary outcome rate of about 4%. In such cases, the inclusion of an NCO might not influence the trade-off by much.

In the simulations presented in Section 3.3, we examine the sensitivity of the ES-CVTMLE estimator to the use of an NCO, as well as to violation of the U-compatibility condition.

An advantage of the ES-CVTMLE is its flexibility to combine either both arms of the RWD or only the control arm. However, one limitation is the interpretation of the estimand. For instance, if the selector combines the RWD in 60% of cross-validation folds, it becomes challenging to interpret the underlying population upon which the treatment effect is assessed.

3.2.2 Weighted combination

An alternative strategy combines RCT and RWD treatment effect estimators, $\hat{\tau}_r$ and $\hat{\tau}_o$, via a weighted average: $\hat{\tau}(\lambda) = (1-\lambda)\hat{\tau}_r + \lambda\hat{\tau}_o$, where $\lambda \in [0,1]$ represents the weight assigned to the RWD effect estimator. This approach assumes that $\hat{\tau}_r$ is a consistent estimator of the true treatment effect, i.e. $\mathbb{E}[\hat{\tau}_r] = \tau$, while $\hat{\tau}_o$ may be biased, with bias denoted by δ. Candidate estimators for $\hat{\tau}_r$ and $\hat{\tau}_o$ include outcome regression, AIPW and TMLE.

Oberst et al. (2023) proposed selecting the optimal weight λ^* by minimizing the MSE of the combined estimator,

$$\begin{aligned}\lambda^* &= \underset{\lambda\in[0,1]}{\arg\min}\, \mathbb{E}\left[\left(\hat{\tau}(\lambda) - \tau\right)^2\right] \\ &= \frac{\mathrm{Var}(\hat{\tau}_r) - \mathrm{Cov}(\hat{\tau}_r, \hat{\tau}_o)}{\delta^2 + \mathrm{Var}(\hat{\tau}_r - \hat{\tau}_o)}.\end{aligned}$$

To implement this in practice, they proposed using plug-in estimates for the variance and covariance terms. In particular, they estimate δ by $\hat{\delta} = \hat{\tau}_o - \hat{\tau}_r$ and use this to compute an estimate $\hat{\lambda}$ for the MSE-minimizing weight. Notably, while $\hat{\tau}_r - \hat{\tau}_o$ provides an unbiased estimate for δ, the plug-in estimator $\hat{\lambda}$ is not unbiased for λ^*, due to Jensen's inequality. Specifically, the plug-in estimator $\hat{\lambda}$ underestimates λ^* on average when $\text{Cov}(\hat{\tau}_r, \hat{\tau}_o) = 0$ due to the convexity of the expression. As a result, when the true bias δ is small, the method may overly discount the observational data relative to what is theoretically optimal. By design, this method requires that both the treatment and control arms are present in the RWD. Thus, its applicability is limited to settings where RWD includes only control patients, a common scenario in real-world clinical studies.

Chen et al. (2021) introduced a dichotomous weighted average approach with a preliminary hypothesis testing step to account for bias. The approach hinges on a comparison of bias against a pre-determined threshold. If the bias is below this threshold, the weight is selected to optimize the MSE, as though there is no bias at all; otherwise, the integration is optimized with an estimate of the bias, similar to the MSE-minimizing estimator in Oberst et al. (2023). Their dichotomous procedure to estimate bias sets

$$\hat{\delta}_\gamma = \text{sgn}(\hat{\delta}) \left(|\hat{\delta}| - \gamma \cdot \sqrt{\hat{\sigma}_r^2 + \hat{\sigma}_o^2} \right)_+ ,$$

where $A_+ := \max\{A, 0\}$. The selection of γ, a critical hyper-parameter, impacts the extent of bias adjustment, with a smaller γ means that the bias term is less aggressively shrunk to zero and thus is more conservative. The authors proved that, under a certain data generating process, their estimator achieves a minimax optimal rate, up to a poly-log factor, of convergence to the oracle, where the true bias δ is known. However, the determination of γ's value, which significantly influences the combination and lacks a predefined range, presents a practical challenge due to its arbitrary nature and the difficulty in justifying its selection to stakeholders.

Two significant advantages of the two described linear combination methods include: i) ease of implementation—requiring only $\hat{\tau}_r$, and $\hat{\tau}_o$ and their variances, which can be obtained from estimators like IPW. The methods only require straightforward calculations with minimal coding and does not demand extensive computational resources; ii) flexibility—the simplicity of these approaches allows for easy adaptation to alternative estimands, such as the odds ratio, by merely adjusting the definitions and estimators of τ_r, and τ_o. This inherent flexibility makes the method adaptable to scenarios where the outcome is binary.

Two limitations of the two methods are: i) uncertainty quantification—their model-free nature implies that confidence intervals cannot rely on assumptions inherent in parametric models. Neither set of authors made any asymptotic claims about the variance, leading users to resort to nonparametric bootstrap methods for inference; ii) applicability–by design, the two methods can only apply to integrating RWD with both treatment and control groups are available. This restricts their utility in scenarios where treatment, such as a novel drug, is not yet available in the market and absent in the RWD.

3.2.3 Shrinkage

Stein's Paradox demonstrates that, for three or more parameters, there exists a combined estimator that has a lower overall expected MSE than the traditional method of estimating each parameter independently (Stein, 1956; James and Stein, 1961).

Theorem 3.1. *(Stein's paradox) Let* $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_p)$*, for* $p \geq 3$*, the estimator* $\hat{\boldsymbol{\theta}}_0(\boldsymbol{X}) = \boldsymbol{X}$ *is inadmissible; the James–Stein shrinkage estimator* $\hat{\boldsymbol{\theta}}_{\mathrm{JS}}(\boldsymbol{X})$ *strictly dominates* $\hat{\boldsymbol{\theta}}_0(\boldsymbol{X})$ *under quadratic loss:*

$$\hat{\boldsymbol{\theta}}_{\mathrm{JS}}(\boldsymbol{X}) \;=\; \left(1 - \frac{(p-2)\sigma^2}{\|\boldsymbol{X}\|^2}\right)_+ \boldsymbol{X}, \qquad (a)_+ := \max\{0, a\}.$$

This result means that we gain in terms of a lower MSE by simultaneously estimating multiple parameters.

Green and Strawderman (1991) and Green et al. (2005) extended this idea to combining unbiased and possibly biased estimators and proposed an estimator that shrinks the biased estimators towards the unbiased estimators:

$$\hat{\boldsymbol{\tau}}_{\mathrm{S1}} = \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} + \left(\boldsymbol{I}_K - \frac{a\boldsymbol{\Sigma}_r^{-1}}{\left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)^T \boldsymbol{\Sigma}_r^{-2} \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)}\right)_+ \cdot \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right),$$

where $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ and $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ are now vectors of dimension K, and each element is the ATE from one of K strata in the corresponding dataset. $\boldsymbol{\Sigma}_r$ is the variance-covariance matrix of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$; note that the components of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ (and $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$) are assumed mutually independent so $\boldsymbol{\Sigma}_{\boldsymbol{r}}$ is a diagonal matrix. The shrinkage parameter a is customarily set as $K - 2$. The plus sign denotes taking the element-wise positive part of the matrix.

In the same vein, Rosenman et al. (2023a) proposed a general procedure for deriving shrinkage estimators in this setting, by minimizing Stein's Unbiased Risk Estimate (SURE).

Definition 3.1. (Stein's Unbiased Risk Estimate) Consider an estimator of $\boldsymbol{\theta}$ of the form:

$$\hat{\boldsymbol{\theta}}_0(\boldsymbol{X}) = \boldsymbol{X} + g(\boldsymbol{X}),$$

then

$$\text{SURE}(g, \boldsymbol{X}) = p\sigma^2 + \|\boldsymbol{g}(\boldsymbol{X})\|^2 + 2\,\Sigma_{j=1}^{p}\sigma\partial_j g_j(\boldsymbol{X})$$

is unbiased for $\mathbb{E}\|g(\boldsymbol{X}) - \boldsymbol{\theta}\|^2$.

They developed a new estimator,

$$\hat{\boldsymbol{\tau}}_{\text{S2}} = \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} + \left(\boldsymbol{I}_K - \frac{\text{Tr}\left(\boldsymbol{\Sigma}_r^2\right)\boldsymbol{\Sigma}_r}{\left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{r}}{}^2 \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)}\right)_+ \cdot \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right).$$

Both methods require at least three strata and independence among stratum estimates, to ensure a reduction in overall MSE. Simulations by Rosenman et al. (2023a) revealed that the choice of stratification variables and the number of strata affect performance. However, their findings remain inconclusive: it is unclear what the optimal number of strata is, whether it is better to have more or fewer strata, and no single estimator consistently outperforms the others.

For inference, both papers proposed empirical Bayes (EB) confidence intervals for the kth stratum-specific estimate, $\hat{\tau}_{\text{S1},k}$ and $\hat{\tau}_{\text{S2},k}$, based on the result of the James-Stein estimator variance in Morris (1983). A key difference between EB CIs and conventional frequentist CIs is that EB CIs aim to achieve an average coverage rate for the entire $\hat{\boldsymbol{\tau}}$, which means that they may under-cover the causal effects for some strata and over-cover for others. Simulations in Rosenman et al. (2023a) confirmed that for large sample sizes, their 90% EB CIs met the desired average coverage rate, although the coverage rates varied across individual strata.

Combining the weighted-combination and shrinkage concepts, Rosenman et al. (2023b) constructed a double shrinkage approach, which intuitively speaking, involves first constructing a MSE-minimizing weighted average of $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ and $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$, and then shrinking it towards zero. To derive the estimator, the authors introduced a Bayesian hierarchical model for the true value of $\boldsymbol{\tau}$, the bias $\boldsymbol{\delta}$, and the estimates $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ conditional on $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$. The expression for the kth entry of their double-shrinkage estimator is as follows:

$$\begin{aligned}
&\tau_{DS,k}\left(\phi^2, \omega^2\right) \\
&= \underbrace{\left\{\frac{\omega^2\left(\phi^2 + \sigma_{ok}^2 + \sigma_{rk}^2\right)}{\sigma_{rk}^2\left(\phi^2 + \sigma_{ok}^2\right) + \omega^2\left(\phi^2 + \sigma_{ok}^2 + \sigma_{rk}^2\right)}\right\}}_{a_k} \cdot \left\{\underbrace{\frac{\left(\phi^2 + \sigma_{ok}^2\right)}{\phi^2 + \sigma_{ok}^2 + \sigma_{rk}^2}}_{\lambda_k}\hat{\tau}_{rk} + \underbrace{\frac{\sigma_{rk}^2}{\phi^2 + \sigma_{ok}^2 + \sigma_{rk}^2}}_{1-\lambda_k}\hat{\tau}_{ok}\right\},
\end{aligned}$$

where ϕ and ω are the prior variances of $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$, respectively. The second term resembles the form of a weighted-combination estimator, while the first term, a_k, acts as a shrinkage factor towards zero. Notably, ϕ and ω are crucial in regulating the combination, and since they are unknown, the authors suggested three estimation methods: moment matching, empirical Bayes MLE, and URE minimization. Their simulations indicated that estimating ϕ and ω using moment matching and EB MLE generally outperform URE minimization. Like the findings in Rosenman et al. (2023a), simulations indicated that the performance of the double-shrinkage estimator is influenced by the selection of stratification variables and the number of strata, though the mechanism behind this influence remains unclear.

3.2.4 Bias correction

Another idea is to use the RCT data, which offers unbiased causal estimates, to correct the bias in the RWD estimate. The two methods in Kallus et al. (2018) and Yang et al. (2020) are in the same vein, and the latter is an extension of the former. By contrasting the RCT and RWD, these methods learn the impact of hidden confounding, that is, violation of Assumption 3.1. Both of the methods assume RWD comparability (Assumption 2.3) to hold.

Kallus et al. (2018) aimed at learning the impact of confounding that results in the HTE function estimated from the RWD to be biased. That is,

$$\begin{aligned}\xi(X) = &\{\mathbb{E}[Y(1) \mid X, S=0] - \mathbb{E}[Y(1) \mid X, A=1, S=0]\} \\ &- \{\mathbb{E}[Y(0) \mid X, S=0] - \mathbb{E}[Y(0) \mid X, A=0, S=0]\},\end{aligned}$$

and when ignorability (Assumption 3.1) holds for the RWD, $\xi(X) = 0$. Kallus et al. (2018) proposed to obtain samples of $\xi(x)$ by calculating the difference between the CATE estimated in the RCT and RWD. As an illustration, they assumed a linear structure of $\xi(x)$ and derived the least-squares coefficient estimates. The resulting $\hat{\xi}(x)$ serves as the bias-correction term in this procedure termed *experimental grounding*. They demonstrated that under the linearity assumption, the bias-corrected CATE estimate $\hat{\tau}_{\mathrm{XG}}(x)$ is root-n consistent, although efficiency gains are not theoretically proven.

Yang et al. (2020) summarized the impact of unmeasured confounding on the potential outcomes into the *confounding function*:

$$\varphi(X) = \mathrm{E}[Y(0) \mid A=1, X, S=0] - \mathrm{E}[Y(0) \mid A=0, X, S=0].$$

They demonstrated that, assuming the parametric forms of the confounding function $\varphi(X)$ and the CATE function $\tau(X)$ are known, both $\varphi(X)$ and $\tau(X)$ are non-parametrically identifiable. They also derived semi-parametric efficient, doubly robust estimators under several examples for the assumed parametric form of $\tau(X)$. In their Theorem 3, the authors proved efficiency gains in estimating $\tau(x)$ compared to using the RCT alone, provided the terms in parametric forms of $\tau(x)$ and $\varphi(x)$ are not co-linear. However, if the terms of $\tau(x)$ are a linear combination of those in $\varphi(x)$, there is no efficiency gain from data fusion. This is empirically observed in our simulations presented later in Section 3.3. The authors provided intuition for this result: when the terms are linearly dependent, all the observational data is consumed to estimate the coefficients of $\varphi(X)$ and hence does not contribute to the estimation of $\tau(x)$. While their method is demonstrated using the additive treatment effect as defined in Definition 1.3, adapting it to binary outcomes, or alternative estimands, such as relative risks and odds ratios, may present additional complexities.

While these two methods do not require RWD ignorability (Assumption 3.1), these methods make assumptions about the parametric structure of the functional forms of CATE and confounding, which are subject to misspecification.

van der Laan et al. (2024) proposed a bias-correction method that targets a slightly different estimand: the average treatment effect averaged over the covariate distribution of the covariate-pooled RCT-RWD data:

$$\mathbb{E}_{\boldsymbol{W}}\{\mathbb{E}[Y(1) - Y(0) \,|\, S = 1, \boldsymbol{W}]\},$$

where $p_{\boldsymbol{W}}(\boldsymbol{w}) = p(\boldsymbol{w} \,|\, S = 1)p(S = 1) + p(\boldsymbol{w} \,|\, S = 0)p(S = 0)$. They decomposed this estimand into the difference between a naively pooled ATE estimand and a bias estimand that accounts for the conditional effect of RCT enrollment on the outcome. To estimate them, they introduced an adaptive targeted minimum loss-based estimation (A-TMLE) framework. Unlike the other two bias-correction methods discussed, the A-TMLE estimator does not rely on parametric models for the biases but instead employs flexible methods for estimation such as the highly adaptive lasso estimator(Benkeser and van Der Laan, 2016).

3.2.5 Prognostic adjustment

The PROCOVA (Schuler et al., 2021) method enhances the efficiency of RCT treatment effect estimation by incorporating a prognostic score—a predicted outcome under control conditions derived from baseline risk factors, as formally defined in Hansen (2008)—into a linear regression model. While prognostic scores have long been used in clinical trials (e.g. Cooney et al., 2009), such as the Framingham risk

score (D'Agostino Sr et al., 2008), PROCOVA specifically applies this to data fusion. It leverages external control data to train the prognostic model, which is then applied to the RCT for covariate adjustment. This prognostic model acts as a medium for transferring information from the RWD to the trial, thereby improving statistical efficiency while maintaining rigorous control over Type I error. The method relies on the ANCOVA II estimator, which, importantly, guarantees consistent treatment effect estimates under randomization, even with model misspecification (Yang and Tsiatis, 2001). Furthermore, the original authors proposed a pre-trial sample size calculation method to complement this data fusion approach.

However, PROCOVA's effectiveness depends on the prognostic score's incremental predictive ability for outcome variability, beyond that already captured by the baseline variables. If baseline covariates already account for a substantial portion of outcome variation, the potential efficiency gains from PROCOVA may be limited. For example, if outcomes are linearly related to covariates, including linearity after transformation such as in generalized linear models (GLMs), no additional efficiency benefit is achievable beyond that from the routine adjustment by pre-treatment covariates. This observation resembles the intuition of the findings in Yang et al. (2020) as discussed in Section 3.2.4; this is because all the information from the raw covariates is already exploitable by the linear model. Moreover, its pre-trial sample size calculations assume the external data provide accurate estimates of outcome variance and prognostic correlation—a potentially strong assumption unless the RWD and RCT populations are highly comparable.

The principal advantage of PROCOVA is its protection against Type I error, a critical consideration for regulatory decision-making. This feature contributed to PROCOVA receiving regulatory endorsement from the EMA in 2022 (European Medicines Agency, 2022). A recent extension by Liao et al. (2025) generalizes the framework to semi-parametric estimators, such as the augmented inverse probability weighting estimator (AIPW) and the targeted maximum likelihood estimator (TMLE) (van der Laan and Rubin, 2006).

3.2.6 A summary table for application

Table 3.1 provides a practical summary of the methods discussed above, serving as a reference for researchers selecting an appropriate method. The *data assumption* column indicates whether a method is parametric (P), nonparametric (NP) or semiparametric (SP). Understanding these assumptions helps users recognize their limitations: semi-parametric methods require correct model specification of the parametric part to avoid biased estimates, while non-parametric methods offer

flexibility but may have poor finite sample performance. The next two columns indicate whether a method is applicable to combining only the control arm or both arms from the RWD. The *inference* column describes how each approach quantifies uncertainty. The following column assesses the difficulty of implementation, specifically coding difficulty. *Easy* denotes straightforward coding in a couple of lines, while methods like the power likelihood approach are marked as "medium" due to the need for users to specify a causal model and assemble basic functions despite the available package. The R packages available are: `causl` (Evans, 2021) for the power likelihood approach, `ElasticIntegrative` (Yang and Gao, 2023) for the elastic integrative estimator, `EScvtmle` (Dang et al., 2023b) for the ES-CVTMLE estimator, and `IntegrativeHTEcf` (Yang, 2020) for the confounding function method. See Appendix C.1 for details. The *alternative estimands* column compares the complexity of adapting methods to estimands such as relative risk. Semi-parametric methods typically require significant modification, such as replacing the efficient influence function, and current packages often lack this flexibility. The final column lists primary limitations for users to consider.

Category	Method	Data assumption	Uses RWD†		Inference	Implementation difficulty	Alt. estimands	Limitations / requirements
			control	*both*				
Bayesian dynamic borrowing	Power likelihood (Lin et al., 2025)	P	✓	✓	Bayesian credible interval	Medium	Easy	Requires full parametric model; correct HTE specification
Test-then-pool	Elastic integration (Yang et al., 2023)	NP		✓	Asymptotics / bootstrap	Package ready	Hard	Correct HTE specification required
	ESCV-TMLE (Dang et al., 2022)	SP	✓	✓	Asymptotics	Package ready	Hard	Interpretation may be challenging
Weighted combination	Anchored thresholding (Chen et al., 2021)	NP			Bootstrap	Easy	Easy	No reversion to RCT-only as bias increases
	MSE-minimising (Oberst et al., 2023)	NP		✓	Bootstrap	Easy	Easy	
Bias correction	Experiment grounding (Kallus et al., 2018)	SP		✓	Bootstrap	Easy	Hard	Assumes linearity; efficiency gains for HTE, not ATE
	Confounding function (Yang et al., 2020)	SP		✓	Asymptotics / bootstrap	Package ready	Hard	Correct HTE specification; gains for HTE, not ATE
Covariate adjustment	PROCOVA (Schuler et al., 2021)	P	✓		Sandwich estimator	Easy	Easy	Limited gain if outcome nearly linear
Shrinkage	Green et al. (2005)	NP		✓	Bootstrap	Easy	Easy	Requires multiple independent strata estimates
	Rosenman et al. (2023a)	NP		✓	Bootstrap	Easy	Easy	Requires multiple independent strata estimates

Table 3.1: Comparison of statistical methods for incorporating real-world data (RWD). P = parametric, NP = non-parametric, SP = semiparametric. †A check-mark indicates whether RWD is incorporated only in the control arm or in both arms.

3.3 Simulation studies

In our simulation, we aim to be as realistic as possible by introducing both binary and continuous covariates, explicitly specifying hidden confounding, and accounting for treatment heterogeneity, among other factors. Details of method implementation can be found in Appendix C.1.

3.3.1 Simulation setup

We adopt the *frugal parameterization* proposed in Evans and Didelez (2024) to parameterize and simulate from specified marginal structural models (MSM). The primary advantage is its flexibility: we can simulate exactly the causal effects we would like. We first generate three pre-treatment covariates X_1, X_2 and X_3, where X_3 acts as an effect modifier to introduce treatment effect heterogeneity. We assume that the marginal distribution of X_1 is a standard normal distribution, and the marginal distribution of X_2 follows a Bernoulli distribution with a probability of 0.5. We model the dependency between X_1 and X_2 later using a copula. We assume that the effect modifier X_3 is independent of X_1 and X_2 and follows a Bernoulli distribution with a probability of 0.5.

$$\begin{aligned} X_1 &\sim \mathcal{N}(0,1) \\ X_2 &\sim \text{Bernoulli}(0.5) \\ X_3 &\sim \text{Bernoulli}(0.5) \\ U &\sim \mathcal{N}(0,1.1) \\ Y(a) \mid X_3, U &\sim \mathcal{N}(0.5 + 0.2a + 0.1X_3a + X_3 + U, 1) \end{aligned}$$

Given the mixture of continuous and discrete variables, we model the dependence between X_1, X_2 and Y using a latent Gaussian copula model, as proposed by Fan et al. (2017). Specifically, we employ a trivariate Gaussian copula with correlation coefficients $\rho_{X_1X_2} = 2\,\text{expit}(0.1) - 1 \approx 0.05$, $\rho_{X_1Y} = 2\,\text{expit}(2) - 1 \approx 0.76$ and $\rho_{X_2Y} = 2\,\text{expit}(1) - 1 \approx 0.46$. These coefficients allow us to simulate a weak correlation between X_1 and X_2, a strong influence of the continuous variable X_1 on the outcome and a moderate correlation between the binary variable X_2 and Y.

We simulate treatment assignments in the RCT and in the observational data as below:

$$\begin{aligned} A_{\text{RWD}} &\sim \text{Bernoulli}(\text{expit}\{-0.5 + X_1 + X_2 + \psi U\}) \\ A_{\text{RCT}} &\sim \text{Bernoulli}(0.5). \end{aligned}$$

This parameterization assumes that treatment assignment is completely randomized in the RCT. In this setup, the value of the target estimand, $\tau = \mathbb{E}[Y(1) - Y(0)]$ is 0.25. In contrast, in the observational data, treatment assignment depends on observed covariates X_1, X_2 and X_3 as well as confounding from U. The degree of hidden confounding is controlled by the value of ψ. When $\psi = 0$, U is no longer a confounder, making the target estimand identifiable in the observational data. However, as the absolute value of ψ increases, the bias in the ATE estimated from the RWD correspondingly intensifies. In the simulation, we vary the value of ψ to introduce different magnitudes of hidden confounding into the observational data. Note that the average proportion of treated individuals remains at 50% in all simulations. We set the size of the RCT and observational datasets to 300 and 1200 respectively.

3.3.2 Augmenting both arms of the RWD

In Figure 3.1, each line represents one data fusion estimator discussed in Section 3.2, and the results are averaged over 300 different simulated datasets. Comparing the root mean square error (RMSE) of the fusion estimator $\hat{\tau}$ to that of $\hat{\tau}_r$, the estimator from the RCT alone, is particularly useful to determine the benefits of data fusion. The y-axis of the figure shows the relative RMSE, which is defined similarly to Oberst et al. (2023):

$$r\,\text{RMSE} = \frac{\text{RMSE}(\hat{\tau})}{\text{RMSE}(\hat{\tau}_r)},$$

where as a reference estimator, $\hat{\tau}_r$ is the AIPW estimator using the RCT data only. The fusion estimator $\hat{\tau}$ outperforms τ_r if rRMSE is less than 1 and worse, if rRMSE exceeds 1, which means that the efficiency gain is not enough to offset the inflation in bias.

As a consequence of violation of Assumption 2.3 and 3.1, the RWD bias, defined as $\delta := \mathbb{E}[\hat{\tau}_o] - \tau$, is plotted on the x-axis. Instead of the absolute value of δ, we plot bias on a relative scale. Specifically, we define the relative bias as the multiple of the standard deviation of the RCT-only estimator $\hat{\tau}_r$:

$$r\,\text{Bias} = \frac{\delta}{\text{sd}(\hat{\tau}_r)}.$$

For instance, a relative bias of one indicates that $\mathbb{E}[\hat{\tau}_o]$ is one standard deviation away from τ. The advantage of using relative bias is that it provides an intuitive metric of confounding bias in relation to the variability of $\hat{\tau}_r$.

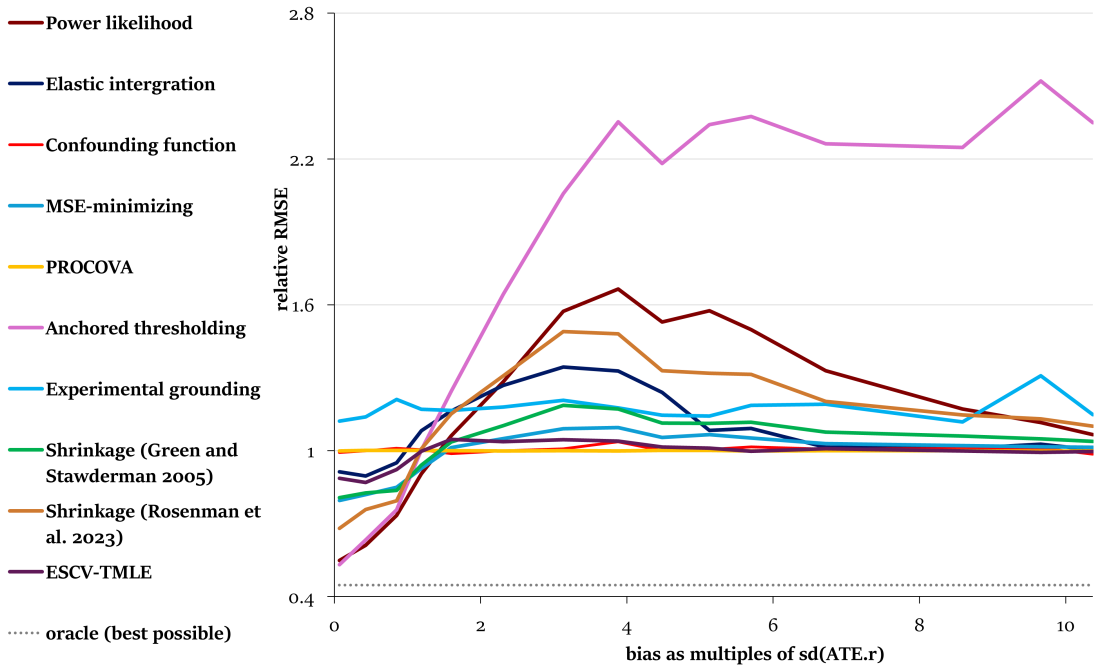


Figure 3.1: RMSE relative to RCT-only approach (without data fusion) by varied magnitude of bias induced by hidden confounding, in the baseline scenario with a continuous outcome

We also include an oracle performance curve for reference, which illustrates the best possible reduction in RMSE that could be achieved if the bias of the estimator $\hat{\tau}_o$ were known when fusing the two datasets. An example is when r Bias $= 0$, indicating that the observational data should be integrated entirely and treated equally as the RCT data. In this scenario, the reduction in RMSE is due to the increase in total sample size, which is approximately equal to $\sqrt{300/(1200+300)} \approx 0.44$ in our simulation setting. We include this curve to demonstrate how close each method can come to the optimal performance. However, note that the oracle curve does not apply to the PROCOVA estimate as it only incorporates the control arm.

Figure 3.2 compares the coverage rates of 95% confidence intervals (credible interval for the power-likelihood estimator), and the ratio of the length of the 95% CI of each data fusion estimator against that of $\hat{\tau}_r$. This comparison illustrates a trade-off: estimators that achieve greater power gains when bias is small are also at risk of poorer worst-case coverage rates. This highlights the trade-off between power gains and coverage rate.

3.3.2.1 Anchored thresholding

As δ increases, almost all the curves converge to the reference line, except for the anchored thresholding estimator in Chen et al. (2021). The reason is that as the

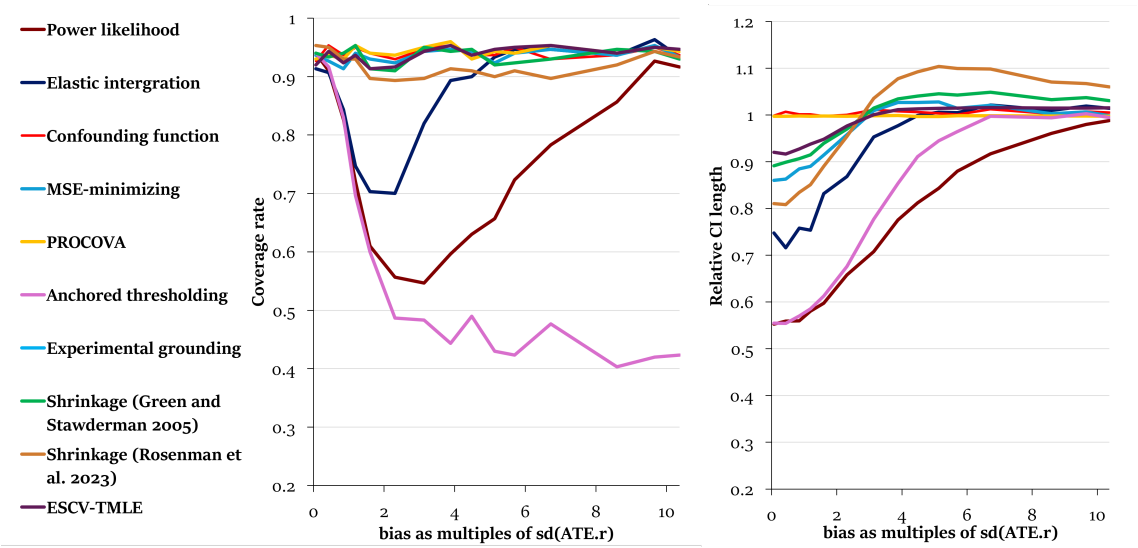


Figure 3.2: Coverage rate and relative CI length of data fusion estimators, by varied magnitude of bias induced by hidden confounding.

RWD bias, δ, approaches infinity, their proposed estimator becomes asymptotically biased. We illustrate this using the following example.

Example 3.1. Let $\hat{\tau}_r \sim N(\tau, \sqrt{\sigma^2/n_e})$ and $\hat{\tau}_o \sim N(\tau + \delta, \sqrt{\sigma^2/n_o})$. We assume that the total sample size is $n = n_r + n_o$ and $n_o = \rho n$, where $\rho \in (0, 1)$.

Following the methodology outlined in Section 3.2 of Chen et al. (2021), the weighting factor becomes $\omega = \frac{\sigma^2/n_r}{\sigma^2/n_r + \sigma^2/n_o} = \rho$, and the estimator of bias $\hat{\delta} = \hat{\tau}_o - \hat{\tau}_r \sim N\left(\delta, \sqrt{\frac{\sigma^2}{n_r} + \frac{\sigma^2}{n_o}}\right)$. We assume δ is a very large positive number, specifically, $\delta \gg \sqrt{\frac{\sigma^2}{n_r} + \frac{\sigma^2}{n_o}}$. After the proposed soft-thresholding procedure, the adjustment for bias they proposed is $\hat{\delta}_\gamma = \hat{\delta} - \gamma\sqrt{\frac{\sigma^2}{n_r} + \frac{\sigma^2}{n_o}} \sim N\left(\delta - \gamma\sqrt{\frac{\sigma^2}{n_r} + \frac{\sigma^2}{n_o}}, \sqrt{\frac{\sigma^2}{n_r} + \frac{\sigma^2}{n_o}}\right)$. Recall that γ is an arbitrary positive hyper-parameter that controls the adjustment. Therefore, their final estimator becomes:

$$
\begin{aligned}
\hat{\tau}_{\text{AT}} &= \omega\hat{\tau}_r + \rho\left(\hat{\tau}_o - \hat{\delta}_\gamma\right) \\
&= (1-\rho)\hat{\tau}_r + \rho\left(\hat{\tau}_r + \hat{\delta} - \hat{\delta} + \gamma\sqrt{\frac{\sigma^2}{n_r} + \frac{\sigma^2}{n_o}}\right) \\
&= \hat{\tau}_r + \rho\gamma\sqrt{\frac{\sigma^2}{n_r} + \frac{\sigma^2}{n_o}} \\
&\sim N\left(\tau + \rho\gamma\sqrt{\frac{\sigma^2}{n_r} + \frac{\sigma^2}{n_o}}, \sigma/\sqrt{n_r}\right).
\end{aligned}
$$

It is obvious that when δ is very large, their anchored-thresholding estimator $\hat{\tau}_{AT}$ is biased by $\rho\gamma\sqrt{\frac{\sigma^2}{n_r} + \frac{\sigma^2}{n_o}}$, with a variance equal to using the RCT data alone. Addi-

tionally, a larger γ leads to greater bias, which explains the empirical observations made in Oberst et al. (2023). This extra bias is the reason why the RMSE curve remains constantly above the reference line, and the coverage rate remains low as δ increases. This means that the anchored thresholding estimator does not offer the assurance that when the RWD is highly biased, we will not be worse off than using the RCT alone. For this reason, despite showing the highest power gains in Figure 3.2, we do not recommend the anchored thresholding method for data fusion.

3.3.2.2 Bias-correction methods

Figure 3.3 focuses on the two bias-corrections methods. It shows that neither the confounding function (Yang et al., 2020) nor the experimental grounding (Kallus et al., 2018) reduces the RMSE, regardless of the level of confounding bias. In fact, using the experimental grounding estimator results in a loss of efficiency, leading to wider confidence intervals.

This is because both methods essentially distribute the difference between $\hat{\tau}_r$ and $\hat{\tau}_o$ across a pre-specified function space defined by the covariates. When marginalized over the RCT samples, the estimate reverts to $\hat{\tau}_r$, yielding no efficiency gains in estimating the ATE. van der Laan et al. (2024) further formalized this point by comparing the non-parametric efficiency bounds, showing that if the target parameter measures the HTE in the RCT averaged over the covariate distribution of the RCT, there may still be no efficiency gain, even if the estimand is efficiently estimated. Additionally, for the experiment grounding estimator, the loss is due to the added noise, which makes the estimation less efficient.

It is worth noting that the confounding function method is designed for efficiency gains in estimating the HTE function, with theoretical proof provided in Theorem 3 in Yang et al. (2020). However, this method is not suitable for estimating the ATE in our context. If we make an extra assumption that the observational sample is representative of the target population, meaning both the RCT and RWD comes from the same population, we can marginalize the estimated HTE function over the pooled RCT + RWD data to achieve power gains by effectively increasing the sample size. Even then, the efficiency gain relies on correct specification of the function spaces.

3.3.2.3 Methods showing a trade-off

Figure 3.4 focuses on estimators that exhibit a trade-off pattern, showing a reduction in RMSE in some scenarios and an increase in others. Specifically, the leftmost panel of Figure 3.4 demonstrates that when the RWD bias is small, typically less than twice

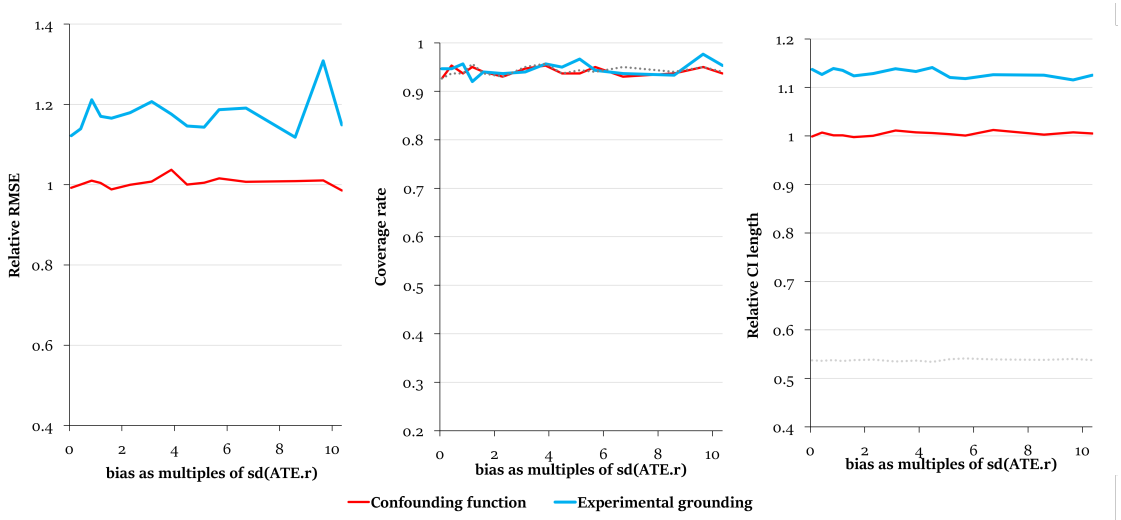


Figure 3.3: Relative RMSE, coverage rate and relative CI length of bias-correction estimators (Kallus et al., 2018; Yang et al., 2020).

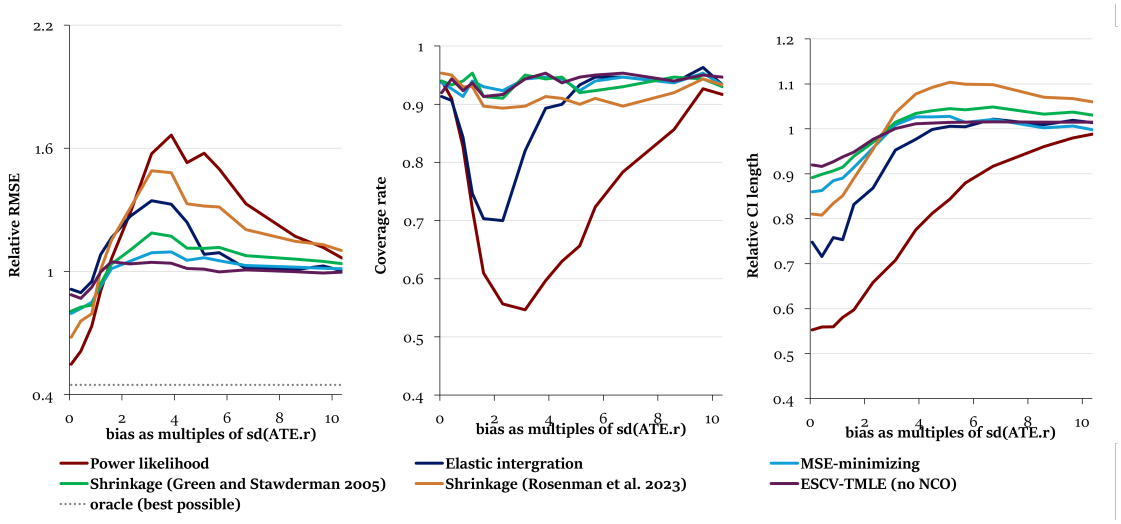


Figure 3.4: Relative RMSE, coverage rate and relative CI length of estimators that exhibit a trade-off (Green et al., 2005; Rosenman et al., 2023a; Oberst et al., 2023; Lin et al., 2025; Dang et al., 2022; Yang et al., 2023).

the standard deviation of $\hat{\tau}_r$, the reduction in variance outweighs the introduced bias leading to a reduction in RMSE. However, as δ exceeds two standard deviations, the RMSE increases compared to not incorporating RWD. This is primarily due to the bias overwhelming the benefit of reduced variance. As the RWD bias continues to increase, most methods gradually converge to the reference line, indicating that these methods recognize the extremely large RWD bias and respond by down-weighting the influence of the RWD to zero, relying solely on the RCT for estimation.

The middle and the rightmost panels of Figure 3.4 provide further details about the trade-off. Since the trade-off in RMSE involves a trade-off between squared bias and variance, it manifests in the coverage rate of the 95% confidence interval and the relative length of these intervals. Additionally, there is a general trade-off between best-case reward and worst-case risk. Methods with larger potential efficiency gains are also more susceptible to higher worse-case RMSE. This is evident in Figure 3.4, where the order of lines from high to low before crossing the reference line is exactly reversed afterwards. This implies that aiming for a larger efficiency gain involves taking on greater risks.

Beyond the presented simulations, we have conducted additional simulations with varying parameter values. The relative behaviour remains somewhat consistent across these simulations. The power-likelihood (Lin et al., 2025), elastic integration (Yang et al., 2023), and the shrinkage estimator by Rosenman et al. (2023a) are among the *ambitious* methods with substantial potential efficiency gains but lower coverage rate in certain ranges of RWD bias. In contrast, the MSE-minimizing and ES-CVTMLE (no NCO) estimators tend to be *conservative*; they offer a smaller potential reduction in uncertainty but maintain a coverage rate closer to the nominal rate.

In this simulation setting, when the RWD is unconfounded, the power-likelihood estimator achieves considerable efficiency gains, nearly reducing the RMSE to the best possible level and shortening the CI length by almost 45%. However, this comes with the risk of CI under-coverage, with the worst-case coverage rate dropping below 60%. As discussed in Section 2.7 in the previous chapter, unlike frequentist methods that directly target the inference of ATE, the power-likelihood approach selects the learning factor η based on optimizing the expected predictive density over the target population. This means η is chosen so that the η-indexed joint posterior of the model parameters best predicts the *true* data. Since this involves a joint distribution, η is influenced not only by the ATE, which is primarily a function of the mean outcomes, but also the other features of the outcome distribution. For example, if the conditional variance of Y differs between the RCT and the RWD—which is likely given

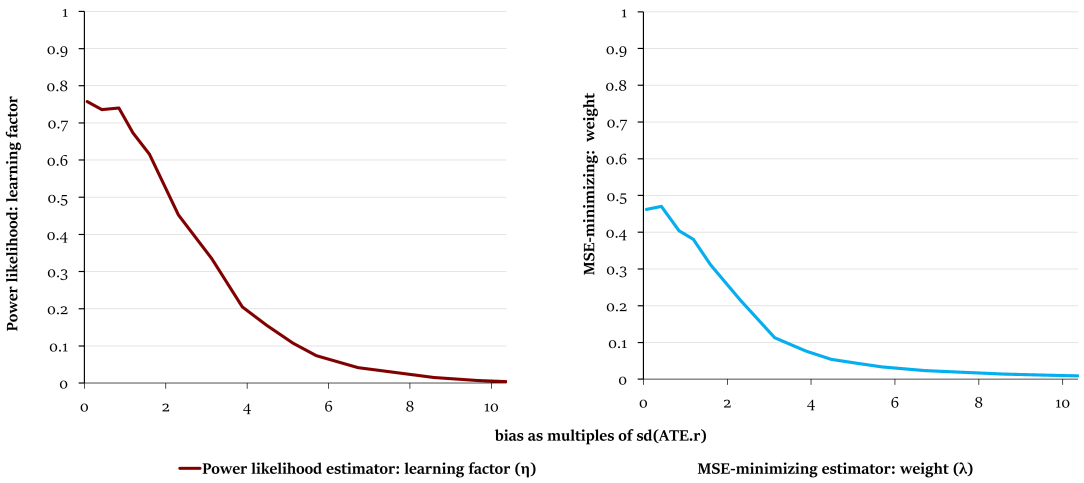


Figure 3.5: The average learning factor η selected by power-likelihood method (Lin et al., 2025) and the average weight λ assigned to $\hat{\tau}_o$ selected by the MSE-minimizing method (Oberst et al., 2023). Note that due to differences in methodologies, the absolute levels of these two parameters are not directly comparable.

the highly controlled environment of the RCT versus the numerous unmeasured uncertainties in RWD—the power-likelihood method accounts for this difference. In such scenarios, stronger unmeasured confounding potentially increases the noise in Y. Consequently, the power-likelihood approach typically selects a smaller η, rendering a more cautious inclusion of information from the RWD. The simulations in Lin et al. (2025) illustrated a scenario where the conditional volatility of Y intensifies with the strength of unmeasured confounding. Their results showed that in such cases, the power-likelihood method achieves efficiency gains while maintaining a coverage rate close to the nominal.

Key to the MSE-minimizing estimator is the weight λ that balances the RCT and RWD estimates. As discussed in Section 3.2.2, plug-in estimator $\hat{\lambda}$ underestimates the MSE-minimizing λ due to Jensen's inequality. Consequently, less RWD is included than is theoretical optimal. This likely contributes to the estimator being more conservative compared to other methods.

The learning factor η in the power-likelihood estimator and the weight λ in the MSE-minimizing estimators play crucial roles in moderating the influence from the RWD. Both methods proposed ways to robustly select these parameters. Indeed, Figure 3.5 demonstrates that both methods effectively downweight the RWD in response to increased bias.

Regarding the shrinkage estimators in Green et al. (2005) and Rosenman et al. (2023a), as discussed in Section 3.2.3, these methods are designed to improve the efficiency of the vector of strata estimates rather than directly enhancing the ATE.

To get an estimate of the target estimand τ, we introduce an extra step to multiply by a weight vector:

$$\hat{\tau}_{\text{ATE},S1} = \boldsymbol{w}^T \hat{\boldsymbol{\tau}}_{S1} \qquad\qquad \hat{\tau}_{\text{ATE},S2} = \boldsymbol{w}^T \hat{\boldsymbol{\tau}}_{S2},$$

where $\boldsymbol{w}^T$ is a $1 \times K$ vector with weights for each stratum in the target population. We use $\hat{w}_k = n_{r,k}/n_r, k = 1, \dots K$ as an estimate.

Theoretically, these shrinkage estimators should reduce the overall RMSE, where *overall* refers to the sum, or a weighted sum, of strata estimator RMSEs. However, this theoretical guarantee of RMSE reduction does not necessarily extend to $\hat{\tau}_{\text{ATE},S1}$ and $\hat{\tau}_{\text{ATE},S2}$. This explains why, in certain ranges of δ, neither the RMSE of the ATE is not consistently reduced, nor the length of CIs. Furthermore, as discussed in Section 3.2.3, Green et al. (2005) and Rosenman et al. (2023a) both suggested using the EB confidence interval for each entry of the vectors $\hat{\boldsymbol{\tau}}_{S1}$ and $\hat{\boldsymbol{\tau}}_{S2}$. However, since our focus is on $\hat{\tau}_{\text{ATE},S1}$ and $\hat{\tau}_{\text{ATE},S2}$, which are linear combinations of the strata shrinkage estimates, the point-wise CIs are not very useful. It is more appropriate to use the nonparametric bootstrap approach, which involves resampling each stratum in the RCT and RWD, respectively, and calculating the confidence intervals based on, in this implementation, 200 bootstrap samples.

3.3.2.4 ES-CVTMLE (Dang et al., 2022)

The ES-CVTMLE method offers the option to incorporate an NCO to enhance the estimation of bias. As discussed in Section 3.2.1, the NCO-integrated estimator relies on the U-compatibility condition. To investigate the sensitivity to this assumption, we modify the simulation parameterization by splitting the unmeasured confounder U into two independent variables, $U_1 \sim \mathcal{N}(0,1)$ and $U_2 \sim \mathcal{N}(0, 0.1)$, such that $U_1 + U_2$ is equivalent to U. Specifically,

$$\begin{aligned} Y(a) \mid X_3, U_1, U_2 &\sim \mathcal{N}(0.5 + 0.2a + 0.1X_3a + X_3 + U_1 + U_2,\ 1) \\ A_{\text{RWD}} &\sim \text{Bernoulli}(\text{expit}\{-0.5 + X_1 + X_2 + X_3 + \psi U_1 + \psi U_2\}). \end{aligned}$$

We further include three NCO variables:

$$\begin{aligned} N_1 &\sim \mathcal{N}(1 + X_2 + X_3 + U_1, 2) \\ N_2 &\sim \mathcal{N}(1 + X_2 + X_3 + U_2, 2) \\ N_3 &\sim \mathcal{N}(0, 2). \end{aligned}$$

In broad terms, N_1 captures 10/11 of the total variation of unmeasured confounding, while N_2 captures the remaining 1/11. We then refer to N_1 as a strong NCO and N_2

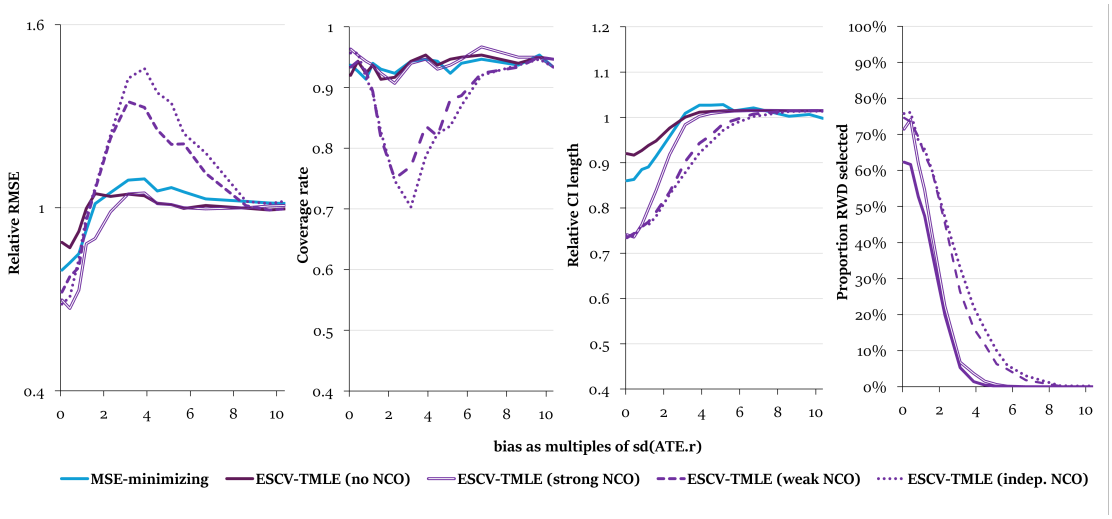


Figure 3.6: Relative RMSE, coverage rate, relative CI length, and average selected proportion of RWD of the ESCV-TMLE estimator (Dang et al., 2022) across scenarios without NCO, with a weak NCO, with a strong NCO and with an irrelevant NCO. The MSE-minimizing estimator (Oberst et al., 2023) is included for reference.

as a weak NCO. N_3 is simulated as a noise variable representing an extreme scenario where an NCO is entirely irrelevant.

Figure 3.6 compares the performance is different versions of the ES-CVTMLE estimator: without an NCO (solid line), with a weak NCO (dashed line), with a strong NCO (double line), and with an irrelevant NCO (dotted line). The MSE-minimizing estimator (blue solid line) is included for reference. Interestingly, the inclusion of NCOs with varying degrees of relevance affects the estimator's behaviour differently. With a strong NCO, which captures most of the unobserved confounding, the performance is optimal. It shows substantial efficiency gains when RWD is unbiased and maintains a coverage rate consistently around 95%. In this scenario, the ES-CVTMLE estimator dominates the MSE-minimizing estimator, as U-comparability is nearly satisfied the NCO strongly effectively enhances the bias estimation in the trade-off. However, if the NCO is weak or totally irrelevant, the estimator become overly ambitious, resulting in higher worst-case RMSE and CIs under-covering. In these cases, U-compatibility is violated, and the algorithm places undue confidence in an unreliable NCO.

The rightmost panel of Figure 3.6 shows the average proportion of the RWD included, which helps explain the different behaviours across scenarios. Compared to not using an NCO, the versions with NCOs tend to select more RWD when the RWD bias is near zero, which is beneficial. However, the curves return back to zero more slowly than when not using an NCO. For the versions with a weak or irrelevant NCO, the proportions selected are consistently higher than when not using an NCO.

This occurs because the squared-bias term in the selector function is underestimated under the illusion of minimal confounding bias due to the misleading NCO.

These findings suggest that when a candidate NCO is available, it is crucial to assess its relevance. A weak or irrelevant NCO can be misleading and may not function as intended. Therefore, using an NCO is not always better than not using one.

3.3.3 Augmenting only the control arm of the RWD

3.3.3.1 Prognostic adjustment (Schuler et al., 2021; Liao et al., 2025)

As described in Section 3.2.5, the PROCOVA method only uses the control arm of the RWD to fit a prognostic score model, $m(x) = \mathbb{E}[Y \mid A = 0, x, S = 0]$, which is then added as a feature in the RCT dataset. Consequently, the interpretation of the x-axis in Figure 3.1 and 3.2 is more nuanced. Since the PROCOVA method does not use $\hat{\tau}_o$, the impact of unobserved confounding is more on the expectation of the outcomes of the external controls . If there is no unobserved confounding, we expect $\mathbb{E}[Y \mid A = 0, X, S = 0] = \mathbb{E}[Y \mid A = 0, X, S = 1]$. The stronger the confounding, the larger the difference between these expectations.

In Figure 3.1 the PROCOVA line overlaps with the reference line, with a coverage rate constantly close to 95% and no power gain, as shown in Figure 3.2. In this case, there is no benefit from data fusion. This is because the simulation setup, as outlined in Section 3.3.1, assumes a linear outcome. As discussed in Section 3.2.5, the additional prognostic feature does not add any extra information if the outcome is linear, as noted by Schuler et al. (2021). Additionally, the relative RMSE curve of the PROCOVA estimator is flat, indicating that the estimator remains unaffected by the difference between $\mathbb{E}[Y \mid A = 0, X, S = 0]$ and $\mathbb{E}[Y \mid A = 0, X, S = 1]$.

To understand how the PROCOVA estimator performs with non-linearity in the outcome, echoing the simulations in Schuler et al. (2021), we introduce another standard normal variable X_4 and add a quadratic term X_4^2 to the outcome:

$$Y(a) \mid X_3, X_4, U_1, U_2 \sim \mathcal{N}\big(0.5 + 0.2a + 0.1X_3a + X_3 + \alpha X_4^2 + U_1 + U_2,\ 1\big)\,.$$

The coefficient α controls the degree of curvature in the outcomes. When $\alpha = 0$, the outcome is linear, as in the main simulation. Figure 3.7 shows that as α increases, the efficiency gain increases without compromising the coverage rate. Note that the rightmost panel of Figure 3.7 plots power on a relative scale, as the ratio against the power of $\hat{\tau}_r$. This is because, as α varies, the variance of Y changes, making it sensible to compare on a relative rather than an absolute scale. In this specific

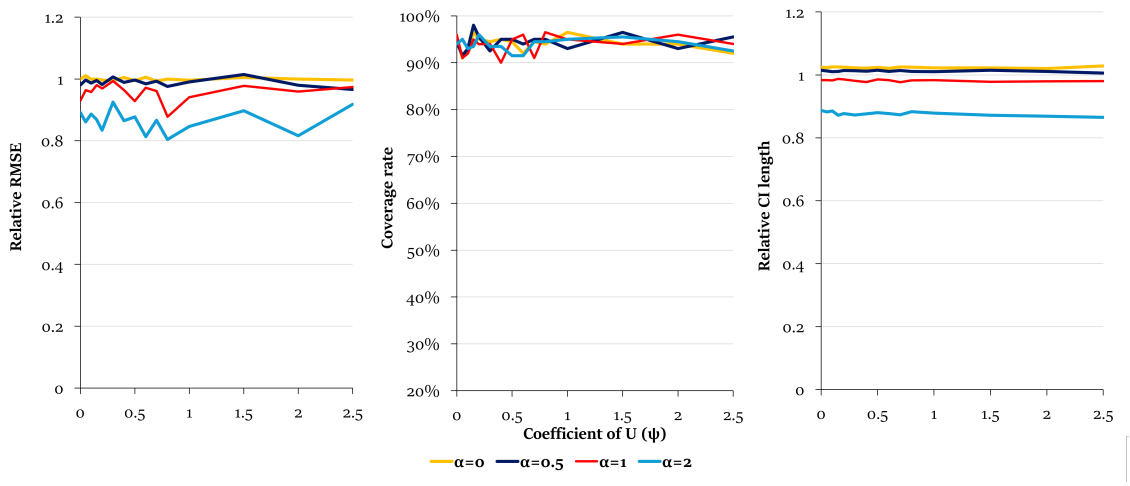


Figure 3.7: Relative RMSE, coverage rate and relative CI length of PROCOVA estimator (Schuler et al., 2021) for outcomes with various degrees of curvature $\alpha = \{0, 0.5, 1, 2\}$.

simulation, when $\alpha = 2$, on average, the PROCOVA estimator shortens the CIs by approximately 12%.

However, the above simulation assumes that the prognostic score is perfect, i.e. $m(X) = \mathbb{E}[Y(0)|X, S = 1]$ using the notations in Schuler et al. (2021). The authors relaxed this assumption in their Theorem 2, replacing it with the condition $|m(X) - \mu_0(X)| \xrightarrow{L_2} 0$, where $\mu_0(X) \equiv \mathbb{E}[Y(0)|X, S = 1]$. To test the sensitivity to the perfect prognostic score assumption, we distinguished the curvature coefficient α for the RCT and RWD models by α_r and α_o. Note that in this setting, the condition $|m(X) - \mu_0(X)| \xrightarrow{L_2} 0$ still holds. Table 3.2 gives the relative RMSE in different (α_o,α_r) scenarios. We observe that the most efficiency gain is generally achieved when $\alpha_r = \alpha_o$, and the benefit diminishes when there is a discrepancy. Interestingly, when the RCT outcome is quadratic w.r.t. X_4 but the RWD is linear without X_4, the RMSE of the PROCOVA estimator is even higher than $\hat{\tau}_r$. In this case, the prognostic score introduces more noise than signal in Y's residual variation. Essentially, in a most simplified multivariate regression scenario, the variance of the regression coefficient of the treatment A can be expressed as:

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_A \,|\, \mathbf{X}) &= \frac{\sigma^2_{Y|X_1 \ldots X_d}}{n \times \sigma^2_{A|X_1 \ldots X_d}} \\ &= \frac{\sigma^2_{Y|X_1 \ldots X_d}}{n_r \times (1 - R^2_{A|X_1 \ldots X_d})\sigma^2_A},\end{aligned}$$

where $\sigma^2_{Y|X_1 \ldots X_d}$ is the residual variance of Y after regressing on $X_1, \ldots, X_d$ and $R^2_{A|X_1 \ldots X_d}$ denotes the R^2 of regressing A on all covariates. Although in an RCT, the

treatment is randomized and $R^2_{A|X_1...X_d}$ is expected to be zero in theory, this may not hold true in finite samples, especially in small RCTs. Adding an extra variable to the regression, regardless of its relevance, $\sigma^2_{Y|X_1...X_d}$ will either decrease or stay the same, and $R^2_{A|X_1...X_d}$ will either increase or remain the same. If the additional variable fails to reduce the residual variance of Y, but instead increases $R^2_{A|X_1...X_d}$, then variance of $\hat{\beta}_A$ will be higher as a result. This explains why, when $\alpha_r > 0$ but $\alpha_o = 0$, the relative RMSE exceeds 100%.

Relative RMSE	$\alpha_r = 0$	$\alpha_r = 0.5$	$\alpha_r = 1$	$\alpha_r = 2$
$\alpha_o = 0$	100%	112%	130%	177%
$\alpha_o = 0.5$	100%	98%	96%	89%
$\alpha_o = 1$	100%	100%	96%	88%
$\alpha_o = 2$	100%	100%	92%	84%

Table 3.2: Relative RMSE of PROCOVA estimator when the prognostic score is imperfect.

The PROCOVA method proposes a way to calculate the expected reduction in required RCT sample size prior to conducting the trial. The calculation is based on the expected correlation between the prognostic score and the residuals of Y in the RCT. The estimation of the correlation relies entirely on the RWD and is sensitive to the assumption that the prognostic score is perfect. For example, when $\alpha_o = 2$ the estimated correlation is approximately 0.22, and one would conclude a roughly 4% reduction in the required RCT sample size. However, if $\alpha_r = 0.5$, the actual correlation between the prognostic score and the RCT outcome residuals is close to 0, indicating that the utility of adding a prognostic score is overestimated. If we actually have reduced the RCT size by 4% as suggested by the the procedure, we may end with inadequate power. This example highlights the limitation of the proposed sample size calculation procedure.

In the above discussion, we explored the sensitivity of the PROCOVA estimator to the degree of outcome non-linearity and the assumption of perfect prognostic score. It maintains the nominal coverage rate regardless of the magnitude of confounding, safeguarding against Type I error, which is crucial for informing high-stakes decisions. The ability to calculate the expected sample size reduction also makes it a practical choice among data fusion methods.

3.3.3.2 Power-likelihood (Lin et al., 2025)

Figure 3.8 presents the performance of the power-likelihood method. The pattern is consistent with its application in integrating both arms: when confounding bias is

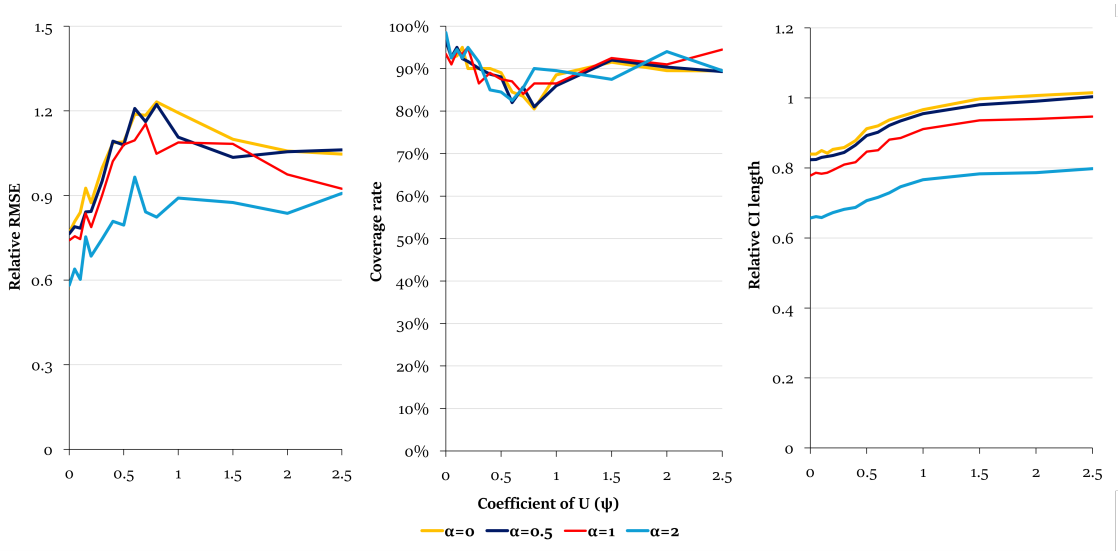


Figure 3.8: Relative RMSE, coverage rate and relative CI length of the power-likelihood estimator (Lin et al., 2025) for outcomes with various degrees of curvature $\alpha = \{0, 0.5, 1, 2\}$.

small, the estimator effectively reduces variance while maintaining a coverage rate close to the nominal. However, as RWD bias intensifies, the estimator's RMSE exceeds that of $\hat{\tau}_r$ due to the excessive bias introduced by including RWD, resulting in CI undercoverage. As confounding bias continues to increase, making the RWD increasingly dissimilar to the RCT, the selected $\hat{\eta}$ to optimize ELPD is reduced to zero. This avoids extreme bias in the estimate, causing the RMSE to return to the reference line.

Interestingly, the reduction in variance compared to $\hat{\tau}_r$ is more pronounced as α grows. This might be due to the fact that with stronger non-linearity or more complex outcome models, the correctly-specified parametric outcome model outperforms nonparametric machine learning methods within the AIPW estimator, especially with small sample sizes. However, potential model misspecification in practice must be carefully considered.

3.3.3.3 ES-CVTMLE (Dang et al., 2022)

Figure 3.9 compares different versions of the ES-CVTMLE estimator. The findings mostly align with those in Section 3.3.2.4. Including a weak or irrelevant NCO can expose the estimation to higher risk of bias.

Notably, compared to PROCOVA and power-likelihood estimators, the benefits of the ES-CVTMLE (no NCO) estimator are limited, especially as the outcome curvature intensifies. In the scenario where $\alpha = 2$, there is no reduction in the RMSE, and in fact, the ES-CVTMLE estimator performs worse than the RCT-only

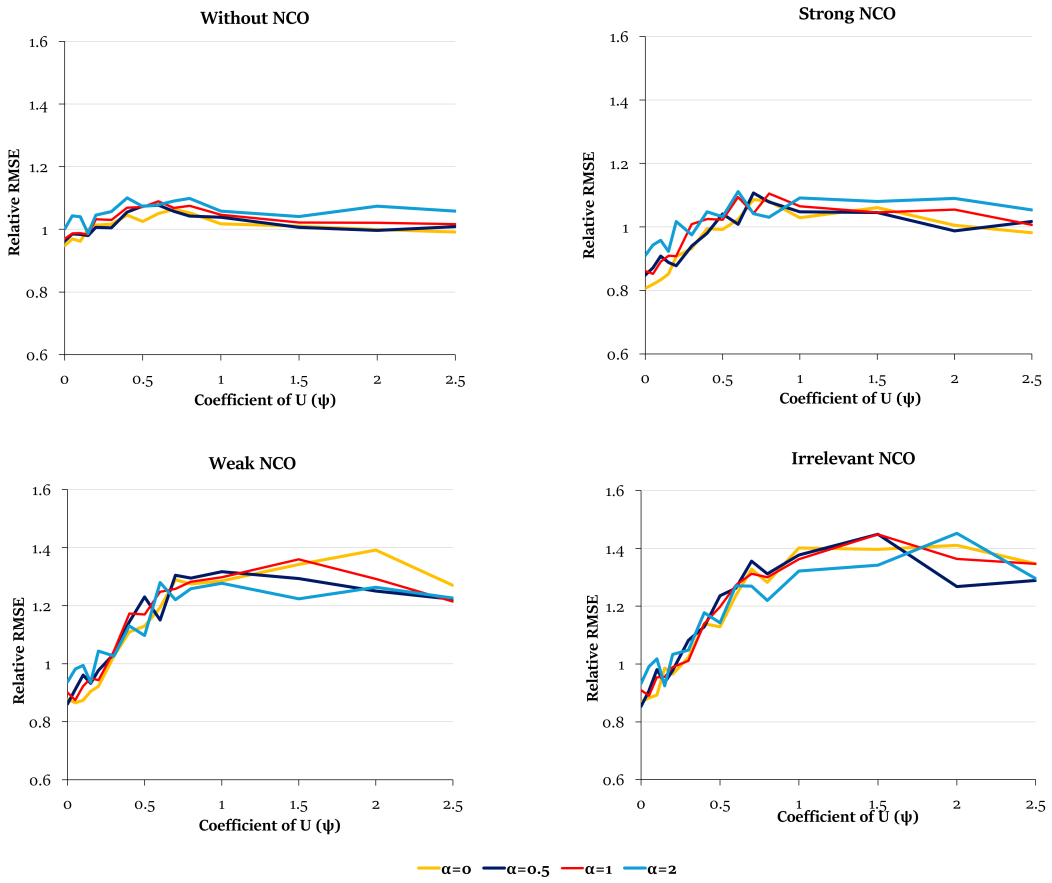


Figure 3.9: Relative RMSE, coverage rate and relative CI length of different versions of the ES-CVTMLE estimator (Dang et al., 2022): without NCO, with a strong NCO, with a weak NCO and with an irrelevant NCO for outcomes with various degrees of curvature $\alpha = \{0, 0.5, 1, 2\}$.

estimator, exhibiting a higher RMSE. van der Laan et al. (2024) reached a similar conclusion in their simulations. In their scenario (c) and (d), where complex non-linear terms were introduced into the generative model, the ES-CVTMLE failed to achieve any efficiency gain. They suspected that this might be due to the ES-CVTMLE not accurately estimating the bias structure or its cross-validation method not fully leveraging the data.

3.4 Real-world application: PIONEER 6 Trial

Patients with type 2 diabetes are at significantly higher risk of developing cardiovascular (CV) diseases. To help manage glycemic levels, semaglutide—a glucagon-like peptide-1 (GLP-1) receptor agonist—has been approved for clinical use. Its cardiovascular safety was assessed in the PIONEER 6 trial (NCT02692716), a random-

ized, double-blind study in patients with type 2 diabetes at high CV risk. The trial enrolled 3,183 participants across 21 countries between 2016 and 2018, with the primary outcome being the time to the first major adverse CV event (MACE). While the study established cardiovascular safety (HR 0.79, 95% CI [0.57–1.11]) (Husain et al., 2019), it was underpowered for efficacy assessment.

While randomized trials remain the gold standard, their long duration precluding access to modern treatments motivates the exploration of alternative evidence synthesis methods. We examine whether integrating existing real-world data (RWD) with PIONEER 6 trial findings—using the fusion methods reviewed earlier—can provide high quality evidence circumventing the need for a new long duration superiority trial.

3.4.1 Target trial emulation

The calibration of observational RWD against a randomized trial is known as target trial emulation, with the goal of estimating the causal quantity of interest (Hernán and Robins, 2016). We emulate PIONEER 6 using real-world claims data from Optum's Clinformatics® Data Mart (CDM), following the framework set out in Hernán and Robins (2016) and Hernán et al. (2022) to specify the key components. We consider the one-year incidence rate of MACE as the outcome[1]. Eligibility criteria were aligned with the trial protocol, and we carefully defined *time zero* for follow-up to avoid issues such as immortal time bias. Since untreated patients may differ systematically from those receiving semaglutide, we adopt an active comparator design, using initiation of a DPP4 inhibitor as a placebo proxy due to its similar glycemic control indications but neutral MACE risk (Scirica et al., 2013; Green et al., 2015).

We consider two data fusion designs: one incorporating only the control arm of the RWD and another incorporating both arms. In Design 1, we align the RWD sampling window with the recruitment period of PIONEER 6 (2016–2018), including only individuals who met the trial's eligibility criteria and initiated DPP-4i during that timeframe. This yielded 16,111 RWD records.

Design 2, which incorporates both treatment and control arms from the RWD, presented challenges due to the late availability of semaglutide in the US. Injectable semaglutide received FDA approval in December 2017 (U.S. Food and Drug Administration, 2020), but substantial uptake in the CDM dataset only began in 2019.

[1]While MACE is traditionally defined as a composite of CV death, nonfatal myocardial infarction (MI), and nonfatal stroke, we extend the definition by replacing CV death with all-cause mortality. This modification is necessary because the real-world data source we use do not reliably distinguish CV-related deaths from other causes. Throughout this section, "MACE" refers to this modified definition.

Aligning the RWD recruitment window with PIONEER 6 would have excluded the treatment group entirely, necessitating a broader data window. However, extending data collection into 2020 introduced additional complications, as COVID-19 was associated with a significantly elevated one-year risk of major adverse cardiovascular events (MACE; HR 1.55, 95% CI: [1.50, 1.60]) (Xie et al., 2022). To balance comparability and sample size, we selected a two-year window before and after PIONEER 6 recruitment, resulting in 89,655 eligible RWD records—of which only approximately 2% were on semaglutide.

To mimic the random assignment of treatment strategies at baseline, it is necessary to control for confounding factors to achieve comparability (or exchangeability) between the treatment and control groups. The Optum CDM dataset offers a wide range of covariates, such as demographics (age, sex, race), biomarkers (e.g. LDL, HDL), medication use (e.g. insulin), and medical history (e.g. heart failure). We utilized feature selection methods, including regularization techniques (LASSO, Ridge regression, and elastic net) and Gradient Boosted Machines (GBM) to identify covariates that predict both treatment assignment and the outcome. As one-year MACE outcomes are missing for about 15.9% of the records, identifying variables predictive of outcome missingness and adjusting for them in the analysis is also necessary.

We further refined the RWD sample by excluding records with incomplete key predictors and trimming those with covariate values outside the 5th and 95th percentiles of PIONEER 6's distribution. After these steps, Design 1 retained 3,652 records, while Design 2 retained 19,206 records, including only 463 semaglutide-treated patients. These procedures closely follow the methodology of Dang et al. (2023a). Full details of the target trial emulation procedure are presented in Appendix C.2.

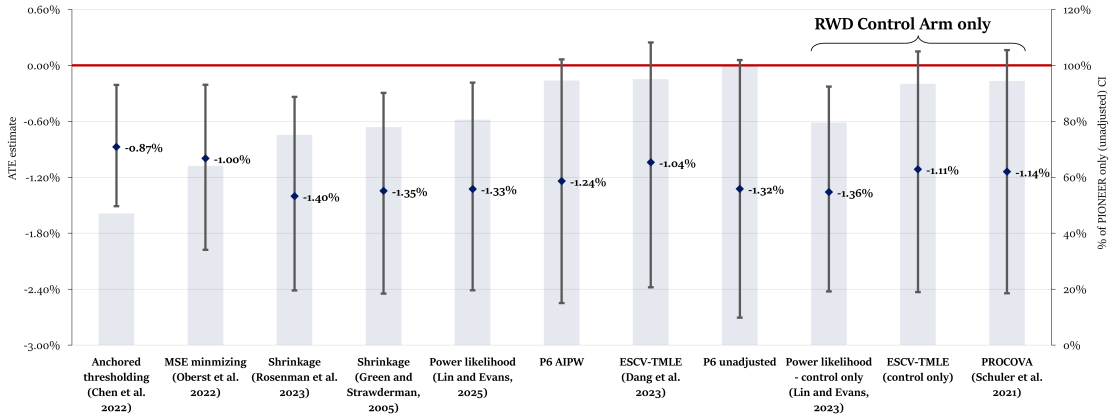
Notably, the procedures are technically outcome-blind—that is, they do not rely on the outcomes from PIONEER 6 and can therefore be implemented prior to their observation. As shown in Table C.7 and C.9 (Appendix C), the resulting RWD cohorts are well-aligned with PIONEER 6, particularly in terms of MACE rates across treatment arms, despite the fact that outcomes were not used during sample construction under either Design 1 or Design 2.

This emulation framework enhances the comparability between the randomized trial and real-world data, increasing the plausibility of transportability (Assumption 2.3).

3.4.2 Initial analysis: PIONEER 6 alone

Figure 3.4.2 presents the results. Using PIONEER 6 data alone, we estimated an unadjusted risk difference of -1.32% between treatment and control groups, with a 95% confidence interval of $[-2.71\%, 0.06\%]$, which includes zero.

Williamson et al. (2014) showed that inverse probability-of-treatment weighting (IPTW) to adjust for baseline characteristics always improves the precision of treatment effect estimation. Accordingly, we applied the Augmented Inverse Propensity Weighting (AIPW) (Robins et al., 1994; Rotnitzky et al., 1998), which involves an outcome model and a propensity model. We adjusted for key predictors of MACE, including age, HbA1c, high-density lipoprotein (HDL), low-density lipoprotein (LDL), creatinine, albumin, SGLT2i use at baseline, prior revascularization, and heart failure history in the outcome and propensity models, fitted using the SuperLearner (van der Laan et al., 2007). This adjustment yielded an ATE estimate of -1.24% with a 95% CI of $[-2.55\%, 0.06\%]$ that is 5.25% narrower, compared to the unadjusted estimator.



	Risk difference (%) (95% CI)	% of P6 (unadjusted) CI length
P6 unadjusted	−1.32 [−2.71, 0.06]	100.00%
P6 AIPW	−1.24 [−2.55, 0.06]	94.75%
Anchored-thresholding	−0.87 [−1.51, −0.21]	47.05%
MSE-minimizing	−1.00 [−1.98, −0.21]	64.07%
Shrinkage (Rosenman et al., 2023a)	−1.40 [−2.41, −0.34]	75.15%
Shrinkage (Green et al., 2005)	−1.35 [−2.45, −0.29]	77.94%
Power likelihood	−1.33 [−2.41, −0.19]	80.62%
ESCV-TMLE	−1.04 [−2.38, 0.25]	95.06%
Power likelihood (control arm only)	−1.36 [−2.42, −0.23]	79.54%
ESCV-TMLE (control arm only)	−1.15 [−2.38, 0.14]	91.16%
PROCOVA (control arm only)	−1.19 [−2.51, 0.12]	95.08%

Figure 3.10: The bars represent estimated difference in 1-year MACE risk and 95% confidence/credible intervals. The columns show the width of 95% confidence/credible intervals relative to that of the unadjusted P6 estimate. P6 unadjusted: Wald test of the MACE rate in the treatment and the placebo arms of PIONEER 6. P6 AIPW: AIPW estimate after adjusting for key covariates in PIONEER 6.

3.4.3 Design 1: Incorporating control arm only

Due to the inherent variability introduced by cross-validation, we repeat the ESCV-TMLE method five times and average the results. On average, the RWD is accepted for inclusion in 10% of cross-validation splits, yielding a risk difference estimate of -1.15% (95% CI: $[-2.38\%, 0.14\%]$). This confidence interval is 9% narrower than that of the unadjusted estimator using PIONEER 6 data only.

In applying the PROCOVA method, we first follow the sample size calculation procedure in Schuler et al. (2021) to assess the potential efficiency gain from incorporating a prognostic score calibrated from RWD. The prognostic score exhibits a weak correlation with the outcome ($\rho = 0.04$), translating to an expected reduction of only 0.18% in the required sample size. The final estimated risk difference is -1.19% (95% CI: $[-2.51\%, 0.12\%]$), with precision comparable to standard covariate adjustment using trial data alone. In this case, incorporating a prognostic score from RWD adds minimal value, and the sample size calculation provides a reasonable indication of the expected efficiency gain in advance.

The power likelihood method relies on a parametric model specification. Based on the feature selection outlined in Section 3.4.1, the model includes six key variables: age, serum albumin, serum creatinine, LDL, HDL, and history of heart failure. To address outcome missingness, we apply multiple imputation (Gelman et al., 1995). Across five imputed datasets, the average selected η was 0.74. Pooling the corresponding η-powered posterior samples, the estimated ATE was -1.36% (95% CI: $[-2.42\%, -0.23\%]$), with a confidence interval approximately 20% narrower than that of the unadjusted estimator using PIONEER 6 alone. These findings provide evidence supporting the superiority of oral semaglutide over standard care for MACE prevention.

3.4.4 Design 2: Incorporating both arms

Incorporating both arms reveals the issue of treatment effect heterogeneity. To enhance the plausibility of Assumption 2.3 and maximize the efficiency gains from data fusion, we sought to minimize differences in the ATE between the RCT and RWD arising from disparities in effect modifier distributions. We apply the Causal Forest method (Wager and Athey, 2018) to the PIONEER 6 trial data, which is assumed to be free from unmeasured confounding. In practice, to maintain outcome-blindness in target trial emulation and the analyses it involves, a historical trial can be used as a proxy. We identified LDL, insulin use, serum creatinine, and HbA1c as key predictors of treatment effect heterogeneity. However, none reached statistical significance (smallest p-value: 0.17), likely due to the limited sample size.

Even if statistically significant, these variables should be interpreted as proxies for treatment effect variation rather than causal determinants. Incorporating them into the matching process, we constructed a final 1:1 matched RWD sample.

In this emulated RWD sample, we use the AIPW estimator for the ATE. The resulting ATE estimate is -0.24% (95% CI $[-2.13\%, 1.66\%]$). Although this AIPW estimate from the RWD was higher than the corresponding estimate from PIONEER 6, it remained within the RCT's 95% confidence interval. Notably, the ATE estimate from the emulated RWD sample exhibits greater variability compared to the RCT estimate. As discussed in Section 3.4.1, this uncertainty arises primarily from the limited number of semaglutide-treated patients in the RWD.

With both RWD arms incorporated, all methods except ESCV-TMLE narrowed the confidence interval by 19–53%, leading to 95% CIs below zero. The anchored thresholding estimator showed the largest gain. However, as noted in Section 3.3.2, caution is warranted when interpreting its results due to its sensitivity to violations of the comparability and ignorability assumptions of the RWD. The MSE-minimizing estimator showed the second largest efficiency gain, assigning a λ-weight of 0.26 to the RWD estimate. The two shrinkage estimators by Rosenman et al. (2023a) and Green et al. (2005) show similar outcomes in terms of estimates and reduction in confidence interval.

The power likelihood method demonstrates consistency with Design 1. The average learning factor η is 0.86, resulting in a 20% reduction in the credible interval width, further supporting its robustness across different designs. ESCV-TMLE incorporates RWD in only 2% of cross-validation folds, yielding a 5% efficiency gain. This improvement is comparable to the improvement achieved by adjusting baseline covariates in PIONEER 6 alone. This result aligns with simulation findings presented in Section 2.5, which show that the ESCV-TMLE, without the inclusion of NCOs, exhibits relatively conservative performance compared to alternative methods.

This application highlights the critical role of carefully designed simulation studies in method evaluation, providing the necessary context to interpret real-world findings appropriately.

3.5 Summary

This chapter reviewed recent data fusion methods aimed at improving efficiency of treatment effect estimation by combining RCTs and RWD. We began by highlighting two central challenges: violations of transportability and ignorability in the RWD.

These violations can bias the estimated average treatment effect (ATE) away from the target. It is therefore essential to assess the plausibility of these assumptions in any application and to consider how each method responds to their potential violations.

While methodological developments typically offer asymptotic guarantees, the simulations in Section 3.3 reveal important patterns in finite-sample behaviour. In particular, the trade-off between bias and efficiency is prominent. Efficiency gains when RWD is minimally biased may come at the cost of inflated bias and Type I error when assumptions are violated—especially within a moderate bias range (2–5 standard deviations). This trade-off is reflected in coverage rates and confidence interval lengths.

No single method dominates across all scenarios. Each carries its own trade-offs. Bias-correction methods (Kallus et al., 2018; Yang et al., 2020) improve efficiency in estimating HTEs, but these gains often disappear when marginalized over the RCT sample to estimate the ATE. The anchored thresholding approach (Chen et al., 2021) is not recommended due to its lack of robustness against bias from violations of transportability and ignorability. As bias increases, this method fails to revert to using the RCT alone.

We also discussed conservatism versus ambition in fusion methods. Conservative methods guard against bias but may yield minimal gains even in favourable settings. If a method cannot deliver meaningful power improvements when bias is absent, it may not serve its intended purpose.

Table 3.1 outlines practical considerations for method selection. Figure 3.11 provides a decision guide based on research contexts. A key initial consideration is whether treated subjects are present in the RWD. If not—e.g., for new treatments or policies—the relevant methods include the power likelihood (Lin et al., 2025), ES-CVTMLE (Dang et al., 2022), and PROCOVA (Schuler et al., 2021; Liao et al., 2025). PROCOVA, while offering type I error control and sample size calculation, may have limited power gains.

If including both arms from the RWD, the type of outcome dictates the next step. For continuous outcomes where the typical estimand is the difference in outcomes, viable options include power likelihood (Lin et al., 2025), ES-CVTMLE (Dang et al., 2022), elastic integration Yang et al. (2023), shrinkage estimators (Green et al., 2005; Rosenman et al., 2023a), and the MSE-minimizing approach (Oberst et al., 2023). The choice among these should consider the trade-off between efficiency and bias, as well as implementation complexity.

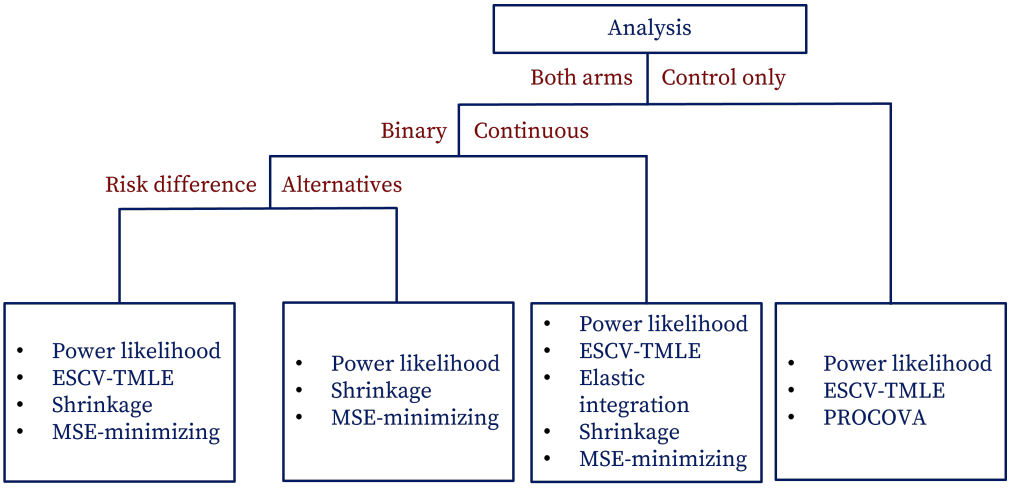


Figure 3.11: Decision framework for selecting appropriate data fusion methodologies based on analysis type, outcome type, and estimand in clinical research.

For binary outcomes, the choice of method depends on the estimand. While the current implementation of ES-CVTMLE (Dang et al., 2022) is based on the risk difference, other methods including the power likelihood (Lin et al., 2025), the two shrinkage methods (Green et al., 2005; Rosenman et al., 2023a), and the MSE-Minimizing approach (Oberst et al., 2023) can adapt to alternative estimands like relative risk.

Our application to the PIONEER 6 trial demonstrates how to emulate a target trial and construct a high-quality RWD cohort in an outcome-blind manner. This example provides a practical template for implementing target trial emulation and applying data fusion methods under realistic constraints.

This chapter offers a structured overview of efficiency-focused data fusion methods, summarising their assumptions, trade-offs, and empirical performance. It aims to support applied researchers in selecting appropriate methods and to inform ongoing methodological development in this area.

Chapter 4

Exact simulation of longitudinal data from marginal structural models

Abstract. *Simulating longitudinal data from specified marginal structural models is a crucial but challenging task for evaluating causal inference methods and informing study design. While data generation typically proceeds in a fully conditional manner using structural equations according to a temporal ordering, it is difficult to ensure alignment between conditional distributions and the target marginal causal effects, which presents a fundamental challenge. To address this, we propose a flexible and efficient algorithm for simulating longitudinal data that adheres exactly to a specified marginal structural model. Our approach accommodates time-to-event outcomes and extends naturally to survival settings, which are prevalent in applied research. Compared to existing approaches, it offers several advantages: it enables exact simulation from a known causal model rather than relying on approximations; avoids restrictive assumptions about the data-generating process; and remains computationally efficient by requiring only the evaluation of analytical expressions, rather than Monte Carlo methods or numerical integration. Through simulation studies replicating realistic scenarios, we validate the method's accuracy and utility. Our method will facilitate researchers in effectively simulating data with target causal structures for their specific scenarios.*

Contents

4.1 Introduction 92
4.2 Related literature 94
4.3 Copulas 95
4.3.1 Pair-copula construction 96
4.3.2 Simulating from pair-copula constructions 97
4.4 Simulation algorithm 98
4.4.1 The fully-conditional approach 99
4.4.2 Frugal parameterization for outcome simulation 99
4.4.3 Simulation algorithm 101
4.4.4 Illustrative example 102
4.5 Simulating survival outcomes 103
4.5.1 Simulation algorithm 106
4.5.2 Illustration: Concomitant treatment 111
4.5.3 Competing risks 115
4.6 Discussion 118

4.1 Introduction

Longitudinal data are fundamental for studies analyzing the causal effect of a treatment or exposure, especially for research into chronic diseases such as diabetes or HIV. Unlike cross-sectional studies, longitudinal designs capture the dynamic nature of treatment regimens, which often vary over time due to clinical decisions, patient adherence, or external factors. Even in clinical trials with strict adherence protocols, intercurrent events—such as treatment discontinuation or supplementary medication use—are common. By accounting for these treatment dynamics, longitudinal data enable the estimation of meaningful causal quantities, such as the "on-treatment" estimand, which quantifies the treatment effect assuming treatment is administered as intended.

A central challenge in analyzing longitudinal data is addressing time-varying confounding, where a time-dependent covariate predicts both the outcome and future treatments, while itself being influenced by past treatments. For example, in the case study by Robins et al. (2000) on the effect of zidovudine (AZT) treatment on HIV infection, the biomarker CD4 lymphocyte count is both affected by prior AZT treatment and is predictive of future therapy initiation and HIV progression. The presence of such time-varying confounding renders conventional adjustment methods, like regression and stratification, generally biased.

Marginal structural models, introduced by Robins (1997), have emerged as a powerful tool for estimating causal effects under time-varying confounding. Model parameters can be consistently estimated using inverse probability of treatment weighting (Robins, 1999, 2000). Among alternative methods, including structural nested models estimated through g-estimation (Robins, 1989, 1992; Robins and Tsiatis, 1991), marginal structural models have gained prominence due to their interpretability and ease of implementation. A systematic review by Clare et al. (2019) highlights their widespread use in applied research.

As causal inference methods for longitudinal data are continually proposed to address complex real-world settings, simulation studies are essential to evaluate their effectiveness for practical use. These studies assess important properties such as bias, confidence interval coverage, relative efficiency, and robustness to violations of the assumptions used, guiding both methodological development and application. Sample size calculations also benefit from simulations. Prior to conducting a trial or an observational study, it is crucial to determine the number of participants or observed units required to achieve adequate statistical power. Although standard formulas can sometimes be used for time-fixed treatments, more complex scenarios with time-varying treatments make deriving analytical formulas impractical or impossible. In such cases, simulation is invaluable, allowing researchers to estimate sample sizes based on realistic data-generating processes and causal structures.

A fundamental requirement for both method evaluation and sample size determination is the ability to simulate data that conform to a specified causal model. However, simulating longitudinal data from a pre-specified marginal structural model is not a straightforward task. While generating data sequentially from structural equations—creating covariates, treatment status, and then the outcome conditioned on past information—may seem intuitive, these models rely on conditional rather than marginal specifications. This makes it difficult to sequentially simulate from conditional distributions that align with a pre-specified marginal model; in particular, it is extremely challenging to sample from a null hypothesis—the so-called *g-null paradox* (Robins and Wasserman, 1997).

The remainder of this paper is structured as follows: Section 4.2 reviews existing methods for simulating longitudinal data from causal models. Section 4.3 introduces key preliminaries, including the frugal parameterization (Evans and Didelez, 2024) and copulas. In Section 4.4, we introduce our proposed methodology. Section 4.5 extends the approach to time-to-event data, enabling simulations from survival models such as Cox marginal structural models, and Section 4.5.2 illustrates the

method through a simulation study replicating a real-world clinical scenario involving concomitant treatment in a trial. We conclude with a discussion in Section 4.6. Example code for the simulations in Sections 4.4.4 and 4.5.2 is available at `https://github.com/XiLinStats/LongSiMSM`.

4.2 Related literature

As noted in Section 4.1, the challenge lies in ensuring that the simulated data adhere to a specified *marginal* model, rather than simply following sequential conditional distributions in a structural equation model. Existing approaches address this issue through two strategies:

(i) Simulating from conditional models and deriving or approximating the corresponding marginal models post hoc.

(ii) Bypassing conditional simulations and simulating directly from the target marginal model.

Several studies have explored the first approach. Xiao et al. (2010) used sequential conditional models to approximate Cox marginal structural models under the rare disease assumption, i.e., very low failure rates, which mitigates the impact of non-collapsibility but assumes the absence of unmeasured confounding between time-varying confounders and outcomes. Young and Tchetgen Tchetgen (2014) employed the g-formula to bridge conditional and marginal models in discrete-time Cox marginal structural models. While theoretically feasible for discrete time-varying covariates, deriving marginal models becomes intractable for continuous covariates due to the need for numerical integration. Simplifications, such as the Markov property or rare disease assumptions, are often necessary but limit generalizability. Sofrygin et al. (2017) developed the `simcausal` R package for flexible simulation from nonparametric structural equation models. The workflow involves specifying structural equations and sequentially simulating data from conditional distributions. However, deriving analytic expressions for marginal causal quantities under this conditional parameterization is often intractable and prone to model misspecification. To address this, users typically specify low-dimensional *working* marginal structural models (Neugebauer and van der Laan, 2007) and estimate their parameters via Monte Carlo simulations of counterfactual data. This introduces two limitations: first, the true marginal model remains unknown; second, reverse-engineering simulations to match a target causal function is challenging. Keogh et al. (2021) described simulation from additive-hazard marginal structural models (Aalen, 1989). They utilized

the result that if the conditional hazard model follows an additive form, the equivalent marginal hazard model will also be additive. While the marginal model retains the correct form, true parameter values are generally intractable and require Monte Carlo approximations. A further limitation is the difficulty in ensuring non-negative hazards for continuous covariates.

Alternative methods focus on simulating directly from the target marginal model. Havercroft and Didelez (2012) simulated directly from a discrete Cox marginal structural model, introducing time-varying confounding indirectly via a shared latent variable. However, this approach assumes no direct effect of time-varying confounders on outcomes, which may not hold in practice, e.g., the effect of CD4 count on HIV survival (Robins et al., 2000). Evans and Didelez (2024) proposed a *frugal parameterization* that prioritizes the causal margin and parameterizes the surrounding variables accordingly, enabling direct simulation from the exact marginal model while explicitly modeling time-varying covariates. This addresses key limitations of earlier methods. Building on this, Seaman and Keogh (2024) applied the frugal parameterization to simulate survival outcomes, using a one-dimensional "risk score" to summarize multi-dimensional covariates. While simplifying dependence structures, this method relies on computationally expensive Monte Carlo approximations to estimate the cumulative distribution functions of the risk score. In settings with continuous treatments or baseline variables, these approximations must be performed separately for each individual at each time point, substantially increasing the computational burden.

The aims for a new simulation approach therefore include the ability to: simulate exactly from a specified marginal structural model; include multiple time-varying covariates, treatments, outcomes, and censoring; have good control over their distributions; avoid restrictive assumptions about the data-generating process. In addition, we would like it to scale well in the number of time-points and variables, and applicable in real-world settings.

4.3 Copulas

We can think of continuous multivariate distributions as the joining together of the univariate marginals (or margins) and the dependence structure. This structure can be represented by a copula, which is a multivariate distribution function $C : [0,1]^d \to [0,1]$ with margins uniformly distributed on $(0,1)$. According to Sklar's theorem (Sklar, 1959, 1973), let F be a multivariate distribution with margins $F_1, \ldots, F_d$.

Then there exists a copula C such that for all $(x_1, \dots x_d)$ in $\mathbb{R}^d$,

$$F(x_1, \dots, x_d) = C\left\{F_1(x_1), \dots, F_d(x_d)\right\} = C(u_1, \dots, u_d),$$

where $u_i = F_i(x_i)$, $i = 1, \dots, d$. Throughout, we refer to such u_i as *ranks* for simplicity, although strictly speaking they correspond to probability integrals under the Probability Integral Transform (Rosenblatt, 1952). If $F_1(x_1), \dots, F_d(x_d)$ are all continuous, then C is unique and takes the form

$$C(u_1, \dots, u_d) = F\{F_1^{-1}(u_1), \dots, F_d^{-1}(u_d)\}, \qquad (u_1, \dots, u_d) \in [0,1]^d.$$

Otherwise, C is only uniquely determined on $\operatorname{Ran} F_1 \times \cdots \times \operatorname{Ran} F_d$, where $\operatorname{Ran} F_i$ denotes the range of F_i. The choice of copula can be considered independently of the univariate margins.

4.3.1 Pair-copula construction

While there exist a wide variety of parametric bivariate copulas, the options for higher-dimensional copulas are limited, with Gaussian and Student t-copulas being common choices. Extending symmetric bivariate Archimedean copulas to multivariate settings is also possible but imposes strong restrictions, such as requiring the same level of dependence among all possible pairs. Building on the work of Joe (1996), Bedford and Cooke (2001, 2002) introduced pair-copula constructions, a flexible and intuitive approach to constructing multivariate copulas by modeling multivariate data as a series of bivariate copulas. This hierarchical structure incorporates additional variables at each level, using pair-copulas as the fundamental building blocks.

Following the general formula given in Joe (1997), and assuming that the joint distribution of $(Y, X_1, \dots, X_d)$ is absolutely continuous with continuous margins, the conditional distribution of Y given d other variables can be decomposed into d bivariate copula density functions, each of which can be specified flexibly:

$$f(y \,|\, \boldsymbol{x}) = c_{Y X_j | \boldsymbol{X}_{-j}} \left\{F(y \,|\, \boldsymbol{x}_{-j}), F(x_j \,|\, \boldsymbol{x}_{-j})\right\} \cdot f(y \,|\, \boldsymbol{x}_{-j}), \tag{4.1}$$

where $\boldsymbol{x}$ is a d-dimensional vector; x_j is an arbitrarily chosen component of $\boldsymbol{x}$, and $x_{-j} := (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d)$ denotes $\boldsymbol{x}$ with its jth entry removed. Applying (4.1) recursively to the conditional densities, we have:

$$f(y \,|\, \boldsymbol{x}) = f(y) \cdot \prod_{i=1}^{d} c_{Y X_{d+1-i} | \boldsymbol{X}_{1:d-i}} \{F(y \,|\, \boldsymbol{x}_{1:d-i}), F(x_{d+1-i} \,|\, \boldsymbol{x}_{1:d-i})\}, \tag{4.2}$$

where each copula in the product in (4.2) is bivariate; this defines a *pair-copula*. In principle, any ordering of the conditioned variables $x_1, x_2, \ldots, x_d$ can be selected, although some orderings may be easier to use than others. In our proposed method, we choose an ordering that is convenient and efficient for simulation.

In conclusion, under appropriate regularity conditions, a multivariate density can be expressed as a product of pair-copulas, acting on different conditional distributions.

4.3.2 Simulating from pair-copula constructions

The pair-copula construction includes conditional distributions of the form $F(y \mid \boldsymbol{x})$. Assuming that the joint distribution of $(Y, X_1, \ldots, X_d)$ is absolutely continuous with continuous margins, and the bivariate copula $C_{YX_j|\boldsymbol{X}_{-j}}$ is differentiable, Joe (1996) demonstrated the following result:

$$F(y \mid \boldsymbol{x}) = \frac{\partial C_{YX_j|\boldsymbol{X}_{-j}} \{F(y \mid \boldsymbol{x}_{-j}), F(x_j \mid \boldsymbol{x}_{-j})\}}{\partial F(x_j \mid \boldsymbol{x}_{-j})}, \tag{4.3}$$

where $j \in \{1, \ldots, d\}$. For notational convenience, we introduce the *h-function* $h(u_1, u_2; \Theta)$ from Aas et al. (2009). Let u_1 and u_2 be random variables with a uniform distribution on $(0, 1)$, for a sufficiently regular copula $C_{12}(u_1, u_2)$, the h-function is defined as

$$h(u_1, u_2; \Theta) := F(u_1 \mid u_2) = \frac{\partial}{\partial u_2} C_{12}(u_1, u_2; \Theta)$$

where the second argument in $h(\cdot; \Theta)$ is the conditional variable, and Θ represents the set of copula parameters. The h-function is effectively the conditional copula $C(u_1 \mid u_2; \Theta)$. We further define $h^{-1}(u_1, u_2; \Theta)$, be the inverse of the h-function in the first variable u_1. Equivalently, $h^{-1}(u_1, u_2; \Theta) := F^{-1}(u_1 \mid u_2)$.

Remark 4.1. Here we define a copula C as *sufficiently regular* if it has non-zero density everywhere on its domain. Under this condition, the associated h-functions exist and are invertible. For standard parametric families—including the Gaussian, Student-t, Clayton and Frank copulas—the analytical form of $h(\cdot)$ and its inverse $h^{-1}(\cdot)$ can be derived directly from their definitions. See Appendix C of Aas et al. (2009).

The h-functions are foundational for simulating from $F(y \mid \boldsymbol{x})$. Similar to (4.2), by imposing an ordering $1, \ldots, d$ on the components of $\boldsymbol{x}$, we can recursively apply h-functions to obtain:

$$F(y \mid \boldsymbol{x}) = h\{F(y \mid \boldsymbol{x}_{1:d-1}), F(x_d \mid \boldsymbol{x}_{1:d-1}), \Theta_d\}$$

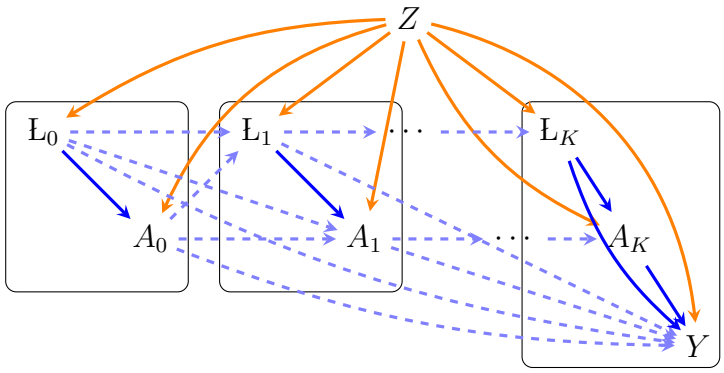


Figure 4.1: An example directed acyclic graph showing temporal relationships in a longitudinal discrete setting, with baseline covariates Z; treatments $A_0, \ldots, A_K$; time-varying covariates $L_0, \ldots, L_K$ and an outcome Y observed at the end of the K follow up intervals. Given the topological ordering $Z, L_0, A_0, L_1, A_1 \ldots, L_K, A_K, Y$, any variable may depend upon any subset of the previous variables.

$$
\begin{aligned}
&= h[h\{F(y \,|\, \boldsymbol{x}_{1:d-2}), F(x_{d-1} \,|\, \boldsymbol{x}_{1:d-2}; \Theta_{d-1}\}, F(x_d \,|\, \boldsymbol{x}_{1:d-1}); \Theta_d] \\
&\ldots \\
&= h[\ldots h\{h(F(y), F(x_1); \Theta_1), F(x_2 \,|\, x_1); \Theta_2\}, \ldots, F(x_d \,|\, \boldsymbol{x}_{1:d-1}); \Theta_d]. \quad (4.4)
\end{aligned}
$$

Intuitively, the bivariate copulas "weave" the dependence between the outcome and each of the covariates. Thus, given a value for $\boldsymbol{X}$ we can simulate Y by first drawing a uniform $U \sim \mathrm{U}(0, 1)$, and then evaluating:

$$
F(Y) = h^{-1}[\ldots h^{-1}\{h^{-1}(U,\ F(X_1); \Theta_1), F(X_2 \,|\, X_1); \Theta_2\}, \ldots, F(X_d \,|\, X_1, \ldots, X_{d-1}); \Theta_d]. \quad (4.5)
$$

This means that to simulate Y given a d-dimensional vector $\boldsymbol{X}$, we recursively calculate the d inverse h-functions, forming a pair-copula construction. This underpins our proposed simulation method.

4.4 Simulation algorithm

We first consider a longitudinal study with n individuals observed at discrete time points, $\mathcal{K} = \{0, \ldots, K\}$. At each follow-up k, treatment status A_k, and a set of time-varying covariates L_k are recorded. Additionally, Z represents a set of static baseline covariates. Latent variables, if present, are written as $\{H_k\}_{k \in \mathcal{K}}$. To indicate the history of a variable up to time k, we use bar notation. For instance, $\bar{A}_k = \{A_0, \ldots, A_k\}$ and $\bar{L}_k = \{L_0, \ldots, L_k\}$, with the convention $\bar{A}_{-1} = \emptyset$ and $\bar{L}_{-1} = \emptyset$. The outcome Y is observed at the final time point K. Figure 4.1 is a graphical representation of this setup, which is common in medical and epidemiological research. Examples include

examining the effects of an HIV treatment on infection status assessed at the end of the study period (Robins et al., 2000), and assessing the impact of iron supplement use during pregnancy on the odds of anemia at delivery (Bodnar et al., 2004). In Section 4.4.4, we illustrate our method through a simulation study that replicates such settings.

Following the notation of Evans and Didelez (2024), we use an asterisk to represent causal or interventional distributions where treatment is set to a specific value $A_k = a$ via intervention. In graphical terms, this corresponds to removing all arrows directed into A_k.

4.4.1 The fully-conditional approach

To outline our approach, we initially assume the time-varying covariate L_k is one-dimensional and omit any latent process $\{H_k\}_{k\in\mathcal{K}}$; the extension to multi-dimensional time-varying covariates and latent variables is discussed in Appendix D.1.

The standard method for simulation of longitudinal data $(Z, \bar{L}_K, \bar{A}_K, Y)$ is by time step in blocks using the Rosenblatt transformation (Rosenblatt, 1952). Specifically:

1. **Baseline simulation**: Simulate static baseline variables, z following a specified factorization, e.g. $F_{Z_1}, F_{Z_2|Z_1}, \ldots$.

2. **Time-step simulation**: At each time point $k = 0, \ldots, K$, simulate $l_k = F^{-1}_{L_k|Z,\bar{L}_{k-1},\bar{A}_{k-1}}(w_{k,1}, \mid z, \bar{l}_{k-1}, \bar{a}_{k-1})$ and $a_k = F^{-1}_{A_k|Z,\bar{L}_k,\bar{A}_{k-1}}(w_{k,2}, \mid z, \bar{l}_k, \bar{a}_{k-1})$, where $w_{k,1}, w_{k,2} \sim \mathrm{U}(0,1)$.

3. **Outcome simulation**: Given the simulated covariates and treatments, generate the outcome $y = F^{-1}_{Y|Z,\bar{L}_K,\bar{A}_K}(w_y \mid z, \bar{l}_K, \bar{a}_K)$, with $w_Y \sim \mathrm{U}(0,1)$.

Our proposed method diverges from this fully-conditional approach at Step 3. Instead of specifying the conditional distribution $F_{Y|Z\bar{L}_K\bar{A}_K}$—which inherently introduces misalignment between conditional and the desired marginal models—we introduce a margin-preserving parameterization for outcome simulation.

4.4.2 Frugal parameterization for outcome simulation

Our method uses the frugal parameterization to simulate outcomes, avoiding potentially intractable analytical calculations to obtain the marginal structural model. Evans and Didelez (2024) showed that the fully-conditional distribution of the outcome can be *frugally* parameterized as:

$$p_{Y|\bar{L}_K,\bar{A}_K}(y \mid \bar{l}_k, \bar{a}_K) = p^*_{Y|\bar{A}_K}(y \mid do(\bar{a}_K)) \cdot c^*_{Y\bar{L}_K|\bar{A}_K}(u^*_Y, u_{L_0}, \ldots, u_{L_K} \mid do(\bar{a}_K)), \quad (4.6)$$

where $u_Y^* = F_{Y|\bar{A}_K}(y \,|\, do(\bar{a}_K))$ and $u_{L_k} = F_{L_k|\bar{A}_K}(l_k \,|\, do(\bar{a}_K))$ for $k = 0, \ldots, K$ are the marginal ranks under interventions that set $\bar{A}_K = \bar{a}_K$.

Remark 4.2. For brevity, we omit conditioning on baseline variables Z in all quantities in (4.6), treating them as if no baseline variables exist. This omission will continue hereafter, though all expressions can be extended to condition on Z by including it in their respective conditioning sets.

This decomposition implies that we can simulate Y through the multivariate copula and its interventional margin. Applying the pair-copula construction in (4.2) to (4.6), we obtain:

$$p(y \,|\, \bar{l}_K, \bar{a}_K) = p(y \,|\, do(\bar{a}_K)) \cdot \prod_{k=0}^{K} c^*\{F(y \,|\, \bar{l}_{k-1}, do(\bar{a}_K)), F(l_k \,|\, \bar{l}_{k-1}, do(\bar{a}_K)); \Theta_k\}. \tag{4.7}$$

Note that we omit some subscripts on densities for conciseness where doing so does not introduce ambiguity. This decomposes the conditional density into the product of $K+1$ pair-copulas and the target marginal causal density. The pair-copula linking Y and L_k is parameterized by Θ_k, which may depend on intervened treatments $do(\bar{A}_k)$, past time-varying confounders $\bar{L}_{k-1}$ and baseline covariates $\boldsymbol{Z}$. Under intervention, graphically speaking, all arrows into nodes $\bar{A}_k$ are removed in the directed acyclic graph, rendering L_k independent of subsequent interventions. Consequently, $F(l_k \,|\, \bar{l}_{k-1}, do(\bar{a}_K)) = F(l_k \,|\, \bar{l}_{k-1}, \bar{a}_{k-1})$ for $k = 1, \ldots, K$ and $F(l_0 \,|\, do(\bar{a}_K)) = F(l_0)$. Following the simulation strategy from pair-copula constructions outlined in Section 4.3.2, the fully-conditional distribution of the outcome is alternatively parameterized as:

$$F(y \,|\, \bar{l}_K, \bar{a}_K) = h\big[\ldots h\big\{h(F^*_{Y|\bar{A}_K}(y \,|\, do(\bar{a}_K)),\, F(l_0); \Theta_0),\, F(l_1 \,|\, l_0, a_0); \Theta_1\big\} \ldots; \Theta_K\big], \tag{4.8}$$

which is centered around the causal margin $F^*_{Y|\bar{A}_K}$.

We write $h_k(u_1, u_2) = h(u_1, u_2; \Theta_k)$, so as to avoid including the parameters Θ_k. By the inversion method, it is straightforward that given $(\bar{l}_K, \bar{a}_K)$ and $w \sim \mathrm{U}(0,1)$,

$$y = F^{*-1}_{Y|\bar{A}_K}(h_0^{-1}\big[\ldots h_{K-1}^{-1}\big\{h_K^{-1}(w, F(l_K \,|\, \bar{l}_{K-1}, \bar{a}_{K-1})), F(l_{K-1} \,|\, \bar{l}_{K-2}, \bar{a}_{K-2})\big\}, \ldots, F(l_0)\big])$$

follows the distribution parameterized by (4.7).

This result means that following the approach outline in Section 4.4.1, given the simulated baseline variables and time-varying variables, we can simulate the outcome in two steps:

1. Calculate the marginal rank of the outcome, u_Y^*, as:

$$u_Y^* = h_0^{-1}\big[\ldots h_{K-1}^{-1}\big\{h_K^{-1}(w_Y, F(l_K \,|\, \bar{l}_{K-1}, \bar{a}_{K-1})), F(l_{K-1} \,|\, \bar{l}_{K-2}, \bar{a}_{K-2})\big\}\ldots, F(l_0)\big] \tag{4.9}$$
$$= h_0^{-1}\big[\ldots h_{K-1}^{-1}\big\{h_K^{-1}(w_Y, w_{K,1}), w_{K-1,1}\big\}, \ldots, w_{0,1}\big],$$

   where $w_Y \sim \mathrm{U}(0,1)$.

2. Obtain y through the inverse cumulative distribution function of the marginal distribution: $y = F^{*-1}_{Y|\bar{A}_K}(u_Y^* \,|\, do(\bar{a}_K))$.

An intuitive way to understand the calculation in Step 1 is that the inverse h-functions "unravel" the dependencies between Y and each of $L_K, L_{K-1}, \ldots L_0$ sequentially, and remove the conditioning of Y on $\bar{L}_K$ until the marginal rank of Y is obtained. The ranks in the inverse h-functions, such as $F(l_k \,|\, \bar{l}_{k-1}, \bar{a}_{k-1})$, correspond exactly to the random standard uniform variables $w_{k,1}$ used to simulate l_k at time step k. This enables direct substitution of the l_k ranks into (4.9) without additional computation.

4.4.3 Simulation algorithm

In summary, to simulate the data $(Z, \bar{L}_K, \bar{A}_K, Y)$, the user specifies:

- the population distribution of the static baseline covariates Z: F_Z, either as a joint distribution, or through a conditional factorization;
- the development of the time-varying confounder L_k as influenced by $\boldsymbol{Z}$ and $\bar{A}_{k-1}$: $F_{L_k|Z,\bar{L}_{k-1},\bar{A}_{k-1}}$;
- treatment assignment based on the time-varying confounder $\bar{L}_k$, past treatment and baseline covariates Z: $F_{A_k|\boldsymbol{Z},\bar{L}_k,\bar{A}_{k-1}}$;
- the target causal marginal structural model: $F^*_{Y|\bar{A}_k,Z}(\cdot \,|\, do(\bar{a}_K), z)$;
- $K+1$ bivariate copulas describing dependencies between outcome Y and $L_0, \ldots, L_K$:

$$C_{YL_k|Z,do(\bar{A}_k),\bar{L}_{k-1}}(\cdot;\Theta_k) \quad \text{for} \quad k = 0, \ldots, K.$$

Section 4.4.4 provides an example of specifying and interpreting these inputs through a simulation study that captures real-world complexities.

Algorithm 2 below details the simulation procedure:

Extending the simulation algorithm to incorporate multiple time-varying confounders—whether assumed to be observed or unobserved — is straightforward. The key

Algorithm 2 Simulation algorithm for an individual i over K follow-ups

Sample $w_z \sim U(0,1)$
$z \leftarrow F_Z^{-1}(w_z)$
Let $k = 0$
while $k \leq K$ **do**
  Sample $w_{k,1}, w_{k,2}$ independent standard uniform variables
  $l_k \leftarrow F^{-1}_{L_k|Z,\bar{L}_{k-1},\bar{A}_{k-1}}(w_{k,1} \mid z, \bar{l}_{k-1}, \bar{a}_{k-1})$
  $a_k \leftarrow F^{-1}_{A_k|Z,\bar{L}_k,\bar{A}_{k-1}}(w_{k,2} \mid z, \bar{l}_k, \bar{a}_{k-1})$
  $k \leftarrow k + 1$
end while
Sample $w_y \sim \mathrm{U}(0,1)$
Let $\nu_0 = w_y$
for $j = 0$ to K **do**
  $\nu_{j+1} \leftarrow h^{-1}_{K-j}(\nu_j, w_{(K-j),1})$
end for
$u_y^* \leftarrow \nu_{K+1}$
$y \leftarrow F^{*-1}_{Y|\bar{A}_k,Z}(u_y^* \mid do(\bar{a}_K), z)$
return $(z, \bar{l}_K, \bar{a}_K, y)$

modification lies in the outcome simulation: the marginal rank of Y is obtained by sequentially removing the conditioning on all the time-dependent confounders at each step, through inverse h-functions. We provide an illustrative example in Appendix D.1, outlining the simulation procedure for incorporating an additional latent process $\{H_k\}_{k\in\mathcal{K}}$.

4.4.4 Illustrative example

We demonstrate our simulation method by replicating the setting in Bodnar et al. (2004), which examines the causal effect of iron supplement use during pregnancy on the odds of anemia at delivery. Participants are randomized into treatment and control groups and followed up over four visits leading up to delivery. Treatment is not time-fixed as participants are assumed to adhere to assigned treatment up to $k = 2$, after which standard clinical protocols dictate further treatment based on hemoglobin levels.

The simulation considers baseline variables $Z = \{\mathrm{B} = \text{Pre-pregnancy obesity}, \mathrm{C} = \text{Age}\}$ and time-varying covariate $L = \{\text{Hemoglobin level (g/dL)}\}$. The time-dependent treatment, A_k, indicates whether an iron supplement is prescribed. The outcome Y is measured at delivery and $Y = 1$ indicates anemia at delivery.

The distribution of baseline variables is: $B \sim \text{Bernoulli}(0.1)$, $C \sim U(25, 35)$, $L_0 \sim N(11 - 0.05\,B - 0.02\,C,\, 0.5)$ and $A_0 \sim \text{Bernoulli}(0.5)$. For subsequent follow-up

visits $k > 0$ we have $L_k \sim \mathrm{N}(L_{k-1} + 0.5\, A_{k-1},\, 0.1)$, and $A_2 = A_1 = A_0$ with $A_k \sim \text{Bernoulli}(\text{expit}(10 - L_k + 0.1\, A_{k-1}))$ for $k = 3, 4$. We specify a logistic discrete-time hazard marginal structural model:

$$\mathbb{P}[Y = 1 \,|\, do(\bar{A}_4), B, C] = \text{expit}(-2 + 0.1\, B + 0.02\, C - \beta\ \text{cum}(\bar{A}_4)),$$

where $\text{cum}(\bar{A}_4) = \Sigma_{i=0}^{4} A_i$. Here, $\exp(\beta)$ represents the causal odds ratio for an incremental dose, and $\exp(4\beta)$ corresponds to the "always treated" efficacy estimand.

We use a t-copula $T_{\rho,\text{df}}$ (with $\rho = -0.5$ and $\text{df} = 5$) across all pair-copula construction components to model the dependency between Y and each of L_0, L_1, L_2, L_3, and L_4. Compared to a Gaussian copula, the t-copula captures heavier-tailed dependency, capturing stronger associations in the tails reflecting stronger associations in extreme scenarios (e.g., severely low hemoglobin levels and anemia). Our algorithm offers full flexibility, allowing users to choose any copula family and specify different parameters for each bivariate copula in the pair-copula construction.

We simulated scenarios with varying causal coefficients $\beta = \{-0.5, -0.3, 0\}$, representing strong, moderate and no effect, respectively and samples sizes $n = \{500, 1000, 2000\}$. For each combination, we generated 400 datasets. The model $Y = \text{expit}(\alpha + \gamma\, B + \theta\, C - \beta\ \text{cum}(\bar{A}_4))$ is fitted using stabilized inverse probability of treatment weights, and an unweighted approach. For each fit, we estimated 95% confidence intervals for β using both the robust sandwich estimator and non-parametric bootstrap.

Results are summarized in Table 4.1, 4.2 and 4.3. The inverse-probability-weighted estimator shows minimal bias, with magnitude well below Monte Carlo standard errors. In contrast, the unweighted estimator remains significantly biased across sample sizes, showing no convergence. Both the sandwich and bootstrap CIs of the inverse-probability-weighted estimator approach 95% coverage, though with minor under-coverage. The unweighted estimator, however, exhibits poor coverage due to bias. Statistical power for testing the treatment effect (null hypothesis: $\beta = 0$) increases with effect size and sample size for the weighted estimator. In summary, these results confirm that data generated using our simulation algorithm accurately reflect the intended causal effects and time-varying confounding is introduced.

4.5 Simulating survival outcomes

We have described and demonstrated our method to simulate longitudinal data where the outcome is measured at the final time point. Since time-to-event and

β	n	**IPTW**	**Unweighted**
-0.5	500	0.027 (0.186)	0.270 (0.115)
	1000	0.012 (0.140)	0.269 (0.079)
	2000	0.009 (0.096)	0.274 (0.055)
-0.3	500	0.015 (0.179)	0.236 (0.103)
	1000	0.004 (0.119)	0.234 (0.073)
	2000	0.009 (0.088)	0.235 (0.051)
0	500	0.017 (0.138)	0.189 (0.083)
	1000	0.015 (0.087)	0.188 (0.059)
	2000	0.014 (0.072)	0.193 (0.043)

Table 4.1: Bias and Monte Carlo standard errors (in parentheses, defined as $sd(\hat{\beta})$) for inverse-probability-of-treatment-weighted ('IPTW') and unweighted estimators ('unweighted') across different β and n scenarios.

β	n	**IPTW**		**Unweighted**
		Sandwich	**Bootstrap**	
-0.5	500	91.8%	93.0%	31.8%
	1000	92.0%	91.8%	7.5%
	2000	92.3%	92.8%	0.0%
-0.3	500	91.0%	91.5%	35.3%
	1000	94.3%	92.5%	8.0%
	2000	94.3%	93.8%	0.3%
0.0	500	92.8%	93.0%	41.8%
	1000	95.5%	94.3%	12.0%
	2000	92.8%	90.0%	0.5%

Table 4.2: Coverage of 95% confidence intervals for IPTW-sandwich, IPTW-bootstrap, and unweighted estimators across scenarios with varying β and n values.

survival outcomes are very common in medical and epidemiological contexts, we shall now demonstrate how our method can be extended to simulate longitudinal data from survival models.

In Figure 4.2, each plate represents a single interval. For the kth interval, where $k = 0, \ldots, K$, the observations include L_k, A_k at the start of the interval, and the survival status Y_{k+1} at the end. The length of each interval is set to 1. In addition to the notation in Section 4.4, let T denote an individual's failure time. The outcome is represented be a sequence $\{Y_k\}_{k \in \mathcal{K}^+}$, where $\mathcal{K}^+ = \{1, \ldots, K+1\}$ and $Y_{k+1} = I(T < k+1)$ indicates failure.

We describe the algorithm for the case where the time-varying confounder is

β	n	**IPTW**		**Unweighted**
		Sandwich	**Bootstrap**	
-0.5	500	83.3%	79.5%	55.3%
	1000	97.5%	97.8%	82.5%
	2000	100.0%	100.0%	98.3%
-0.3	500	45.8%	43.0%	12.0%
	1000	80.0%	77.8%	15.3%
	2000	95.8%	96.5%	26.8%
0.0	500	7.3%	7.0%	58.3%
	1000	4.5%	5.8%	88.0%
	2000	7.3%	10.0%	99.5%

Table 4.3: Statistical power to reject the null hypothesis $\beta = 0$; that is, no causal effect.

univariate and no latent processes are present. The algorithm can be extended analogously to the case of ordinary longitudinal outcomes. The illustrative example at the end of this section demonstrates the simulation with two time-varying confounders.

Depending on the chosen marginal structural model, there are various ways to model the outcomes. Examples include:

- Discrete-time marginal structural model, where the binary survival outcome Y_k is modeled directly:

$$P(Y_{k+1} = 1 \,|\, do(\bar{A}_k = \bar{a}_k), Z = z, Y_k = 0) = f(\bar{a}_k, z; \alpha), \tag{4.10}$$

  where f can be any real-valued functions bounded between 0 and 1, for example, an expit function for a logistic marginal structural model.

- Continuous-time hazard MSM, where we introduce an auxiliary continuous variable, $\tilde{Y}_{k+1} = T - k \,|\, T > k$, representing the incremental failure time given survival at the start of the kth interval. We model incremental survival time $\tilde{Y}_{k+1}$ and truncated it to obtain the observed binary survival outcome Y_{k+1}. Given survival at k, the time to failure is $T \,|\, T > k = k + \tilde{Y}_{k+1}$. Examples of marginal models for $\tilde{Y}_{k+1}$ include:

  - Cox proportional hazard:

$$\lambda(\tilde{y}_{k+1} \,|\, do(\bar{a}_k), z, y_k = 0) = \lambda(\tilde{y}_{k+1} \,|\, do(\bar{0}_k), z, y_k = 0) \exp\{g(\bar{a}_k, z; \beta)\},$$

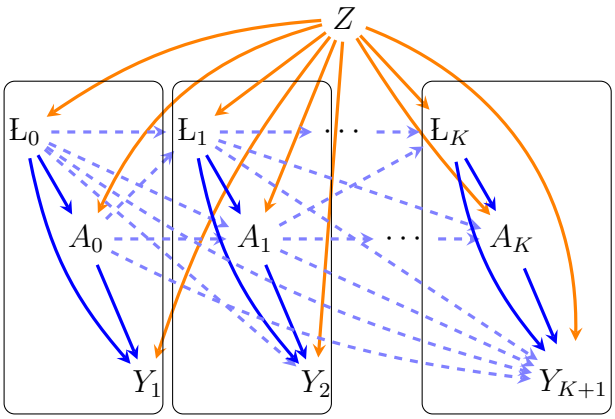


Figure 4.2: Example directed acyclic graph showing temporal relationships in a longitudinal survival setting. The sequence of outcomes observed over time is denoted as $\{Y_k\}_{k\in\mathcal{K}^+}$, where $\mathcal{K}^+ = \{1, \ldots, K+1\}$.

– Additive hazard:

$$\lambda(\tilde{y}_{k+1} \,|\, do(\bar{a}_k), z, y_k = 0) = \lambda(\tilde{y}_{k+1} \,|\, do(\bar{0}_k), z, y_k = 0) + g(\bar{a}_k, z; \gamma).$$

We focus on describing the simulation algorithm for time-to-event outcomes, with a remark on adapting the approach for discrete survival outcomes, which follows a similar but simpler framework.

4.5.1 Simulation algorithm

The general approach remains similar to the non-survival case, but with the key difference that at each time step k, simulation is now conditioned on survival up to the start of the current time interval. For brevity, we continue to omit conditioning on Z.

Unlike the derivation of outcome simulation for the non-survival case in Section 4.4.2, all quantities involved in the copulas must now be conditioned on survival up to this point. Specifically, when simulating the incremental failure time, $\tilde{y}_{k+1}$, given $Y_k = 0$, (4.7) becomes:

$$\begin{aligned} p(\tilde{y}_{k+1} \,|\, \bar{l}_k, \bar{a}_k, y_k = 0) =& p^*(\tilde{y}_{k+1} \,|\, do(\bar{a}_k), y_k = 0) \cdot \\ & \prod_{j=0}^{k} c^*_{\tilde{Y}_{k+1}, L_j | \bar{L}_{j-1}, \bar{A}_k, Y_k = 0} \big\{ F(\tilde{y}_{k+1} \,|\, \bar{l}_{j-1}, do(\bar{a}_k), y_k = 0)), \\ & F(l_j \,|\, \bar{l}_{j-1}, do(\bar{a}_k), y_k = 0); \Theta_{k,j} \big\}. \end{aligned} \tag{4.11}$$

It is important to note the distinction between the subscripts j and k in the product. To reiterate, our goal is to simulate the outcome at the end of the kth

interval, which corresponds to the kth panel in Figure 4.2. Each copula in (4.11) associates $\tilde{y}_{k+1}$ with all the prior values of the time-varying confounder L_j, for $j = 0, \ldots, k$.

In the non-survival case discussed in Section 4.4.2, the ranks $F(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1})$, for $j = 0, \ldots, k$, are directly available at each simulation at step j and can be simply plugged into the iterative application of inverse h-functions without further computation. However, this convenience does not extend when the ranks are conditioned on survival at later time points. For instance, suppose that at $k = 0$, we simulate baseline health, $P(L_0) \sim \mathrm{N}(0, 1)$. Conditioning on survival after 12 months, however, alters the distribution. The baseline health of the survivors will almost certainly not be representative of the initial population, meaning that the distribution is no longer standard normal. This means that the conditional ranks of L_j that we need are not directly available and require additional computation. Fortunately, employing copulas to define the dependence structure offers a solution to deriving these conditional distributions.

Lemma 4.1. In the setting described above, suppose $Y_k = 0 \Leftrightarrow \tilde{Y}_k > 1$, then the conditional distribution $F_{L_j \mid \bar{L}_{j-1}, do(\bar{A}_k), Y_k=0}$ is given by

$$F(l_j \mid \bar{l}_{j-1}, do(\bar{a}_k), y_k = 0) = \frac{v_{k-1,j} - C_{\tilde{Y}_k L_j \mid \bar{L}_j, do(\bar{A}_k), Y_{k-1}=0}(v_{k-1,j}, q_{k-1,j}; \Theta_{k-1,j})}{1 - q_{k-1,j}}, \tag{4.12}$$

where $v_{k-1,j} = F(l_j \mid \bar{l}_{j-1}, do(\bar{a}_k), y_{k-1} = 0)$ and $q_{k-1,j} = F(\tilde{y}_k = 1 \mid \bar{l}_{j-1}, do(\bar{a}_k), y_{k-1} = 0)$.

The proof is given in Appendix D.2. Here we emphasize that $F_{L_j \mid \bar{L}_{j-1}, do(\bar{A}_k), Y_k=0}$ is the *renormalized* distribution in the population who have survived up to time-point k. This is distinct from simply conditioning on an individual's previous incremental failure time, which would give $F_{L_j \mid \bar{L}_{j-1}, do(\bar{A}_k), \tilde{Y}_k}$. The former corresponds to the expectation of the latter with respect to $\tilde{Y}_k$ given $\tilde{Y}_k > 1$.

The proposed simulation algorithm combines an iterative and recursive structure and is formally presented and validated using an inductive approach.

Proposed simulation method: At time step k, we specify the cumulative distribution functions:

$$F_{L_k \mid \bar{L}_{k-1}, \bar{A}_{k-1}, Y_k=0} \qquad F_{A_k \mid \bar{L}_k, \bar{A}_{k-1}, Y_k=0} \qquad F_{\tilde{Y}_{k+1} \mid do(\bar{A}_k), Y_k=0}.$$

Additionally, we define $k+1$ bivariate copulas $C_{\tilde{Y}_{k+1} L_j \mid \bar{L}_{j-1}, do(\bar{A}_k), Y_k=0}$, parameterized by $\Theta_{k,j}$ for $j = 0, \ldots, k$. These copulas capture the dependence between $\tilde{Y}_{k+1}$ and

each preceding time-varying covariate. Draw $w_{k,1}$, $w_{k,2}$, $w_{k,3} \sim \mathrm{U}(0,1)$ and compute:

$$l_k = F^{-1}_{L_k|\bar{L}_{k-1},\bar{A}_{k-1},Y_k=0}(w_{k,1} \,|\, \bar{l}_{k-1}, \bar{a}_{k-1}) \quad a_k = F^{-1}_{A_k|\bar{L}_k,\bar{A}_{k-1},Y_k=0}(w_{k,2} \,|\, \bar{l}_k, \bar{a}_{k-1}), \tag{4.13}$$

to obtain a sample (l_k, a_k). We initialize by setting $F(\tilde{y}_{k+1} \,|\, \bar{l}_k, do(\bar{a}_k), y_k = 0) = w_{k,3}$. Then, for $j = k-1, \ldots, 0$, we iteratively apply the transformations:

$$F(l_j \,|\, \bar{l}_{j-1}, do(\bar{a}_k), y_k = 0) = \frac{v_{k-1,j} - C_{\tilde{Y}_k L_j|\bar{L}_j, do(\bar{A}_k), Y_{k-1}=0}(v_{k-1,j}, q_{k-1,j}; \Theta_{k-1,j})}{1 - q_{k-1,j}}, \tag{4.14}$$

$$F(\tilde{y}_{k+1} \,|\, \bar{l}_j, do(\bar{a}_k), y_k = 0) = h^{-1}\big\{F(\tilde{y}_{k+1} \,|\, \bar{l}_{j+1}, do(\bar{a}_k), y_k = 0), \\ F(l_{j+1} \,|\, \bar{l}_j, do(\bar{a}_k), y_k = 0); \Theta_{k,j+1}\big\}. \tag{4.15}$$

Expression (4.14) follows Lemma 4.5.1, by induction, $v_{k-1,j}$ and $q_{k-1,j}$ are available from the previous time step $k-1$.

When the recursion reaches $j = 0$, we obtain the rank $F(\tilde{y}_{k+1} \,|\, l_0, do(\bar{a}_k), y_k = 0)$. Applying the h-function one final time, we set:

$$u^*_{k+1} = F(\tilde{y}_{k+1} \,|\, do(\bar{a}_k)) = h^{-1}\big\{F(\tilde{y}_{k+1} \,|\, l_0, do(\bar{a}_k), y_k = 0), \\ F(l_0 \,|\, do(\bar{a}_k), y_k = 0); \Theta_{k,0}\big\}.$$

A sample of $\tilde{y}_{k+1}$ is then obtained via inversion:

$$\tilde{y}_{k+1} = F^{-1}_{Y_{k+1}|do(\bar{A}_k),Y_k=0}(u^*_{k+1}),$$

with the corresponding binary outcome $y_{k+1} = I(\tilde{y}_{k+1} < 1)$. If $y_{k+1} = 1$, the simulation stops; otherwise, before proceeding to $k+1$, we store and compute the following quantities for the next step:

- Store $v_{k,j} = F(l_j \,|\, \bar{l}_{j-1}, do(\bar{a}_k), y_k = 0)$, for $j = 0, \ldots, k$, which have been computed in the current step via (4.14).
- Compute $q_{k,j} = F(\tilde{y}_{k+1} = 1 \,|\, \bar{l}_{j-1}, do(\bar{a}_k), y_k = 0)$: initialize $q_{k,0} = F(\tilde{y}_{k+1} = 1 \,|\, do(\bar{a}_k), y_k = 0)$, then for $j = 1, \ldots k$, recursively update:

$$\begin{aligned} q_{k,j} &= F(\tilde{y}_{k+1} = 1 \,|\, \bar{l}_{j-1}, do(\bar{a}_k), y_k = 0) \\ &= h\big\{F(\tilde{y}_{k+1} = 1 \,|\, \bar{l}_{j-2}, do(\bar{a}_k), y_k = 0), F(l_{j-1} \,|\, \bar{l}_{j-2}, do(\bar{a}_k), y_k = 0); \Theta_{k,j-1}\big\} \\ &= h\big\{q_{k,j-1}, v_{k,j-1}; \Theta_{k,j-1}\big\}. \end{aligned} \tag{4.16}$$

This completes the kth time step, yielding the sample $(l_k, a_k, \tilde{y}_{k+1}, y_{k+1})$, along with the quantities required for the next step. We formalize this claim as Proposition 4.5.1.

Algorithm 3 Simulation algorithm for an individual i up to the Kth follow-up

$k \leftarrow 0$
while $k \leq K$ **do**
  Sample $w_{k,1}, w_{k,2}, w_{k,3}$ independent standard uniform variables
  $l_k \leftarrow F^{-1}_{L_k|Z,\bar{L}_{k-1},\bar{A}_{k-1},Y_k=0}(w_{k,1} \,|\, z, \bar{l}_{k-1}, \bar{a}_{k-1})$
  $a_k \leftarrow F^{-1}_{A_k|Z,\bar{L}_k,\bar{A}_{k-1},Y_k=0}(w_{k,2} \,|\, z, \bar{l}_k, \bar{a}_{k-1})$
  if k = 0 **then**
    $v_{k,j} = w_{0,1}$
  else
    for $j \leftarrow 1, \ldots, k$ **do** ▷ L_j dist. in survivors per (4.14)
      $v_{k,j} \leftarrow \frac{v_{k-1,j} - C(v_{k-1,j}, q_{k-1,j}; \Theta_{k,k-j})}{1 - q_{k-1,j}}$
    end for
  end if
  $\xi_{k+1} \leftarrow w_{k,3}$
  for $j \leftarrow 0, \ldots, k$ **do** ▷ 'Unweaving' as per (4.15)
    $\xi_{k-j} \leftarrow h^{-1}(\xi_{k+1-j}, v_{k,k-j}; \Theta_{K-j})$
  end for
  $\tilde{y}_{k+1} \leftarrow F^{-1}_{\tilde{Y}_{k+1}|\bar{A}_k,Y_{k-1}=0}(\xi_0 \,|\, do(\bar{a}_k))$
  $y_{k+1} = I(\tilde{y}_{k+1} > 1)$
  if $y_{k+1} = 1$ **then**
    Stop
  else ▷ Prepare qs for the next time step per (4.16)
    $q_{k,0} \leftarrow F_{\tilde{Y}_{k+1}|\bar{A}_k,Y_{k-1}=0}(\tilde{Y}_{k+1} = 1 \,|\, do(\bar{a}_k))$
    for $j \leftarrow 1, \ldots k$ **do**
      $q_{k,j} \leftarrow h\{q_{k,j-1}, v_{k,j-1}; \Theta_{k,j-1}\}$
    end for
    $k \leftarrow k + 1$
  end if
end while
return $(\bar{l}_k, \bar{a}_k, \tilde{y}_{k+1}, \bar{y}_{k+1})$

Crucially, our method avoids computing the full conditional distribution of $\tilde{Y}_{k+1} \mid \bar{L}_k, do(\bar{A}_k), Y_k = 0$, but rather requiring just a rank corresponding to our simulated value; this is central to the efficiency of our approach. (4.14) involves k calculations of the distributions of each $L_j \mid \bar{L}_{j-1}$ among those who survived to time k, in which all the quantities, including ranks and copulas, are available from the previous step $k-1$. In (4.15), we recursively apply $k+1$ inverse h-functions to obtain the marginal rank of $F(\tilde{Y}_{k+1} \mid do(\bar{a}_k), Y_k = 0)$. Finally, we simulate $\tilde{Y}_{k+1}$ from the target marginal structural model via inversion. Before proceeding to the next time step $k+1$, we prepare the $q_{k,j}$ through $k+1$ recursive computations in (4.16). In total, this time step requires $3k+5$ functions, including the simulation of L_k and A_k.

We proceed to show that the proposed method works using mathematical induction.

Proposition 4.1. *Given the ranks $w_{k,1}, w_{k,2}, w_{k,3}$ for each $k = 0, \ldots, K-1$ and associated univariate distributions, together with copula parameters $\Theta_{k,j}$ for each $k = 0, \ldots, K$ and $j = 0, \ldots, k$, we can use the procedure outlined above to simulate a survival outcome according to the specified distributions.*

The proof is given in Appendix D.3.

In the derivation and proof presented above, we define the latent variable $\tilde{Y}$ as the incremental time to failure which naturally aligns with the Cox marginal structural model that we use as a working example. Importantly, a key strength of our approach lies in its generality: our proposed algorithm is fully agnostic to the choice of survival model and holds for any specification of the latent variable and its distribution. This is a fundamental result that enables the direct simulation of survival outcomes from any causal marginal structural model, whether continuous or discrete.

Remark 4.3. To illustrate this point, consider the case of simulating a binary survival outcome from a discrete marginal structural model, such as that defined in (4.10). Similar to the framework in Fan et al. (2017), we assume that there exists a continuous latent random variable, S_{k+1}, whose distribution under the intervention $do(\bar{A}_k)$ is $F_{S_{k+1}|do(\bar{A}_k),Y_k=0}$, such that the binary survival indicator is obtained by dichotomizing S_{k+1} at a threshold C. This formulation allows dependencies between Y_{k+1} and past covariates to be modelled via the latent variable S_{k+1}, and the simulation procedure follows the algorithm described above, replacing $\tilde{Y}_{k+1}$ by S_{k+1}. Given the marginal rank u_{k+1}, inversion yields $s_{k+1} = F^{-1}_{S_{k+1}|do(\bar{A}_k),Y_k=0}(u_{k+1} \mid do(\bar{a}_k), y_k = 0)$ Finally, the binary outcome is obtained as $y_{k+1} = I(s_{k+1} > C)$.

In this basic setting, simulating K time steps for each individual requires $\sum_{k=0}^{K-1}(3k+5) = 3K^2/2 + 7K/2$ calculations.This implies that the computational complexity scales quadratically with K.

If additional time-varying confounders—either observed or unobserved—are incorporated, such that $\boldsymbol{L}$ becomes a p-dimensional vector, the number of iterations needed of (4.14), (4.15) and (4.16) scales proportionally with p. As a result, the total number of computations becomes $\sum_{k=0}^{K-1}(3pk+2p+3) = K(3K+1)p/2+3K$. Thus, computational complexity grows linearly with the number of confounders p. In the example presented in Appendix D.5, we replicate the simulation study by Seaman and Keogh (2024), simulating data featuring two time-varying confounders.

However, it should be noted that if an individual experiences failure at any time step, the simulation terminates early. Hence, these stated complexities represent the maximum possible computational cost per individual. All of these calculations require evaluations of h-functions and their inverses and Copula distribution functions. For commonly used copulas, such evaluations are computationally efficient.

4.5.2 Illustration: Concomitant treatment

We illustrate how the proposed simulation framework can support sample size calculation under complex, real-world scenarios. To do so, we construct a synthetic study inspired by the SOUL trial (**S**emaglutide cardi**O**vascular o**U**tcomes tria**L**) (McGuire et al., 2023). The SOUL trial is a randomized controlled study assessing the effects of oral semaglutide, a glucagon-like peptide-1 receptor agonist (GLP-RA), on cardiovascular risk in individuals with type 2 diabetes, and cardiovascular and/or kidney diseases. The primary endpoint is time to first occurrence of major adverse cardiovascular events (MACE), analyzed under the intention-to-treat principle.

A design challenge in SOUL arises from the permitted use of concomitant medications such as SGLT2 inhibitors (SGLT2-i), and there is demonstrably more uptake among participants in the control arm. The uneven 'drop-in' treatment between the semaglutide and placebo arms may dilute the observed intention-to-treat effect, because both groups end up receiving some active treatment, blurring the distinction between them, and potentially undermining statistical power if unaccounted for during sample size calculation. Our goal is to show how the proposed simulation approach can capture such intercurrent events while preserving a predefined causal estimand.

We simulate data over seven half-yearly follow-up periods. Let $B \in \{0,1\}$ indicate baseline cardiovascular disease, L_k denote HbA1c at time k, $A \in \{0,1\}$ the randomized treatment (semaglutide), $S_k \in \{0,1\}$ SGLT2-i use at time k and Y_k

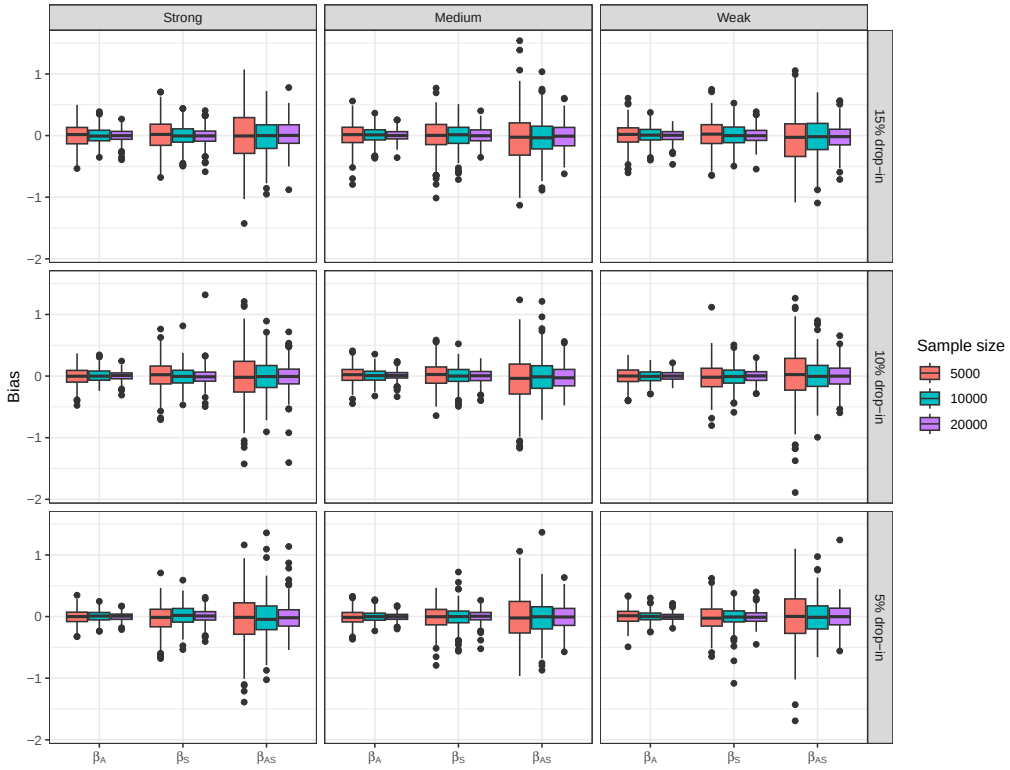


Figure 4.3: Boxplots of the bias for β_A, β_S and β_{AS} by inverse probability weighting (IPTW) across different drop-in scenarios.

the binary survival status at time k. We assume perfect adherence to A and model event times using a Cox marginal structural model:

$$\lambda(\tilde{y}_{k+1} \,|\, do(A, \bar{S}_k), Y_k = 0) = \lambda_0 \exp\left(\beta_A\, A + \beta_S\, S_k + \beta_{AS}\, AS_k\right), \tag{4.17}$$

where λ_0 is a baseline hazard assumed to be constant between follow-up visits and the survival indicator is calculated as $Y_{k+1} = I(\tilde{Y}_{k+1} < 1)$. Baseline characteristics and model parameters are loosely based on the design assumptions of the SOUL trial (McGuire et al., 2023), including an annual MACE rate of 3.5% and a hazard ratio of 0.83 for semaglutide (i.e. $\beta_A = \log(0.83)$). We construct 10 simulation scenarios by varying both the prevalence (0%, 5%, 10% and 15% annual drop-in) in the placebo group and efficacy (hazard ratios of $0.95, 0.9, 0.85$) of SGLT2-i use. These hazard ratios reflect estimates reported for SGLT2 inhibitors in McGuire et al. (2021). Full specifications of the data-generating model and scenario parameters are provided in Appendix D.4. For each scenario and trial sizes of $n = \{5{,}000, 10{,}000, 20{,}000\}$, we simulate 400 datasets and fit a Cox MSM using inverse probability of treatment weights with stabilized weights. Naïve unweighted estimators are also computed for comparison. For reference, the total randomized population in SOUL is 9,650.

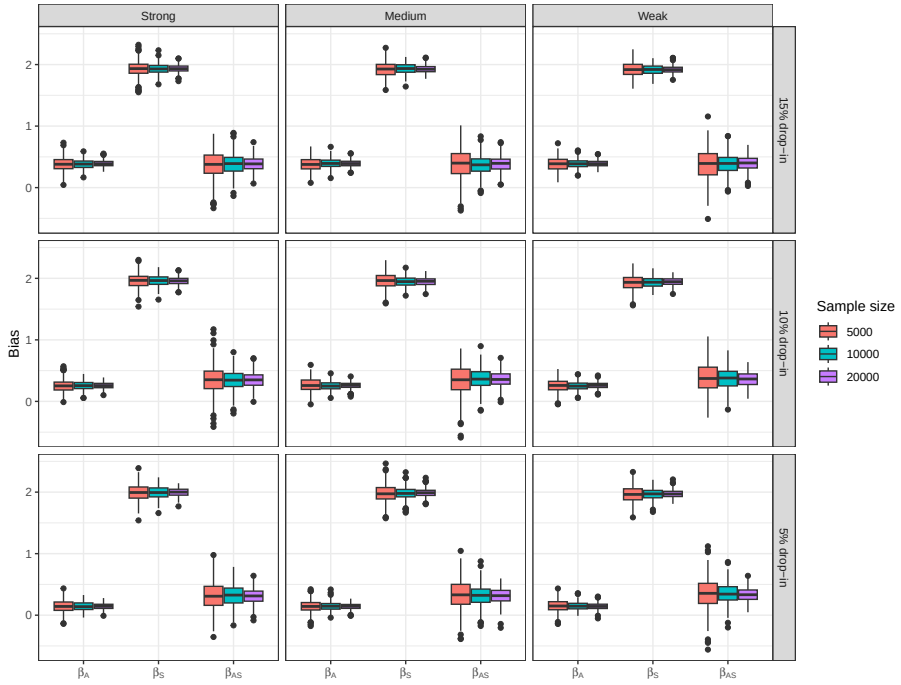


Figure 4.4: Boxplots of the bias for β_A, β_S and β_{AS} by unweighted Cox regression across different drop-in scenarios.

Figure 4.3 and Figure 4.4 display boxplots of the estimation biases for the coefficients β_A, β_S and β_{AS} by inverse-probability-weighted and unweighted Cox regression, respectively. Across all scenarios, the inverse-probability-weighted estimates exhibit biases that are centered at zero. As the sample size increases, the distribution of these biases becomes narrower and more tightly concentrated around zero, indicating consistency. In contrast, unweighted estimates show substantial and persistent bias across all scenarios, with no evidence of convergence toward zero as the sample size increases. Table 4.4 presents the empirical coverage: the inverse-probability-weighted method achieves coverage close to the nominal 95% level, whereas the unweighted approach yields empirical coverage well below the nominal level, reflecting the impact of the time-varying confounding introduced to the data.

These results confirm that the simulated data adhere to the marginal structural model specified in (4.17), and that time-varying confounding is induced. We now demonstrate how the proposed simulation framework can inform trial design, specifically for sample size determination under realistic scenarios involving intercurrent events.

The SOUL trial was designed to evaluate the effect of semaglutide on CV outcomes under the intention-to-treat principle, contrasting the outcomes of individuals randomized to semaglutide ($A = 1$) versus placebo ($A = 0$). Accordingly, we fit the

Effect of SGLT2-i	Prevalence in placebo	n	β_A		β_S		β_{AS}	
			IPTW	Unweighted	IPTW	Unweighted	IPTW	Unweighted
Strong effect	5% drop-in	5000	96.5%	65.0%	94.0%	0.0%	93.0%	74.0%
		10000	94.8%	47.3%	95.0%	0.0%	94.3%	49.0%
		20000	95.0%	18.5%	95.0%	0.0%	95.0%	25.8%
	10% drop-in	5000	94.0%	28.0%	95.0%	0.0%	92.3%	64.5%
		10000	95.8%	4.5%	96.5%	0.0%	95.3%	42.5%
		20000	96.5%	0.0%	95.8%	0.0%	96.3%	16.3%
	15% drop-in	5000	92.8%	4.8%	94.0%	0.0%	94.0%	60.8%
		10000	94.5%	0.0%	96.0%	0.0%	94.3%	34.0%
		20000	93.0%	0.0%	95.0%	0.0%	94.0%	7.5%
Moderate effect	5% drop-in	5000	95.3%	68.0%	96.5%	0.0%	95.5%	69.3%
		10000	94.5%	43.3%	93.8%	0.0%	94.3%	50.8%
		20000	95.3%	17.5%	96.0%	0.0%	93.8%	25.0%
	10% drop-in	5000	96.0%	25.5%	97.3%	0.0%	95.5%	63.8%
		10000	95.0%	4.5%	96.3%	0.0%	93.5%	40.8%
		20000	95.8%	0.3%	96.3%	0.0%	95.3%	14.8%
	15% drop-in	5000	94.3%	4.8%	95.5%	0.0%	95.8%	58.0%
		10000	95.3%	0.0%	94.5%	0.0%	94.0%	37.0%
		20000	92.0%	0.0%	95.3%	0.0%	94.0%	10.0%
Weak effect	5% drop-in	5000	96.3%	65.5%	95.3%	0.0%	93.5%	63.5%
		10000	96.0%	42.8%	97.0%	0.0%	95.3%	45.3%
		20000	95.0%	13.3%	95.0%	0.0%	95.0%	18.5%
	10% drop-in	5000	96.0%	26.8%	96.0%	0.0%	94.5%	59.8%
		10000	96.3%	5.0%	96.5%	0.0%	95.0%	37.0%
		20000	95.5%	0.0%	96.8%	0.0%	96.3%	11.5%
	15% drop-in	5000	93.3%	5.3%	95.8%	0.0%	93.8%	57.5%
		10000	93.3%	0.0%	93.5%	0.0%	92.5%	34.8%
		20000	95.5%	0.0%	95.5%	0.0%	95.0%	7.3%
	No drop-in	5000	96.0%	96.0%	–	–	–	–
		10000	94.5%	94.5%	–	–	–	–
		20000	93.5%	93.5%	–	–	–	–

Table 4.4: Empirical coverage rates of 95% confidence intervals of inverse-probability-weighted (IPTW) and unweighted Cox regression methods.

following Cox proportional hazards model:

$$\lambda(t \mid A) = \lambda_0 \exp(\beta_{ITT} A),$$

where t is time to the first occurrence of MACE and $\exp(\beta_{ITT})$ represents the hazard ratio associated with being assigned semaglutide, irrespective of SGLT2-i use. Table 4.5 reports the estimated intention-to-treat effects and empirical power based on a one-sided type I error rate of 0.025. In the hypothetical scenario with no SGLT2-i use in either arm, the estimated intention-to-treat effect closely approximates the true hazard ratio for semaglutide (0.83). Under this scenario, a trial with 10,000 participants achieves approximately 85% power. However, in all other scenarios, the intention-to-treat estimates are attenuated due to drop-in and consequently reduced power. To maintain at least 80% power, our simulations suggest that a sample size of 20,000 is sufficient but may be unnecessarily large and resource-intensive. Con-

versely, a trial with 5,000 participants is consistently underpowered, even in the absence of drop-in. A sample size of 10,000 appears to strike a reasonable balance: it achieves adequate power in scenarios with weak SGLT2-i effects, but falls short when the effect of SGLT2-i is strong or if the prevalence of drop-in in the placebo arm exceeds 5%. For the moderate scenarios, where SGLT2 inhibitors have a moderate effect, 10,000 participants is borderline sufficient, but if drop-in rates in the placebo arm exceed 5%, power may fall below acceptable thresholds. These findings suggest that 10,000 participants may be an appropriate design choice, but the sample size should be adjusted based on the plausibility of drop-in and effect size assumptions.

The 10 scenarios considered here are illustrative. The proposed simulation framework is highly flexible. For example, users can vary the marginal effect of semaglutide by adjusting β_A, alter the baseline incidence rate, or introduce more time-varying covariates. Because the algorithm relies on analytic function evaluations, it is computationally efficient: simulating 10,000 individuals over seven time steps requires approximately 40 seconds on a single core. This efficiency makes it practical and feasible to conduct large-scale simulation studies using parallel computation to support power analyses and inform robust trial design.

4.5.3 Competing risks

An important consideration in research concerning time-to-event outcomes is that of competing risks; these are events whose occurrence prevents the observation of the outcome of interest—for example, non-cardiovascular death in studies of MACE. Classical literature defines various statistical estimands for competing risk settings, including the marginal, cause-specific and subdistribution hazards (see, for example, Pintilie, 2007; Geskus, 2016; Young et al., 2020). These estimands differ in interpretation and causal relevance.

Our proposed simulation framework offers the flexibility to efficiently simulate data from marginal structural models that target the *controlled direct effect*:

$$F(\tilde{Y}_{k+1} \,|\, do(\bar{A}_k = \bar{a}_k, \bar{D}_k = \bar{0}_k), Z = z, Y_k = 0) = f(\bar{a}_k, z; \alpha), \tag{4.18}$$

where $\{D_k\}_{k \in \mathcal{K}_+}$ denote the sequence of competing events and $do(\bar{D}_k = \bar{0}_k)$ represents a hypothetical intervention eliminating these events. This aligns with the definition of the marginal hazard in Young et al. (2020). Under this framework, competing risks are essentially treated as equivalent to censoring. To generate the competing event process, two approaches can be considered. The first specifies a fully conditional distribution: $F(\tilde{D}_{k+1} \,|\, \bar{L}_k, \bar{A}_k, Z, Y_k = 0)$. Under this approach, the outcome process $Y_{1:K}$ is simulated according to Algorithm 3, targeting the causal

Effect of SGLT2-i	Prevalence in placebo	n	Estimate	Hazard Ratio	Power
Strong effect		5000	−0.164	0.849	46.5%
	5% drop-in	10000	−0.165	0.848	74.3%
		20000	−0.169	0.845	97.5%
		5000	−0.153	0.858	43.0%
	10% drop-in	10000	−0.148	0.862	65.5%
		20000	−0.151	0.860	93.0%
		5000	−0.143	0.867	37.5%
	15% drop-in	10000	−0.143	0.867	63.8%
		20000	−0.135	0.874	87.8%
Moderate effect		5000	−0.176	0.839	52.0%
	5% drop-in	10000	−0.173	0.841	81.0%
		20000	−0.179	0.836	98.8%
		5000	−0.160	0.852	45.0%
	10% drop-in	10000	−0.164	0.848	75.8%
		20000	−0.161	0.851	97.8%
		5000	−0.161	0.851	46.0%
	15% drop-in	10000	−0.157	0.855	76.3%
		20000	−0.154	0.857	94.5%
Weak effect		5000	−0.174	0.841	52.0%
	5% drop-in	10000	−0.181	0.834	84.5%
		20000	−0.184	0.832	98.3%
		5000	−0.176	0.838	53.3%
	10% drop-in	10000	−0.180	0.835	82.8%
		20000	−0.176	0.839	98.3%
		5000	−0.182	0.834	55.0%
	15% drop-in	10000	−0.176	0.838	82.3%
		20000	−0.173	0.841	98.3%
		5000	−0.187	0.830	56.8%
	No drop-in	10000	−0.183	0.833	85.0%
		20000	−0.187	0.830	99.5%

Table 4.5: Under the intention-to-treat principal, estimates of β_{ITT}, hazard ratios ($\exp \beta_{ITT}$), and power using a one-sided type I error rate of 0.025 for various scenarios and sample sizes (n).

margin defined in (4.18). Once simulation of the failure process is complete—suppose the trajectory terminates at time step K' when $Y_{K'+1} = 1$—the competing event process is then simulated for each $k \in \{0, \ldots, K'\}$. At each of these time steps, $\tilde{D}_{k+1}$ is drawn from the specified conditional distribution $F(\tilde{D}_{k+1} \mid \bar{L}_k, \bar{A}_k, Z, Y_k = 0)$. If a competing event occurs before time K', the individual's trajectory is truncated at the corresponding time step. Censoring events, such as loss to follow-up, can also be simulated using this approach.

Alternatively, a marginal structural model may be specified for the competing event hazard, mirroring the approach used for the primary outcome:

$$F(\tilde{D}_{k+1} \mid do(\bar{A}_k = \bar{a}_k, \bar{Y}_k = \bar{0}_k), Z = z, D_k = 0) = g(\bar{a}_k, z; \beta). \tag{4.19}$$

Additionally, we impose the following conditional independence:

$$\tilde{D}_{k+1} \perp\!\!\!\perp \tilde{Y}_{k+1} \,|\, do(\bar{A}_k, \bar{Y}_k = \bar{0}_k, \bar{D}_k = \bar{0}_k), Z, \bar{L}_k,$$

which states that, under intervention, and conditional on baseline and time-varying covariates, $\tilde{D}_{k+1}$ and $\tilde{Y}_{k+1}$ are independent. Graphically, this corresponds to the absence of a direct edge between Y_{k+1} and D_{k+1} in the underlying causal graph—an edge that would not be identifiable from observed data in any case.

If both the outcome and the competing event are simulated using MSMs, the simulation proceeds as follows:

1. Simulate baseline covariates Z. Set $Y_0 = D_0 = 0$.
2. For time point $k = 0, \ldots, K$:
   (a) If $Y_k = 0$ or $D_k = 0$, simulate L_k, A_k; otherwise, stop;
   (b) If $Y_k = 0$, simulate Y_{k+1} following Algorithm 3; otherwise set Y_{k+1} to null;
   (c) If $D_k = 0$, simulate D_{k+1} following Algorithm 3, substituting Y with D, along with the corresponding distributions and copulas; otherwise, set D_{k+1} to null.
3. Truncate this individual's trajectory at the first occurrence of either the outcome of interest or the competing event.

Notably, in step (b), the decision to simulate Y_{k+1} depends solely on Y_k irrespective of D_k and vice versa for step (c). This reflects the fact that we target a *controlled direct effect*, as specified in (4.18) and (4.19), where the other event is conceptually eliminated through intervention.

This approach is particularly helpful in settings like the estrogen therapy example in Young et al. (2020). By explicitly specifying the direction and magnitude of the direct effects of treatment on both prostate cancer mortality and other-cause mortality, one can simulate data to understand how these effects jointly determine the total effect, i.e. the cause-specific hazard. This framework can inform trial design: clinicians can input plausible direct effect assumptions and compute the sample size required to achieve adequate power for detecting a treatment effect on the cause-specific hazard.

4.6 Discussion

In this paper, we proposed and demonstrated a principled strategy for simulating longitudinal data exactly from marginal structural models, leveraging the frugal parameterization (Evans and Didelez, 2024). Our proposed method is highly flexible and can accommodate a variety of practical scenarios, including multiple treatments (for example, primary and drop-in), dynamic treatment regimes, dose-response analyses, multiple outcomes (such as secondary outcomes and competing risks), and censoring. Additionally, it imposes no restrictions on the data-generating process or marginal structural causal models. In Section 4.5, we extended this method to survival models and time-to-event outcomes, which are common in clinical trials. The method's computational efficiency, combined with its analytic precision, makes it well-suited to practical tasks such as simulation-based method evaluation and sample size determination.

In Section 4.5.3, we discussed how our proposed simulation framework can flexibly generate data with competing risks from marginal structural models targeting the marginal hazard. However, extending the framework to simulate data under alternative estimands—specifically, the cause-specific and subdistribution hazards, which represent total effects—remains nontrivial. Future work may explore how to extend the proposed method to support simulation from marginal structural models that target these alternative causal quantities.

A limitation of the method is difficulty in enforcing Markov-type conditional independences between the outcome and the history of time-varying confounders. For instance, the pair-copula construction we employed in Section 4.4.2 to model the association between Y and $L_0, \ldots, L_k$ makes it difficult to impose restrictions that might be considered natural; for example, that the distribution of the outcome only depends upon the most recent few values of a time-varying covariate. Parameterizing such a model is infeasible unless all the bivariate copulas in the pair-copula construction are Gaussian; in this case, conditional independence can be enforced by appropriately choosing the partial correlation parameters. In practice, it is far easier to justify the presence of a dependency than to convincingly argue for its absence. If enforcing such Markov conditions is absolutely necessary, we would suggest using either this Gaussian approximation or numerical integration. However, numerical integration in high-dimensions is well known to be infeasible.

Chapter 5

Conclusion

5.1 Summary

This thesis addresses three questions that arose in real-world applications of causal inference:

> *(i) How can real-world data (RWD) be used to augment an underpowered randomized controlled trial (RCT)?*
>
> *(ii) Given the number of available data fusion methods, how can one determine which to use in practice?*
>
> *(iii) Since simulation is a powerful tool, how can we simulate longitudinal data from a causal model in a way that preserves its marginal structure?*

These questions did not arise in abstraction but were motivated by collaborations with practitioners and discussions with colleagues in applied settings. The opportunity to develop methods grounded in real-world needs—rather than hypothetical scenarios—has shaped the direction and purpose of this research. In retrospect, this experience underscores the value of methods that are not only statistically principled but also practically usable.

Chapter 2 introduces a power-likelihood approach for augmenting randomized controlled trials with observational data. Building on the framework of power priors (Ibrahim and Chen, 2000), the proposed approach regulates the contribution of potentially biased external data by selecting a learning factor that maximizes expected posterior predictive accuracy. This *empirical Bayes*-like procedure avoids the need for arbitrary a priori specification of hyperparameters. Through simulations and application to the PIONEER 6 cardiovascular trial, the method demonstrates increased statistical power without compromising inferential validity, especially in scenarios where trial data alone are underpowered.

Chapter 3 presents a systematic evaluation of contemporary data fusion methods aimed at efficiency gains. Recognizing the proliferation of techniques—each with distinct assumptions and trade-offs—the work undertakes extensive simulation studies to compare their empirical performance under realistic conditions. Key insights emerge about the trade-off between efficiency and bias, the sensitivity of methods to violations of RWD ignorability and transportability, and the role of design features such as negative control outcomes and shrinkage. This comparative analysis informs practical method selection and contributes a principled understanding of how different approaches behave in finite samples.

Chapter 4 addresses a gap in the simulation of longitudinal data for causal inference. While marginal structural models are widely used to estimate effects in the presence of time-varying confounding, simulating data from these models poses fundamental difficulties. Existing approaches often rely on approximations, impose restrictive assumptions, or suffer from computational inefficiency. In response, we propose a new simulation algorithm that enables exact generation of longitudinal data from a prespecified MSM. The method leverages frugal parameterization and copula constructions to preserve both marginal and joint structures. It is applicable to complex settings involving multiple time-varying covariates, dynamic treatment regimes, and survival outcomes. The proposed algorithm facilitates transparent benchmarking of causal methods and supports realistic power calculations for longitudinal studies.

Together, these contributions form a coherent research programme that seeks to bridge methodological rigour with real-world relevance. Each chapter advances a core objective of causal inference: *to obtain reliable and actionable knowledge from imperfect data*.

5.2 Future work

5.3 A many-data power-likelihood approach

A natural extension of the work in Chapter 2 is to generalize the power-likelihood framework to incorporate multiple external datasets. The current formulation assumes a single observational source and uses a scalar learning rate $\eta \in [0, 1]$ to regulate its influence. However, when several datasets, or distinct subgroups within a single dataset, are available, it is desirable to assign a separate learning rate to each. This leads to a vector-valued $\eta \in \mathbb{R}^K$, allowing differential borrowing across K sources or strata.

In the single-source setting, η is selected via a grid search in $[0, 1]$ that maximizes the expected log predictive density (ELPD). While effective in one dimension, grid search becomes computationally infeasible as dimensionality increases. To address this, we derived an analytic approximation of the ELPD (Proposition 2.2, Chapter 2). This approximation expresses the ELPD as a sum of a constant and a function of η:

$$f(\eta) = (\hat{\phi}_\eta - \hat{\phi}_r)^\top \mathbb{E}(\boldsymbol{A})(\hat{\phi}_\eta - \hat{\phi}_r) + \operatorname{tr}(\mathbb{E}(\boldsymbol{A})\Sigma_\eta). \tag{5.1}$$

where $\hat{\phi}_\eta$ is the η-powered posterior mean, Σ_η its covariance, and $\boldsymbol{A}$ the expected curvature of the likelihood. Optimizing this function provides a scalable alternative to grid search and supports gradient-based methods.

This result can be extended to higher dimensions, where $\boldsymbol{\eta}$ is vector-valued. While the theoretical form of $f(\boldsymbol{\eta})$ appears to generalize naturally, formalizing this extension requires further work to verify regularity conditions and establish asymptotic properties. Nonetheless, the approach is promising in several practical settings:

- **Subgroup-specific learning**: Certain subpopulations in the external data may be more comparable to the trial population. For example, in the PIONEER 6 application, one might hypothesize that insulin users in the observational data are more similar to trial participants than non-users—though this is offered purely as an illustrative example and not a verified empirical claim. Assigning separate η values to different subgroups allows selective borrowing, potentially improving both precision and robustness.
- **Multiple external sources**: When integrating information from several datasets (e.g. historical trials or linked registries), assigning distinct learning rates allows calibrated, source-specific contributions to the posterior.

This many-data extension of the power-likelihood framework represents a natural and practically useful generalization for real-world applications.

5.4 Simulating longitudinal survival data with competing events from MSMs targeting the cause-specific hazard

In Section 4.5.3 in Chapter 4, we outlined how our simulation framework can be adapted to simulate longitudinal survival data with competing risks, targeting the *controlled direct effect*:

$$F(\tilde{Y}_{k+1} \,|\, do(\bar{A}_k = \bar{a}_k, \bar{D}_k = \bar{0}_k), Z = z, Y_k = 0) = f(\bar{a}_k, z; \alpha), \tag{5.2}$$

where $do(\bar{D}_k = \bar{0}_k)$ represents that the competing events are *eliminated* by intervention. One may argue that the "eliminating competing event" element of the marginal hazard is unrealistic, and other causal quantities should be considered.

A more interpretable estimand in many settings is the *cause-specific hazard*, which conditions on survival from both the event of interest and the competing event, rather than intervening on the latter. Formally, the estimand corresponds to:

$$F(\tilde{Y}_{k+1} \mid do(\bar{A}_k = \bar{a}_k), Z = z, Y_k = 0, D_k = 0) = g(\bar{a}_k, z; \beta). \tag{5.3}$$

The main technical challenge in simulating from this model lies in computing conditional ranks in the subset of survivors from both events:

$$F(l_j \mid \bar{l}_{k-1}, do(\bar{a}_k), y_k = 0, d_k = 0).$$

We anticipate that a derivation analogous to Lemma 4.5.1 can be developed using the properties of pair-copula constructions. While non-trivial, this extension is well-scoped and technically tractable, and could be implemented within the existing simulation framework.

5.5 Realistic emulation of trial data for power and sample size calculations

In Section 4.5.2 of Chapter 4, we presented a case study inspired by the SOUL trial to illustrate how our simulation method can support power and sample size calculations in the presence of concomitant drug use. However, this example was intended as a proof of concept rather than a full emulation of real-world trial data.

A valuable direction for future work is to refine and extend the simulation algorithm to more faithfully reproduce the data-generating processes of specific trials. This may involve:

- Incorporating additional baseline and time-varying covariates, with distributions calibrated to empirical data and published studies;
- Explicitly modelling treatment adherence and other treatment dynamics; and
- Accounting for competing events and, where relevant, time-dependent frailty.

Such a study would be carried out in close collaboration with clinical, regulatory, and epidemiological partners, with the aim of producing actionable guidance for trial design. Each modelling assumption—such as treatment assignment mechanisms, covariate distributions, adherence behaviours, and outcome processes—would need

to be carefully justified using clinical evidence, real-world data, and input from subject-matter experts. The resulting work could form the basis of an applied paper aimed at trialists and epidemiologists, demonstrating the utility of MSM-based simulation for real-world decision-making.

These future directions are at varying stages of development. At the time of submission, some are under active investigation, while others are being scoped for collaborative projects.

Appendix A

Philosophical foundations of causality

Causality lies at the heart of human inquiry. From ancient farmers observing the links between seasons and harvests to modern scientists establishing causal effects in randomized controlled trials (RCTs), the pursuit of causal understanding has driven intellectual progress across millennia. Yet despite its fundamental importance, causality remains a profoundly contested concept, debated by philosophers for centuries. The philosophical study of causality traces back to Aristotle, whose conceptual framework laid the groundwork for centuries of debate. How has causality been understood across different eras, and how do these historical perspectives inform contemporary methodologies?

> *Knowledge is the object of our inquiry, and men do not think they know a thing till they have grasped the 'why' of it (which is to grasp its primary cause).*
>
> —Aristotle, *Physics* II.3 (194b18-20)

As Aristotle asserts in *Physics* II.3 and *Metaphysics* V.2, knowledge requires grasping the *why* of phenomena. His theory of causal pluralism identifies four distinct types of causes: material, formal, efficient, and final. Of particular relevance to modern causal analysis is the *efficient cause*, which concerns the power or mechanism that brings about change, answering the question, "Where does change (or motion) come from?"

By the 18th century, David Hume challenged Aristotelian metaphysics with a radical empiricist critique. In *A Treatise of Human Nature* (1739) and *An Enquiry Concerning Human Understanding* (1748), Hume argued that causality is neither an inherent feature of nature nor a metaphysical necessity but rather a cognitive *habit* arising from the repeated observation of event B following event A (constant

conjunction). He dismissed Aristotle's notion of an *efficient cause* as unobservable, reducing causation to a psychological *custom*:

> *We may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object had not been, the second never had existed.*
>
> —Hume, *An Enquiry Concerning Human Understanding*, (Hume, 1748)

Immanuel Kant sought to reconcile Hume's skepticism by asserting that causality is not merely an empirical habit but an a priori category of the understanding. In *Critique of Pure Reason* (Kant, 1781), Kant posited that causal relationships are necessary structures through which the mind synthesizes sensory data into coherent experience. For Kant, causality is not derived from observation but is a precondition for intelligible perception.

In the 20th century, building on the implicit counterfactual reasoning in Hume's definition, David Lewis (Lewis, 1973) formalized the counterfactual theory of causation, framing causality in terms of difference-making:

> *We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects – some of them, at least, and usually all – would have been absent as well.*
>
> —Lewis, *Causation.* (Lewis, 1973)

These philosophical discourses on causality——ranging from Aristotle's causal pluralism to Hume's empiricist skepticism, and Lewis's counterfactual theory——reveals a fundamental tension: is causality an intrinsic feature of the world, a cognitive construct, or a mere statistical association? These debates have profoundly influenced the statistical perspective on causality, with modern causal inference adopting a principled synthesis of these viewpoints.

Appendix B

Appendix for Chapter 2: Combining experimental and observational data through a power likelihood

Contents

B.1 Proof of Lemma 2.1 126
B.2 Proof of Theorem 2.1 127
B.3 Proof of Proposition 2.2 128

B.1 Proof of Lemma 2.1

Proof. Following the setup outlined in Section 2.3, we have an experimental dataset of size n_r, $X = (X_1, \ldots, X_{n_r})$ and an observational dataset of size n_o, $Y = (Y_1, \ldots, Y_{n_o})$. We assume that they share the same causal parameters ϕ^*.

We write the log-likelihoods as the following estimation equations:

$$\sum_{i=1}^{n_r} f(\phi; X_i) = 0, \tag{B.1}$$

$$\sum_{i=1}^{n_o} g(\phi; Y_i) = 0, \tag{B.2}$$

$$\sum_{i=1}^{n_r} f(\phi; X_i) + \eta \sum_{i=1}^{n_o} g(\phi; Y_i) = 0, \tag{B.3}$$

where $\mathbb{E}f(\phi^*; X_i) = 0$ and $\mathbb{E}g(\phi^* + \delta; Y_i) = 0$ for some unknown bias δ. From the standard theory of estimation equations, we know that the solution to (B.1), $\hat{\phi}_r$, is

asymptotically normal following $\sqrt{n_r}(\hat{\phi}_r - \phi^*) \to N(0, \boldsymbol{V})$, and likewise, $\sqrt{n_o}(\hat{\phi}_o - \phi^* - \delta) \to N(0, \boldsymbol{W})$.

Let the solution to (B.3) be $\hat{\phi}_\eta$. We can approximate each sum in (B.3) with its Taylor expansion and use the fact that $\sum_{i=1}^{n_r} f(\hat{\phi}_r; X_i) = 0$ and $\sum_{i=1}^{n_o} g(\hat{\phi}_o; X_i) = 0$, then

$$
\begin{aligned}
n_r \boldsymbol{V}^{-1}(\hat{\phi}_\eta - \hat{\phi}_r) + \eta n_o \boldsymbol{W}^{-1}(\hat{\phi}_\eta - \hat{\phi}_o) &= o_p(n_r + \eta n_o) \\
\left(n_r \boldsymbol{V}^{-1} + \eta n_o \boldsymbol{W}^{-1}\right) \hat{\phi}_\eta &= n_r \boldsymbol{V}^{-1} \hat{\phi}_r + \eta n_o \boldsymbol{W}^{-1} \hat{\phi}_o + o_p(n_r + \eta n_o).
\end{aligned}
$$

Then we have an expression for $\hat{\phi}_\eta$ as

$$
\hat{\phi}_\eta = \left(n_r \boldsymbol{V}^{-1} + \eta n_o \boldsymbol{W}^{-1}\right)^{-1} \left(n_r \boldsymbol{V}^{-1} \hat{\phi}_r + \eta n_o \boldsymbol{W}^{-1} \hat{\phi}_o\right) + o_p(1).
$$

Let $\mathcal{I}(\hat{\phi}_r)$ and $\mathcal{I}(\hat{\phi}_o)$ be the Fisher information matrices (FIMs) for the experimental and observational data respectively, then the combined FIM is

$$
\begin{aligned}
\mathcal{I}(\hat{\phi}_\eta) &= -\mathbb{E}\left[\frac{\partial}{\partial \phi}\left(\sum_{i=1}^{n_r} f(\phi; X_i) + \eta \sum_{i=1}^{n_o} g(\phi; Y_i)\right)\right] \\
&= -\mathbb{E}\left[\frac{\partial}{\partial \phi} \sum_{i=1}^{n_r} f(\phi; X_i)\right] - \eta\, \mathbb{E}\left[\frac{\partial}{\partial \phi} \sum_{i=1}^{n_o} g(\phi; Y_i)\right] \\
&= \mathcal{I}(\hat{\phi}_r) + \eta\, \mathcal{I}(\hat{\phi}_o) \\
&= n_r \boldsymbol{V}^{-1} + \eta n_o \boldsymbol{W}^{-1}.
\end{aligned}
$$

That is, the combined Fisher information is a weighted sum of the two datasets. Therefore, we have, asymptotically

$$
\begin{aligned}
&\sqrt{n_r + \eta n_o}\left\{\hat{\phi}_\eta - \left(n_r \boldsymbol{V}^{-1} + \eta n_o \boldsymbol{W}^{-1}\right)^{-1} \left(n_r \boldsymbol{V}^{-1} \hat{\phi}_r + \eta n_o \boldsymbol{W}^{-1} \hat{\phi}_o\right)\right\} \\
&\qquad \xrightarrow{d} \mathcal{N}\left(0,\, \left(\frac{n_r}{n_r + \eta n_o} \boldsymbol{V}^{-1} + \frac{\eta n_o}{n_r + \eta n_o} \boldsymbol{W}^{-1}\right)^{-1}\right).
\end{aligned}
$$

□

B.2 Proof of Theorem 2.1

Theorem 1.1 (Consistency). Let $\delta = \delta(n_o) = \delta^*/n_o^k$ denote the bias in $\mathcal{D}_o$.

(a) If $0 \le k < 1/2$ and $\delta^* \neq 0$, then $\hat{\eta}$ converges to 0, and $\hat{\phi}_{\hat{\eta}} = \phi^* + o_p(n^{-1/2})$.

(b) If $k \ge 1/2$ or $\delta^* = 0$ then $\hat{\eta}$ does not converge.

Proof. From (B.5) in the previous proof, we observe that for sufficiently large data samples, the posterior predictive distribution does not depend upon the prior distribution that we choose. This observation is also noted in Dunsmore (1976).

As discussed in Remark 2, ELPD is a proper scoring rule, so it will maximize the expected score at the distribution that is closest to the truth. That is, as sample size tends to infinity, the value of ϕ that maximizes ELPD converges to ϕ^*.

From Lemma 2.1 we have that

$$\begin{aligned}\hat{\phi}_\eta &= \left(n_r V^{-1} + \eta n_o W^{-1}\right)^{-1} \left(n_r V^{-1}\hat{\phi}_r + \eta n_o W^{-1}\hat{\phi}_o\right) \\ &= \left(n_r V^{-1} + \eta n_o W^{-1}\right)^{-1} \left\{n_r V^{-1}\left(\phi^* + o_p(n^{-1/2})\right) + \eta n_o W^{-1}\left(\phi^* + \delta + o_p(n^{-1/2})\right)\right\} \\ &= \phi^* + \left(V^{-1} + \frac{\eta}{\rho}W^{-1}\right)^{-1} W^{-1}\frac{\eta}{\rho}\delta^* n^{-k} + o_p(n^{-1/2}),\end{aligned}$$

where $\rho = n_r/n_o$ and is assumed to be constant. In scenario (a), $k < \frac{1}{2}$ and we have that

$$\hat{\phi}_\eta = \phi^* + \left(V^{-1} + \frac{\eta}{\rho}W^{-1}\right)^{-1} W^{-1}\frac{\eta}{\rho}\delta^* n^{-k} + o_p(n^{-1/2}).$$

Then, since ELPD is a proper scoring rule, and the bias is statistically non-negligible, the optimal η will asymptotically be zero.

In scenario (b), $k \geq \frac{1}{2}$ or $\delta^* = 0$ and we have that

$$\hat{\phi}_\eta = \phi^* + \left(V^{-1} + \frac{\eta}{\rho}W^{-1}\right)^{-1} W^{-1}\frac{\eta}{\rho}\delta^* n^{-k} + o_p(n^{-1/2}) = \phi^* + o_p(n^{-1/2}),$$

and η has no asymptotically and statistically relevant effect on the ELPD at all. Hence we would not expect η to converge. □

B.3 Proof of Proposition 2.2

Proof. If we expand the predictive density function, $p(\tilde{x}\,|\,x)$, about $\hat{\phi}_r$, which is the MLE estimate of the experimental data, by Taylor's theorem, ignoring higher order terms, we have

$$\begin{aligned}p(\tilde{x}\,|\,x) &= \int_\Phi p(\tilde{x}\,|\,\phi)p_\eta(\phi\,|\,x_e, x_o)d\phi \\ &= \int_\Phi \left[p(\tilde{x}\,|\,\hat{\phi}_r) + \left(\phi - \hat{\phi}_r\right)^\intercal \frac{\partial p(\tilde{x}\,|\,\phi)}{\partial\phi}\Big|_{\phi=\hat{\phi}_r} + \frac{1}{2}\left(\phi - \hat{\phi}_r\right)^\intercal \frac{\partial^2 p(\tilde{x}\,|\,\phi)}{\partial\phi^2}\Big|_{\phi=\hat{\phi}_r}\left(\phi - \hat{\phi}_r\right)\right. \\ &\qquad \left. + \ldots\right] p_\eta(\phi\,|\,x_e, x_o)\,d\phi \end{aligned} \tag{B.4}$$

$$= p(\tilde{x} \mid \hat{\phi}_r) + \frac{1}{2} \int_{\Phi} \left(\phi - \hat{\phi}_r\right)^{\intercal} \frac{\partial^2 p(\tilde{x} \mid \phi)}{\partial \phi^2}\Big|_{\phi=\hat{\phi}_r} \left(\phi - \hat{\phi}_r\right) p_\eta(\phi \mid x_e, x_o) d\phi + \dots. \tag{B.5}$$

Note that $\tilde{x}$ represents a new data point from the 'true' distribution, which is regarded as equivalent to the experimental data, and $p_\eta(\phi \mid x_e, x_o)$ is the η-powered posterior distribution of ϕ.

Following Lemma 2.1, we can approximate the posterior $p_\eta(\phi \mid x_e, x_o)$ by a normal distribution:

$$p_\eta(\phi \mid x_e, x_o) \sim N\left(\hat{\phi}_\eta, \boldsymbol{\Sigma}_\eta\right),$$

where

$$\hat{\phi}_\eta = \left(n_r \boldsymbol{V}^{-1} + \eta n_o \boldsymbol{W}^{-1}\right)^{-1} \left(n_r \boldsymbol{V}^{-1} \hat{\phi}_r + \eta n_o \boldsymbol{W}^{-1} \hat{\phi}_o\right), \tag{B.6}$$

$$\boldsymbol{\Sigma}_\eta = \left(n_r \boldsymbol{V}^{-1} + \eta n_o \boldsymbol{V}^{-1}\right)^{-1}. \tag{B.7}$$

Let matrix $\boldsymbol{A} = \frac{\partial^2 p(\tilde{x} \mid \phi)}{\partial \phi^2}\Big|_{\phi=\hat{\phi}_r}$ then, the vector $\boldsymbol{A}^{1/2}\left(\phi - \hat{\phi}_r\right)$ is a normal vector following

$$\boldsymbol{A}^{1/2}\left(\phi - \hat{\phi}_r\right) \sim N\left(\boldsymbol{A}^{1/2}\left(\hat{\phi}_\eta - \hat{\phi}_r\right), \boldsymbol{A}^{1/2} \boldsymbol{\Sigma}_\eta \boldsymbol{A}^{1/2}\right).$$

Therefore, the integral in (B.5) becomes

$$\begin{aligned} &\int_{\Phi} \left(\phi - \hat{\phi}_r\right)^{\intercal} \boldsymbol{A} \left(\phi - \hat{\phi}_r\right) p_\eta(\phi \mid x_e, x_o) d\phi \\ &= \left(\hat{\phi}_\eta - \hat{\phi}_r\right)^{\intercal} \boldsymbol{A} \left(\hat{\phi}_\eta - \hat{\phi}_r\right) + \operatorname{tr}(\boldsymbol{A}\boldsymbol{\Sigma}_\eta). \end{aligned} \tag{B.8}$$

Substituting (B.8) back to (B.4) yields

$$p(\tilde{x} \mid x) = p(\tilde{x} \mid \hat{\phi}_r) + \left(\hat{\phi}_\eta - \hat{\phi}_r\right)^{\intercal} \boldsymbol{A} \left(\hat{\phi}_\eta - \hat{\phi}_r\right) + \operatorname{tr}(\boldsymbol{A}\Sigma_\eta) + \dots.$$

By definition, we can write the ELPD at a given η as

$$\mathbb{E}_\eta\left[\log p(\tilde{x} \mid x)\right] = \mathbb{E}\left[\log\left(p(\tilde{x} \mid \hat{\phi}_r) + \left(\hat{\phi}_\eta - \hat{\phi}_r\right)^{\intercal} \boldsymbol{A} \left(\hat{\phi}_\eta - \hat{\phi}_r\right) + \operatorname{tr}(\boldsymbol{A}\boldsymbol{\Sigma}_\eta) + \dots\right)\right].$$

Therefore we can see that maximizing ELPD is equivalent to maximizing a function of η:

$$\begin{aligned} f(\eta) &= \mathbb{E}\left[\left(\hat{\phi}_\eta - \hat{\phi}_r\right)^{\intercal} \boldsymbol{A} \left(\hat{\phi}_\eta - \hat{\phi}_r\right) + \operatorname{tr}(\boldsymbol{A}\Sigma_\eta)\right] \\ &= \left(\hat{\phi}_\eta - \hat{\phi}_r\right)^{\intercal} \mathbb{E}(\boldsymbol{A}) \left(\hat{\phi}_\eta - \hat{\phi}_r\right) + \operatorname{tr}(\mathbb{E}(\boldsymbol{A})\Sigma_\eta), \end{aligned} \tag{B.9}$$

and the optimal η^*:

$$\eta^* = \arg\max_\eta f(\eta) = \arg\max_\eta \left[\left(\hat{\phi}_\eta - \hat{\phi}_r\right)^{\intercal} \mathbb{E}(\boldsymbol{A}) \left(\hat{\phi}_\eta - \hat{\phi}_r\right) + \operatorname{tr}(\mathbb{E}(\boldsymbol{A})\Sigma_\eta)\right].$$

Substituting (B.6) and (B.7) into (B.9), the first term becomes

$$\left(\hat{\phi}_o - \hat{\phi}_r\right)^\intercal B^\intercal \, \mathbb{E}(\boldsymbol{A}) \, B \left(\hat{\phi}_o - \hat{\phi}_r\right)$$

where $B = \left(n_r V^{-1} + \eta n_o W^{-1}\right)^{-1} \eta n_o W^{-1}$, and the second term equals

$$\operatorname{tr}\left(\mathbb{E}(\boldsymbol{A}) \left(n_r V^{-1} + \eta n_o W^{-1}\right)^{-1}\right).$$

We observe that all the η terms contain n_o as a factor. This implies that, all else being equal, the optimal η^* scales with n_o. Intuitively, as the size of the observational data doubles, the optimal proportion of information drawn from it halves. □

Appendix C

Appendix for Chapter 3: Data fusion for efficient average treatment effect estimation

Contents

C.1 Implementation details of methods in simulation studies 131
C.1.1 Baseline IPTW estimators 131
C.1.2 Shrinkage estimators 132
C.1.3 Bias-correction methods 133
C.1.4 Power-likelihood method 134
C.1.5 Test-then-pool methods 135
C.1.6 Weighted combination methods 135
C.1.7 Prognostic adjustment 136
C.2 Details of real-world application 136
C.2.1 Target trial emulation 136
C.2.2 Implementation of methods 148

C.1 Implementation details of methods in simulation studies

In this appendix, we provide a detailed description of the implementation of the data fusion methods used in our simulation studies presented in Section 3.3.

C.1.1 Baseline IPTW estimators

We use AIPW estimators $\hat{\tau}_r$ and $\hat{\tau}_o$ as reference. The AIPW estimator involves a propensity model and an outcome model. For both, we employ a discrete `SuperLearner`

(van der Laan et al., 2007) ensemble with a library consisting of a mean model and generalized linear models (GLM) (via R package `speedglm` (Enea et al., 2009)). Given the simple generative models for A and Y, this library should suffice. For simulations in Section 3.3.3 where a quadratic term is added, we also include generalized additive models (Hastie and Tibshirani, 1987) and random forest (via R package `ranger` (Wright and Ziegler, 2017)) into the library to allow for non-linear flexibility. Additionally, we implemented a bound of 0.025 on the propensity score to avoid extreme weights.

We use an empirical sandwich estimation of the variance of $\hat{\tau}_r$ following:

$$\hat{\sigma}_r^2 = \frac{1}{n_r^2} \sum_{i \in \mathcal{R}} \hat{I}_i^2$$

where

$$\hat{I}_i = \left[\frac{A_i \left(Y_i - \hat{m}_{1,r}(X_i)\right)}{\hat{e}_r(X_i)} - \frac{(1 - A_i) \left(Y_i - \hat{m}_{0,r}(X_i)\right)}{1 - \hat{e}_r(X_i)} + \left(\hat{m}_{1,r}(X_i) - \hat{m}_{0,r}(X_i)\right) \right] - \hat{\tau}_r.$$

and similarly for $\hat{\sigma}_o^2$, estimated variance of $\hat{\tau}_o$.

C.1.2 Shrinkage estimators

The shrinkage estimators in Green et al. (2005) and Rosenman et al. (2023a) require the dimension of the vectors, i.e. K, to be greater than two and the elements to be independent. This necessitates stratifying the data. Rosenman et al. (2023a) discussed the performance of shrinkage estimators with different numbers of strata based on empirical evidence from simulations. However, their findings were inconclusive: the relative performance of the shrinkage estimators varied with different numbers of strata, and no single estimator consistently outperformed the others. The authors identified this as a potential topic for future research.

Despite the lack of definitive guidance, Rosenman et al. (2023a) found $\hat{\tau}_{S2}$ performed well with relatively few strata (2, 3, or 5). Based on this, we decided to stratify the RCT and RWD, respectively into four subgroups according to the values of X_1 and X_3. As outlined in Section 3.3.1, X_1 and X_3 are binary variables, and the ATE estimates within these subgroups are independent, meeting the required assumptions.

To construct the elements in the shrinkage estimators $\hat{\boldsymbol{\tau}}_{S1}$ and $\hat{\boldsymbol{\tau}}_{S2}$, we use the AIPW algorithm to get the subgroups ATE estimates of the RCT, forming $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ vector. We apply the same procedure to the RWD. To estimate matrix $\boldsymbol{\Sigma}_r$, we use the

sandwich estimators of variance of τ_{rk}'s. For $\hat{\boldsymbol{\tau}}_{S1}$, there is a hyper-parameter a and we set $a = K - 2$, which is a customary selection for James-Stein style estimators.

As mentioned in Section 3.3.2.3, Green et al. (2005) and Rosenman et al. (2023a) both proposed EB CIs for the elements of $\hat{\boldsymbol{\tau}}_{S1}$ and $\hat{\boldsymbol{\tau}}_{S2}$. However, since we are focused on the ATE, these element-wise CIs are not helpful. Therefore we used non-parametric bootstrap method, separately resampling the RCT and RWD data respectively, and obtained the quantiles from 200 iterations.

C.1.3 Bias-correction methods

C.1.3.1 Experiment grounding (Kallus et al., 2018)

We follow the experimental grounding approach as described in Algorithm 1 in Kallus et al. (2018). Specifically, we implement the following steps:

1. We fit two linear outcome regression models on the observational data: $m_1(X_1, X_2, X_3) = \mathbb{E}[Y \mid X_1, X_2, X_3, A = 1, S = 0]$ and $m_0(X_1, X_2, xX3) = \mathbb{E}[Y \mid X_1, X_2, X_3, A = 0, S = 0]$, and calculate their difference as an estimate of CATE estimate $\hat{\omega}(\boldsymbol{z}, \boldsymbol{c}) = \hat{m}_1(X_1, X_2, X_3) - \hat{m}_0(X_1, X_2, X_3)$.

2. We learn the bias correction function, assuming linearity, $\xi(\boldsymbol{x}) = \beta^T \boldsymbol{x}$ where $\boldsymbol{x}$ is the vector of $(x_1, x_2, X_3)^T$ and

$$\hat{\beta} = \arg\min_{\beta} \sum_{i \in \mathcal{R}} \left(q_i Y_i - \hat{\omega}\left(\boldsymbol{X}_i\right) - \beta^T \boldsymbol{X}_i \right)^2,$$

   where $q_i = 2$ if $T = 1$ and $q_i = -2$ if $T = 0$. That is, we fit $\eta(\boldsymbol{z}, \boldsymbol{c})$ through least square regression on $q_i Y_i - \hat{\omega}(\boldsymbol{V}_i)$ using the randomized data.

3. We then calculate the ATE estimates as

$$\hat{\tau}_{XG} = \frac{1}{n_r} \sum_{i \in \mathcal{R}} \hat{\omega}(\boldsymbol{X}_i) + \hat{\beta}^T \boldsymbol{X}_i.$$

For inference, we create confidence intervals using non-parametric bootstrap approach by resampling the RCT and RWD data respectively, obtaining the quantiles from 200 iterations.

C.1.3.2 Confounding function (Yang et al., 2020)

We implement the confounding function estimator using the R package `IntegrativeHTEcf` (Yang, 2020). This package assumes a linear form for both the confounding and HTE functions, limiting its application to binary outcomes. Nonetheless, the linearity assumption is appropriate for the simulations presented. We specify the covariate set

as (X_1, X_2, X_3) and the HTE covariate set to be X_3, aligning with the correct data generation model.

The main function, `IntHTEcf`, outputs the estimated linear coefficients of the HTE and confounding functions, with variance estimated via an internal non-parametric bootstrap procedure. Since we are interested in the ATE, instead of the uncertainty around the HTE coefficients, we conduct our own bootstrap procedure outside the function to estimate the uncertainty of the ATE estimates.

C.1.4 Power-likelihood method

The implementation of this method draws upon R packages `causl` (Evans, 2021). We follow these steps:

First, following the frugal parameterization, we parameterize the causal model as:

$$\begin{aligned} X_1 &\sim \mathcal{N}(\beta_1, \phi_1) \\ X_2 &\sim \text{Bernoulli}(\text{expit}(\beta_2)) \\ Y \mid do(A), X_3 &\sim \mathcal{N}(\beta_4 + \beta_5 A + \beta_6 X_3 A + \beta_7 X_3,\ \phi_4) \end{aligned}$$

and the dependence between X_1,X_2,Y follow a multi-variate Gaussian copula with constant correlation coefficients ρ_1, ρ_2 and ρ_3. We denote the vector consisting of β, ϕ and ρ parameters as $\boldsymbol{\theta}$. We do not need to assume parametric models for A and X_3 as we are interested in the distribution conditional on treatment and the effect modifier.

Using the `fitCausl` function in `causl`, we find $\hat{\boldsymbol{\theta}}_r$ and $\hat{\boldsymbol{\theta}}_o$, which are the MLEs in the RCT and RWD. Next, we perform a grid search for the optimal learning rate $\eta \in [0, 1]$. For each value of η on the grid, we approximate the η-indexed posterior of $\boldsymbol{\theta}$ as described in Section 4.3 of Lin et al. (2025). We then draw random samples from this posterior distribution, $\boldsymbol{\theta}^{(i)}$, for $i = 0, 1, \ldots, 2000$. For each sample $\boldsymbol{\theta}^{(i)}$, we approximate the ELPD using Leave-One-Out cross validation, implemented via the `loo` function in the R package `loo` (Vehtari et al., 2020). Looping through the grid of η, we select the optimal value $\hat{\eta}$ that maximizes the approximated ELPD.

We then obtain the distribution of the ATE estimate based on $\boldsymbol{\theta}^{(i)}$ samples:

$$\hat{\tau}_{PL}^{(i)} = \frac{1}{n_r} \sum_{j \in \mathcal{R}} \hat{\beta}_5^{(i)} + \hat{\beta}_6^{(i)} \times X_{3j},$$

for $i = 1, 2, \ldots, 2000$. Finally, we derive the estimator $\hat{\tau}_{PL} = \frac{1}{2000} \sum_{i=1}^{2000} \hat{\tau}_{PL}^{(i)}$, as well as the credible interval based on the quantiles.

C.1.5 Test-then-pool methods

C.1.5.1 Elastic integration (Yang et al., 2023)

We implement the elastic integration estimator using the R package `ElasticIntegrative` (Yang, 2020). This package is user-friendly and straightforward to use. The main function is `elasticHTE` and we specify (X_1, X_2, X_3) as the covariate set for the propensity score model and X_3 for the HTE function, aligning with the data generative model. `elasticHTE` implemented flexible modeling through `SuperLearner` (van der Laan et al., 2007) and we specify the same libraries used for the AIPW estimation.

For inference, as the estimator is not regular, the variance estimated through non-parametric bootstrap will be biased; this is formalized with a proof in the supplementary materials of their paper. Therefore, the authors implemented a permutation-based bias estimation within the main function. We use the variance output from the function, although we modified the function slightly to also output the covariance. As we need to marginalize it over the covariate distribution of the RCT, we need to account for the correlation between coefficient estimates.

C.1.5.2 ES-CVTMLE

We implement the ES-CVTMLE estimator using the R package `EScvtmle` (Dang et al., 2023b). This package is user-friendly and straightforward to use. The main function `ES.cvtmle` provides flexibility for combining either both arms or only the control arm of the RWD, the inclusion of a NCO, as well as different distribution families for the outcome and the NCO. We specify these settings accordingly for the simulations in Section 3.3.2 and 3.3.3. For the integrated `SuperLearner` (van der Laan et al., 2007) ensemble, we choose the discrete version with the same libraries used for the AIPW estimation. We use five folds for cross-validation.

We use the variance output from `ES.cvtmle` to construct 95% confidence intervals.

C.1.6 Weighted combination methods

C.1.6.1 MSE-minimizing estimator (Oberst et al., 2023)

We implement the MSE-minimizing estimator as $\hat{\tau}_{\hat{\lambda}} = \hat{\lambda}\hat{\tau}_o + (1-\hat{\lambda})\hat{\tau}_r$, where

$$\hat{\lambda} = \frac{\hat{\sigma}_r^2}{(\hat{\tau}_r - \hat{\tau}_o)^2 + \hat{\sigma}_r^2 + \hat{\sigma}_o^2}.$$

As $\hat{\tau}_e$ and $\hat{\tau}_o$ are independent, the covariance term in the original formula of $\hat{\lambda}$ drops out. For inference, we generate confidence intervals by applying a non-parametric bootstrap method, separately resampling the RCT and RWD data and extracting quantiles from 200 iterations.

C.1.6.2 Anchored-thresholding estimator (Chen et al., 2021)

We implement the anchored-thresholding estimator by first calculating the estimated bias $\hat{\delta}_\gamma$ through soft-thresholding:

$$\hat{\delta}_\gamma = \text{sgn}(\hat{\delta}) \left(|\hat{\delta}| - \gamma \cdot \sqrt{\hat{\sigma}_r^2 + \hat{\sigma}_o^2} \right)_+ .$$

. As discussed in Section 3.3.2.3, the hyper-parameter γ is important in controlling the combination and the larger γ is, the higher the potential bias is. Chen et al. (2021) state that γ should be of asymptotic order $\sqrt{\log\left(\min(n_r, n_o)\right)}$, so we set $\gamma = \sqrt{\log 300}$. We then plug $\hat{\delta}_\gamma$ into the estimator:

$$\hat{\tau}_{AT} = (1 - \hat{\omega})\hat{\tau}_r + \hat{\omega}(\hat{\tau}_o + \hat{\delta}_\gamma),$$

where $\hat{\omega} = \hat{\sigma}_r^2/(\hat{\sigma}_r^2 + \hat{\sigma}_o^2)$. We use a non-parametric bootstrap to obtain the 95% confidence interval.

C.1.7 Prognostic adjustment

The implementation of the PROCOVA approach is relatively straight forward. We first empirically centre the covariates as well as the treatment indicator. Then employing the `SuperLearner` ensemble, we fit a prognostic score model and make predictions on the RCT, creating prognostic scores $\hat{m}(X)$. Then, we estimate the treatment effect $\hat{\beta}_A$ using a linear regression adjusted for the empirically centred covariates, prognostic score, and their interactions with the treatment.

We use the sandwich estimator of variance of the $\hat{\beta}_A$ to construct the 95% confidence interval.

C.2 Details of real-world application

C.2.1 Target trial emulation

The calibration of observational RWD against a randomized trial is known as target trial emulation, with the goal of making an inference about the causal quantity of interest (Hernán and Robins, 2016). In this study, we construct an RWD sample designed to emulate the PIONEER 6 trial. Following the framework outlined by Hernán and Robins (2016) and further elaborated in (Hernán et al., 2022), we specify the key components of the emulation process.

C.2.1.1 Real-World Data Source

The RWD is sourced from Optum's de-identified Clinformatics® Data Mart (CDM) Database, consistent with the choice in the RCT DUPLICATE study Franklin et al. (2021) and in Dang et al. (2023a). This administrative claims database includes patient-level data on enrollment, inpatient and outpatient services, pharmacy dispensing, and selected laboratory results for commercially insured individuals across all 50 U.S. states.

C.2.1.2 Inclusion and Exclusion Criteria

We emulated the inclusion and exclusion criteria from the PIONEER 6 protocol using information available in the Optum CDM. For our two data fusion designs, we designated specific *recruitment windows* for the RWD emulations, which will be elaborated upon in the subsequent subsections.

Eligibility for PIONEER 6 required all of the following:

1. Informed consent obtained before any trial-related activities. Trial-related activities are any procedures that are carried out as part of the trial, including activities to determine suitability for the trial.

2. Male or female diagnosed with type 2 diabetes.

3. 3a. Age ≥ 50 years at screening **and** at least one of the following conditions:
    i. Prior myocardial infarction (MI)
    ii. Prior stroke or transient ischaemic attack (TIA)
    iii. Prior coronary, carotid or peripheral arterial revascularisation
    iv. $> 50\%$ stenosis on angiography or imaging of coronary, carotid, or lower extremity arteries
    v. History of symptomatic coronary heart disease documented by, e.g., positive exercise stress test or any cardiac imaging, or unstable angina pectoris with ECG changes
    vi. Asymptomatic cardiac ischaemia documented by positive nuclear imaging test or exercise stress test or stress echocardiography or any cardiac imaging
    vii. Chronic heart failure (New York Heart Association [NYHA] class II–III)
    viii. Moderate renal impairment (estimated glomerular filtration rate [eGFR] between 30–59 mL/min/1.73 m^2)

or

3b. Age $\geq$ 60 years at screening **and** at least one of the following risk factors:

i. Microalbuminuria or proteinuria

ii. Hypertension and left ventricular hypertrophy by ECG or imaging

iii. Left ventricular systolic or diastolic dysfunction by imaging

iv. Ankle–brachial index < 0.9

For an eligible subject, all exclusion criteria below must be answered "no":

1. Known or suspected hypersensitivity to the trial product or related products.
2. Previous participation in this trial. (Participation is defined as signed informed consent.)
3. Female who is pregnant, breastfeeding, intends to become pregnant, or is of childbearing potential and not using adequate contraceptive methods (as required by local regulations or practice).
4. Receipt of any investigational medicinal product within 90 days before screening.
   For Brazil only: Participation in other clinical trials within one year prior to screening unless there is a direct benefit to the research subject at the investigator's discretion.
5. Participation in another clinical trial of an investigational medicinal product. (Participation in a clinical trial evaluating stents is allowed.)
6. Current or previous treatment (within 90 days prior to screening) with any GLP-1 receptor agonist, DPP-4 inhibitor, or pramlintide.
7. Any disorder that, in the investigator's opinion, might jeopardize the subject's safety or compliance with the protocol.
8. Family or personal history of multiple endocrine neoplasia type 2 (MEN 2) or medullary thyroid carcinoma (MTC).
9. History of pancreatitis (acute or chronic).
10. History of major surgical procedures involving the stomach that could affect absorption of the trial product (e.g., subtotal or total gastrectomy, sleeve gastrectomy, gastric bypass surgery).

11. Subjects presently classified as being in New York Heart Association (NYHA) Class IV heart failure.

12. Planned coronary, carotid, or peripheral artery revascularisation scheduled after screening.

13. Any of the following within the past 60 days prior to screening:

    - Myocardial infarction
    - Stroke
    - Hospitalisation for unstable angina pectoris or transient ischaemic attack

14. Chronic or intermittent haemodialysis or peritoneal dialysis or severe renal impairment (corresponding to eGFR <30 mL/min/1.73 m^2).

15. History or presence of malignant neoplasms within the last 5 years (exceptions may apply for certain skin cancers).

16. History of alcohol or drug abuse within one year prior to screening.

17. Mental incapacity, unwillingness, or language barriers precluding adequate understanding of informed consent or protocol requirements.

The rationale behind these criteria is to ensure the inclusion of individuals with T2D at high risk for cardiovascular events, thereby increasing the event rate and statistical power within a feasible follow-up period. Individuals with recent cardiovascular events were excluded to avoid potential biases from outcomes likely influenced by preceding events or interventions.

While we replicate the inclusion and exclusion criteria as closely as possible in the RWD, we acknowledge important limitations. Some prior conditions and treatments may be imperfectly captured or coded differently in claims data compared to the structured trial case report forms. These discrepancies may introduce measurement error or residual confounding despite careful cohort definition.

C.2.1.3 Defining time zero

Establishing *time zero*—the point at which follow-up begins—is crucial for accurately emulating the PIONEER 6 trial using RWD. In randomized trials, time zero is naturally and unambiguously defined by the moment of randomization, which simultaneously anchors baseline covariate measurement and initiates follow-up. By contrast, observational data lack such a clear temporal reference, particularly for

comparator groups receiving standard care or no treatment. This complicates the task of aligning the observational study design with the target trial protocol.

As highlighted in the RCT DUPLICATE initiative (Franklin et al., 2021), untreated diabetic patients often differ systematically from their treated counterparts in ways that may not be fully captured in observational datasets, likely due to disparities in healthcare access and engagement. To mitigate this bias, we adopt an active comparator design. Specifically, we define time zero in the RWD as the initiation of a dipeptidyl peptidase-4 inhibitor (DPP4i). DPP4i serves as an appropriate comparators because they share semaglutide's indication for glycemic control but are not associated with increased or decreased risk of MACE (Scirica et al., 2013; Green et al., 2015). Thus, DPP4i use serves as a proxy for placebo, enabling a meaningful contrast with semaglutide.

To avoid time-related biases—particularly immortal time bias—we ensure that time zero coincides with both cohort entry and the point at which eligibility criteria are satisfied. This alignment ensures that baseline covariates are measured prior to treatment initiation and that follow-up begins consistently across treatment arms.

C.2.1.4 Understanding the baseline variables

All analyses in this section were conducted using the real-world data (Optum CDM), independent of trial outcomes. This allows the analyses to be performed *a priori*—that is, before any linkage to RCT outcomes—ensuring they are outcome-blind and free from post hoc bias.

The baseline variables included in the variable importance analyses span demographic, clinical, laboratory, and medication-related factors. These comprised sex, age, race (White, Asian, Black), and use of key medications: sodium-glucose co-transporter-2 (SGLT2) inhibitors, insulin, antihypertensive agents, lipid-lowering therapy, biguanides, sulfonylureas, and thiazolidinediones. In addition, we included clinical indicators derived from the PIONEER 6 inclusion criteria: prior myocardial infarction, prior stroke or transient ischaemic attack (TIA), prior revascularization, established coronary or peripheral arterial disease, symptomatic or asymptomatic cardiac ischaemia, chronic heart failure (New York Heart Association class II–III), moderate renal impairment (eGFR 30–59 mL/min/1.73 m^2), albuminuria or proteinuria, hypertension with left ventricular hypertrophy, ventricular dysfunction by imaging, and a low ankle–brachial index.

Predictors of 1-year MACE

Table C.2 summarizes the top baseline predictors of 1-year major adverse cardiovascular event (MACE) risk, identified using penalized regression models—LASSO,

Category	Variables
Demographic factors	Sex, age, race (white, Asian, black, others)
Laboratory measures	HbA1c, HDL, LDL, serum creatinine, serum albumin
Clinical history	History of obesity, history of morbid obesity
Medication use	SGLT2 inhibitors, insulin, antihypertensives, lipid-lowering therapy, biguanides, sulfonylureas, thiazolidinediones
Inclusion criteria indicators	• Prior myocardial infarction • Prior stroke or transient ischaemic attack (TIA) • Prior revascularization • Significant coronary or peripheral arterial disease • Symptomatic or asymptomatic cardiac ischaemia • Chronic heart failure (NYHA class II–III) • Moderate renal impairment (eGFR 30–59 mL/min/1.73 m^2) • Albuminuria or proteinuria • Hypertension with left ventricular hypertrophy • Ventricular dysfunction by imaging • Low ankle–brachial index

Table C.1: Baseline variables considered in variable importance analysis

Ridge, and Elastic Net ($\alpha = 0.5$). Across these models, lower serum albumin, older age, higher low-density lipoprotein (LDL) cholesterol, elevated serum creatinine, and lower high-density lipoprotein (HDL) cholesterol were most strongly associated with increased MACE risk. Additional contributors included elevated HbA1c, history of chronic heart failure, prior stroke or TIA, and prior myocardial infarction. These findings were broadly aligned with the gradient boosting machine (GBM) models, which also prioritized age, HDL cholesterol, serum albumin, LDL cholesterol, and HbA1c. The prominence of continuous variables in GBM may reflect the algorithm's preference for features with greater granularity, which yield more potential split points and higher model gain.

Predictors of semaglutide use

Variable	LASSO	Ridge	Elastic Net ($\alpha = 0.5$)
Serum albumin	−3.616	−3.531	−3.581
Age	2.268	2.144	2.228
LDL	1.761	1.708	1.720
Serum creatinine	1.715	1.815	1.682
HDL	−1.136	−1.141	−1.105
HbA1c	0.596	0.618	0.584
Chronic heart failure	0.499	0.498	0.496
Prior stroke or TIA	0.472	0.476	0.467
Prior MI	0.383	0.376	0.380

Table C.2: Estimated coefficients from LASSO, Ridge, and Elastic Net ($\alpha = 0.5$) models for predicting 1-year MACE occurrence. Top variables are shown, ranked by the absolute magnitude of their LASSO coefficients. Continuous variables were rescaled to the $[0, 1]$ range using min-max normalization prior to model fitting.

Table C.3 summarizes the most influential baseline predictors of semaglutide use among eligible RWD records. Younger age, higher serum albumin, insulin use, and obesity were positively associated with semaglutide initiation, while higher LDL cholesterol, sulfonylurea use, and serum creatinine were negatively associated. Prior revascularization was also modestly linked to lower semaglutide use. These findings were broadly consistent with the GBM results, which identified age, HbA1c, LDL cholesterol, serum creatinine, and HDL cholesterol as key predictors.

Variable	LASSO	Ridge	Elastic Net ($\alpha = 0.5$)
Age	−3.075	−2.895	−3.032
Serum Albumin	1.860	2.008	1.881
Insulin	1.146	1.142	1.143
LDL	−1.113	−1.224	−1.128
SGLT2-i	0.837	0.855	0.840
History of obesity	0.818	0.771	0.807
Sulfonylureas	−0.649	−0.635	−0.645
Serum creatinine	−0.441	−0.881	−0.503
Prior revascularization	−0.315	−0.390	−0.330

Table C.3: Estimated coefficients from LASSO, Ridge, and Elastic Net ($\alpha = 0.5$) models for predicting semaglutide use among eligible RWD records. Top variables are shown, ranked by the absolute magnitude of their LASSO coefficients. Continuous variables were rescaled to the $[0, 1]$ range using min-max normalization prior to model fitting.

Predictors of outcome missingness

Table C.4 presents the top baseline covariates associated with outcome missingness. Older age, prior revascularization, albuminuria, and prior myocardial infarction were linked to a higher likelihood of missing data. Higher LDL cholesterol, HbA1c, and HDL cholesterol were also modestly associated with missingness. Similar patterns were observed in the GBM analysis, which identified age, LDL cholesterol, HDL cholesterol, HbA1c, and serum creatinine as key predictors.

Variable	**LASSO**	**Ridge**	**Elastic Net** ($\alpha = 0.5$)
Age	−1.945	−1.876	−1.939
Prior revascularization	−0.783	−0.746	−0.780
LDL	0.553	0.582	0.557
HbA1c	0.468	0.502	0.472
HDL	0.450	0.468	0.452
Race: Asian	−0.366	−0.365	−0.366
Albuminuria	−0.321	−0.303	−0.321
Prior MI	−0.231	−0.239	−0.231
Race: White	−0.229	−0.223	−0.229

Table C.4: Top variables associated with outcome missingness, based on coefficients from LASSO, Ridge, and Elastic Net ($\alpha = 0.5$) models. Variables are ordered by the absolute magnitude of their LASSO coefficients. Continuous variables were rescaled to the $[0, 1]$ range using min-max normalization prior to model fitting.

Across models predicting outcome missingness, semaglutide use, and 1-year MACE, **age**, **lipid markers (LDL, HDL)**, **serum creatinine**, and **HbA1c** consistently emerged as important baseline predictors based on LASSO, Ridge, and Elastic Net coefficients. GBM results were broadly consistent, with a slight emphasis on continuous variables likely due to their greater capacity to create informative splits during tree construction.

C.2.1.5 Treatment Effect Heterogeneity Analysis

This section investigates the presence of treatment effect heterogeneity—i.e. whether the effect of semaglutide varies across baseline covariates. We apply the Causal Forest method (Athey et al., 2019), implemented via the `grf` package (Tibshirani et al., 2017) to estimate individual-level conditional average treatment effects (CATEs).

It is important to note that Causal Forest detects effect heterogeneity but does not itself establish causality. A standard Causal Forest assumes the absence of unmeasured confounding; therefore, we base our analysis on the PIONEER 6 trial, where treatment assignment is randomized by design. In practice, for outcome

blind or pre-trial analyses, an alternative is to use historical trial data. Notably, the sample size of PIONEER 6 is relatively small, and as a result, some detected effects may not reach statistical significance.

To mitigate overfitting and ensure model stability, we restricted the set of input variables for the Causal Forest to the top 15 predictors of MACE, as identified in a prior GBM model. These include: These variables include: age, HDL, serum albumin, LDL, HbA1c, serum creatinine, indicators for key cardiovascular and renal conditions (prior myocardial infarction, prior stroke or transient ischaemic attack, prior revascularization procedure, significant coronary, carotid, or peripheral arterial stenosis, symptomatic or asymptomatic cardiac ischaemia, heart failure, and moderate renal impairment), sex, race (white), and additional comorbidity and medication indicators (antihypertensive use, biguanide use, and insulin use).

The top five most important variables driving treatment effect heterogeneity are: `ldl`, `hdl`, `creatinine_serum`, `insulin_all`, and `hba1c`.

To better understand the CATE function, we use the `best_linear_projection` function, which projects $\tau(\boldsymbol{x})$ onto the linear space spanned by a set of features:

$$\tau(\boldsymbol{X}) \sim \beta_0 + \beta \boldsymbol{V},$$

where $\boldsymbol{V} \subset \boldsymbol{X}$. If we set $\boldsymbol{V}$ to {`ldl`, `hdl`, `creatinine_serum`, `insulin_all`, and `hba1c`}, we can estimate a simple linear association measure for the CATEs.

Variable	**Estimate**	**Std. Error**	**t value**	**Pr($>\|t\|$)**
Intercept	−0.0137	0.0429	−0.3202	0.7488
LDL	0.0177	0.0132	1.3416	0.1798
HDL	0.0126	0.0131	0.9443	0.3451
Serum creatinine	−0.0156	0.0138	−1.1668	0.2434
Insulin use	−0.0181	0.0142	−1.2751	0.2024
HbA1c	−0.0064	0.0141	−0.4553	0.6489

Estimates are from a best linear projection of the conditional average treatment effect. Confidence intervals are cluster- and heteroskedasticity-robust (HC3).

Table C.5: Best linear projection of the conditional average treatment effect

In Table 5, we present the estimated coefficients from the best linear projection of the CATE function. None of the treatment–covariate interactions were statistically significant, likely reflecting the limited sample size of the PIONEER 6 trial. Notably, age and sex were not identified as treatment effect modifiers in our analysis. This finding is consistent with results previously reported in the literature.

D'Andrea et al. (2020), in a pooled analysis of ten clinical trials involving 89,790 patients, reported that uncontrolled hypertension, obesity, gender, age, and race did not significantly modify the cardiovascular effects of GLP-1 receptor agonists. Similarly, Rentzeperi et al. (2022) found that that evidence on sex-based heterogeneity in treatment response remains inconclusive and merits further investigation.

Overall, it is reassuring that our findings based on PIONEER 6 data are broadly consistent with those reported in existing literature.

C.2.1.6 Emulated RWD samples under Design 1

To construct the emulated real-world data (RWD) sample under Design 1, we first identified individuals with type 2 diabetes mellitus (T2DM) in the Optum CDM who had at least 180 days of prior observation (n = 9,043,190). Among these, 912,167 patients initiated DPP4i therapy. We then applied trial-specific inclusion and exclusion criteria, yielding a cohort of 170,449 individuals. Restricting to those who initiated treatment within the PIONEER 6 recruitment window (January to August 2017) produced a subsample of 16,111 patients.

From this group, we selected individuals with non-missing values for five key biomarkers—HbA1c, HDL cholesterol, LDL cholesterol, serum creatinine, and serum albumin[1]—leaving 4,370 patients. To improve clinical comparability with the trial population, we excluded records with covariate values outside the 2.5th–97.5th percentile range of the PIONEER 6 distributions, resulting in 3,652 individuals. We then applied 1:1 matching on age, HbA1c, HDL, LDL, serum creatinine, and baseline insulin use, yielding a final matched sample of 3,137 individuals.

Table C.6 summarizes the full data flow.

Data Stage	**Number of Observations**
Full database: T2DM patients with at least 180 days of prior observation	9,043,190
Initiated DPP-4i therapy	912,167
Applied inclusion and exclusion criteria	170,449
Within PIONEER 6 recruitment time window	16,111
Complete key baseline biomarkers[a]	4,370
Covariate values within the 2.5th–97.5th percentile range of PIONEER 6 population[b]	3,652
1:1 matching [c]	3,137

[a] Defined as having non-missing measurements for HbA1c, HDL cholesterol, LDL cholesterol, serum creatinine, and serum albumin at baseline.
[b] Based on baseline HbA1c, HDL cholesterol, LDL cholesterol, and serum creatinine distributions in the PIONEER 6 trial population.
[c] 1:1 matching conducted on age, HbA1c, HDL cholesterol, LDL cholesterol, serum creatinine, and baseline insulin use.

Table C.6: Construction of the RWD sample under Design 1

[1] Age and sex were fully observed in the CDM. Medication use and medical history variables were complete by construction. BMI was excluded due to 99% missingness. Serum creatinine was preferred over eGFR due to eGFR's dependence on age, sex, and race.

	RWD control arm (n = 3137)	RCT placebo arm (n = 1540)	RCT semaglutide arm (n = 1553)
MACE rate - %	4.4	4.3	3.0
Age - years, mean (SD)	67.2(7.6)	66.4(7.1)	65.9(7.1)
Female - %	43.5	31.3	31.7
Race			
White - %	44.6	72.7	72.4
Black - %	12.2	6.4	5.4
Other - %	43.2	20.9	22.2
HbA1c -%, mean(SD)	8.2(1.8)	8.2(1.6)	8.2(1.6)
LDL cholesterol-mg/dl,mean(SD)	86.0(34.1)	84.8(32.1)	83.7(34.0)
HDL cholesterol-mg/dl,mean(SD)	42.9(10.8)	41.6(10.7)	41.9(11.0)
Creatinine serum - mg/dL, mean(SD)	1.1(0.3)	1.0(0.3)	1.0(0.4)
Albumin serum - g/dL, mean(SD)	4.2(0.3)	4.2(0.3)	4.2(0.3)
eGFR-ml/min/1.73 m^2,mean(SD)	74.4(21.7)	74.2(20.9)	74.2(21.0)
Insulin - %	13.1	61.0	61.0
SGLT2-i - %	6.9	8.9	10.4
Antihypertensives - %	75.6	93.8	93.9
PreviousMI - %	10.8	36.9	35.5
Previous stroke- %	9.6	16.3	15.2
Previous heart failure - %	12.8	12.5	11.8
Outcome missingness - %	17.9	0.0	0.0

Table C.7: Under Design 1, baseline characteristics, outcome missingness and event rates for PIONEER 6 and matched RWD sample

Table C.7 compares baseline characteristics, outcome missingness, and event rates across the PIONEER 6 semaglutide and placebo arms and the emulated RWD control arm. Despite the outcome-blind matching process, MACE rates were nearly identical between the RWD and RCT placebo groups (4.4% vs. 4.3%), suggesting strong alignment in baseline cardiovascular risk.

The RWD cohort was modestly older, had a higher proportion of female participants, and included fewer individuals identified as White compared to the trial population. Clinical biomarkers—HbA1c, LDL, HDL, serum creatinine, and albumin—were well balanced across groups. In contrast, baseline insulin use was substantially lower in the RWD sample (13.1% vs. 61.0%), with similarly reduced use of SGLT2 inhibitors and antihypertensive medications. Outcome missingness was also notable in the RWD (17.9%), whereas it was fully observed in the RCT.

C.2.1.7 Emulated RWD samples under Design 2

Incorporating both treatment and control arms into the RWD sample introduced additional challenges due to the timing of semaglutide's availability in the U.S. Although injectable semaglutide received FDA approval in December 2017 (U.S. Food and Drug Administration, 2020), meaningful uptake in the Optum CDM did not occur until 2019. As a result, restricting the recruitment window to match PIONEER 6 would have excluded nearly all semaglutide users from the RWD, making

a broader time frame necessary.

Extending data collection into 2020 introduced additional complexities, as COVID-19 significantly increased 1-year MACE risk (HR 1.55, 95% CI: [1.50, 1.60]) (Xie et al., 2022). To balance comparability and sample size, we selected a two-year window before and after PIONEER 6 recruitment, yielding 89,036 eligible RWD records, including 2,081 in the treatment arm. After applying the same filtering criteria as in the control-only scenario, the final RWD sample comprised 19,206 patients.

Design 2 follows the same overall emulation framework as Design 1, with adaptations to accommodate treatment availability and dual-arm structure. A summary of the modified construction steps is provided in Table C.8.

Data Stage	**Number of Observations**
Full database: T2DM patients with at least 180 days of prior observation	9,043,190
Applied inclusion and exclusion criteria	912,167
Within two years before/after PIONEER 6 recruitment window	91,736
Complete key baseline biomarkers[a]	24,616
Covariate values within the 2.5th–97.5th percentile range of PIONEER 6 population[b]	19,206
1:3 matching [c]	9,411

[a] Defined as having non-missing measurements for HbA1c, HDL cholesterol, LDL cholesterol, serum creatinine, and serum albumin at baseline.
[b] Based on baseline HbA1c, HDL cholesterol, LDL cholesterol, and serum creatinine distributions in the PIONEER 6 trial population.
[c] 1:1 matching conducted on age, HbA1c, HDL cholesterol, LDL cholesterol, serum creatinine, and baseline insulin use.

Table C.8: Construction of the RWD sample under Design 2

Table C.9 presents baseline characteristics, outcome missingness, and event rates for the RWD semaglutide and control arms under Design 2, alongside the corresponding PIONEER 6 trial arms. Encouragingly, 1-year MACE rates were closely aligned across groups, despite the outcome not being used in the matching procedure, reflecting the strength of the outcome-blind design. Nevertheless, notable differences persisted. The RWD cohorts had a lower proportion of White patients and a higher proportion classified as "Other", reflecting broader real-world diversity. As race was not identified as a treatment effect modifier in our analysis—or in the wider literature—this difference is unlikely to introduce meaningful bias in effect estimation. Patterns of baseline medication use also diverged. RWD participants had lower usage rates of insulin, SGLT2 inhibitors, and antihypertensive agents compared to their RCT counterparts. Furthermore, outcome missingness remained substantial in the RWD (ranging from 18.8% to 26.6%), whereas outcomes were fully observed in the RCT.

	RWD control arm (n = 9020)	RWD semaglutide arm (n = 391)	RCT control arm (n = 1573)	RWD semaglutide arm (n = 1564)
MACE rate - %	4.4	2.8	4.3	3.0
Age - years, mean (SD)	65.7(8.3)	61.0(7.2)	66.4(7.1)	65.9(7.1)
Female - %	45.0	48.6	31.3	31.7
Race				
White - %	47.3	59.6	72.7	72.4
Black - %	12.9	13.0	6.4	5.4
Other - %	39.9	27.4	20.9	22.2
HbA1c -%, mean(SD)	7.9(1.6)	8.2(1.9)	8.2(1.6)	8.2(1.6)
LDL cholesterol-mg/dl,mean(SD)	84.8(33.7)	84.2(33.7)	84.8(32.1)	83.7(34.0)
HDL cholesterol-mg/dl,mean(SD)	41.9(11.2)	42.6(10.8)	41.6(10.7)	41.9(11.0)
Creatinine serum - mg/dL, mean(SD)	1.0(0.3)	1.0(0.3)	1.0(0.3)	1.0(0.4)
Albumin serum - g/dL, mean(SD)	4.3(0.3)	4.3(0.3)	4.2(0.3)	4.2(0.3)
eGFR-ml/min/1.73 m^2,mean(SD)	77.2(22.1)	81.4(21.3)	74.2(20.9)	74.2(21.0)
Insulin - %	30.2	45.5	61.0	61.0
SGLT2i - %	6.9	18.9	8.9	10.4
Antihypertensives - %	76.6	72.6	93.8	93.9
PreviousMI - %	12.5	14.1	36.9	35.5
Previous stroke- %	10.1	9.7	16.3	15.2
Chronic heart failure - %	17.7	23.0	12.5	11.8
Outcome missingness - %	18.8	26.6	0.0	0.0

Table C.9: Under Design 2, baseline characteristics, outcome missingness and event rates for PIONEER 6 and matched RWD sample

C.2.2 Implementation of methods

This section describes the application of various data fusion methods to combine PIONEER 6 and the emulated datasets constructed in Section C.2.1. Variable selection was informed by the analyses in Sections C.2.1.4 and C.2.1.5, focusing on covariates most predictive of treatment assignment and MACE risk. The final set of covariates used for adjustment—hereafter referred to as the selected covariates—includes:

- Demographics and laboratory measures: age, HbA1c, HDL cholesterol, LDL cholesterol, serum creatinine, and serum albumin;
- Medication use: SGLT2 inhibitor use, baseline insulin use;
- Cardiovascular history: prior myocardial infarction, prior coronary or peripheral revascularization, symptomatic or asymptomatic cardiac ischaemia, and prior stroke or transient ischaemic attack (TIA).

These covariates are consistently considered across all estimators for comparability.

C.2.2.1 AIPW

For the analysis of the ATE in the RCT and RWD samples (under Design 2 only) respectively we use the augmented inverse probability weighting (AIPW) estimator implemented in the `AIPW` R package (Zhong et al., 2021). This doubly robust method leverages both outcome regression and propensity score models, combining them to reduce bias and improve efficiency.

We estimate both the outcome and treatment models using discrete SuperLearners (van der Laan et al., 2007). The libraries include:

- Outcome model:
  - Generalized linear models (GLMs, via `speedglm`; Enea et al. (2009)),
  - Generalized additive models (GAMs; Hastie and Tibshirani (1987)),
  - Random forests (via `ranger`; Wright and Ziegler (2017)).
- Propensity score model: GLMs, via `speedglm`.

To improve generalizability and reduce overfitting, we use three-fold cross-fitting, and truncate estimated propensity scores to the interval [0.025, 0.975] to stabilize inverse weights.

Due to the stochastic nature of SuperLearner ensemble fitting, we repeat the full estimation procedure five times. The final ATE estimate is obtained by averaging across repetitions. From this procedure, we obtain the following point estimates and standard errors:

$$\hat{\tau}_r = -0.0124 \qquad \hat{\sigma}_r = 0.0067$$
$$\hat{\tau}_o = -0.0023 \qquad \hat{\sigma}_o = 0.0099.$$

Here, $\hat{\tau}_r$ and $\hat{\sigma}_r$ denote the estimated ATE and standard error from the RCT data, while $\hat{\tau}_o$ and $\hat{\sigma}_o$ refer to estimates from the RWD. These values serve as inputs to the subsequent fusion estimators, including the MSE-minimizing (Oberst et al., 2023) and the anchored-thresholding estimators (Cheng and Cai, 2021).

As noted in Table C.9, MACE rates were highly comparable across cohorts, despite the outcome not being included in the matching procedure. It is therefore unsurprising that, after adjusting for remaining covariate differences, $\hat{\tau}_r$ and $\hat{\tau}_o$, are not significantly different. The standard error from the RWD sample, $\hat{\sigma}_o$, is larger, reflecting the limited number of semaglutide users in the real-world cohort and greater outcome missingness.

C.2.2.2 Prognostic Adjustment

We apply the PROCOVA method in its extended version, using a targeted maximum likelihood estimator (TMLE) as described in Liao et al. (2025). This method incorporates a prognostic score—an estimate of expected outcomes under control—into a doubly robust framework to improve precision and mitigate residual confounding.

A prognostic score model is first trained on the control arm of the RWD samples, separately under Design 1 and Design 2. Because approximately 18% of outcomes

are missing in the RWD, we first estimate inverse probability of missingness weights using a SuperLearner ensemble. The missingness model includes generalized linear models (GLMs) and generalized additive models (GAMs) in its library. These weights are incorporated into the fitting of the prognostic model to reduce bias from informative missingness.

The prognostic model itself is estimated using SuperLearner, with a library that includes GLM, GAM, and random forest. The fitted model is then applied to the PIONEER 6 trial sample, yielding individual-level prognostic scores for all observations.

We then perform a TMLE analysis using the selected covariates (Section A.2) together with the prognostic score as adjustment variables. TMLE is implemented via the `tmle` R package (Gruber and van der Laan, 2012). The SuperLearner library for the outcome model includes GLM, GAM, and random forest; for the propensity score model, we use only GLM to preserve model stability. The final output includes the point estimate of the risk difference and the associated 95% confidence interval.

C.2.2.3 Shrinkage estimators

We implement shrinkage estimators following the approaches of Green et al. (2005) and Rosenman et al. (2023a), which combine trial and real-world data in a stratum-specific manner to balance bias and variance.

Because these estimators require more than two strata to stabilize shrinkage across heterogeneous subgroups, we stratify both the RCT and RWD samples into four strata based on baseline insulin use and LDL cholesterol. These variables were identified in Section C.2.1.5 as potential effect modifiers and can potentially yield comparable strata in terms of expected treatment effects.

We estimate the ATE within each stratum using the AIPW estimator described in Section C.2.2.1, using the same configuration (e.g., SuperLearner libraries, cross-fitting, and truncation). We apply the shrinkage estimators using the formulas set out in Section 3.2.3. Inference is based on nonparametric bootstrap with 300 replications to construct 95% confidence intervals.

Notably, these estimators are only applicable under Design 2, as they require the presence of semaglutide users in the RWD for within-stratum treatment effect estimation.

C.2.2.4 MSE-minimizing estimator

We implement the mean squared error (MSE)-minimizing estimator proposed by Oberst et al. (2023), which adaptively combines RCT and RWD estimates to mini-

mize overall estimation error. The estimator takes the form

$$\hat{\tau}_{\hat{\lambda}} = \hat{\lambda}\hat{\tau}_o + (1 - \hat{\lambda})\hat{\tau}_r,$$

where

$$\hat{\lambda} = \frac{\hat{\sigma}_r^2}{(\hat{\tau}_r - \hat{\tau}_o)^2 + \hat{\sigma}_r^2 + \hat{\sigma}_o^2} = 0.2566.$$

Confidence intervals are obtained via non-parametric bootstrap with 300 iterations. This estimator is only used under Design 2, as it requires both RCT and RWD ATE estimates based on semaglutide exposure.

C.2.2.5 Anchored thresholding estimator

We implement the anchored-thresholding estimator (Chen et al., 2021) by first calculating the estimated bias $\hat{\delta}_\gamma$ through soft-thresholding:

$$\hat{\delta}_\gamma = \text{sgn}(\hat{\delta}) \left(|\hat{\delta}| - \gamma \cdot \sqrt{\hat{\sigma}_r^2 + \hat{\sigma}_o^2} \right)_+,$$

where $\gamma = 0.5 \times \sqrt{\log(3137)}$ and $\hat{\delta} = \hat{\tau}_e - \hat{\tau}_o$. The hyperparameter threshold γ penalizes large discrepancies between trial and RWD estimates while allowing for some borrowing. We then plug $\hat{\delta}_\gamma$ into the final estimator:

$$\hat{\tau}_{AT} = (1 - \hat{\omega})\hat{\tau}_r + \hat{\omega}(\hat{\tau}_o + \hat{\delta}_\gamma),$$

where $\hat{\omega} = \hat{\sigma}_r^2/(\hat{\sigma}_r^2 + \hat{\sigma}_o^2)$ weights the two estimates based on their relative precision. We use non-parametric bootstrap with 300 iterations to obtain the 95% confidence interval. This method is applicable only under Design 2, as it requires observed semaglutide treatment in the RWD.

C.2.2.6 Adaptive power likelihood

Following the frugal parameterization, we parameterize the causal model as:

$$\begin{aligned} X_1 &\sim \mathcal{N}(\beta_1, \phi_1) \\ X_2 &\sim \mathcal{N}(\beta_2, \phi_2) \\ Y \mid do(A), \boldsymbol{C} &\sim \text{Bernoulli}\left\{ \text{expit}\left(\alpha_0 + \sum_{j=1}^{5} \alpha_j C_j + \alpha_6 A + \sum_{j=1}^{5} \alpha_{6+j} C_j A\right)\right\}, \end{aligned}$$

where $\boldsymbol{X} = \{\text{Age, Serum albumin}\}$ and $\boldsymbol{C} = \{\text{Insulin, LDL, HDL, Serum creatinine, HbA1c}\}$, selected based on their predictive importance and their role as potential effect modifiers (Section C.2.1.4 - C.2.1.5).

The dependence between X_1,X_2,Y follow a multivariate Gaussian copula with constant correlation coefficients ρ_1, ρ_2 and ρ_3. We do not model the treatment A or effect modifiers $\boldsymbol{C}$ parametrically, since inference focuses on $p(Y \mid do(A), \boldsymbol{C})$. Approximately 18% of RWD outcomes are missing. To address this, we use multiple imputation (Gelman et al., 1995) via the `mice` R package (van Buuren and Groothuis-Oudshoorn, 2011). We generate 5 imputed datasets and perform 10,000 posterior draws of the parameter vector $\boldsymbol{\theta} = \{\beta_1, \beta_1, \alpha_0, \ldots, \alpha_{11}, \phi_1, \phi_2, \rho_1, \rho_2\rho_3\}$

For each posterior sample $\hat{\boldsymbol{\theta}}^{(i)}$, we calculate the ATE:

$$\widehat{\tau}_{\mathrm{PL}}^{(i)} = \frac{1}{n_r} \sum_{j \in \mathcal{R}} \left\{ \mathrm{expit}\left(\boldsymbol{\alpha}^{(i)\top} \mathbf{V}_1^{(j)}\right) - \mathrm{expit}\left(\boldsymbol{\alpha}^{(i)\top} \mathbf{V}_0^{(j)}\right) \right\},$$

where

$$\boldsymbol{\alpha}^{(i)} = \begin{pmatrix} \alpha_0^{(i)} \\ \vdots \\ \alpha_{11}^{(i)} \end{pmatrix}, \quad \mathbf{V}_1^{(j)} = \begin{pmatrix} 1 \\ \mathbf{C}^{(j)} \\ \mathbf{C}^{(j)} \end{pmatrix}, \quad \mathbf{V}_0^{(j)} = \begin{pmatrix} 1 \\ \mathbf{C}^{(j)} \\ \mathbf{0} \end{pmatrix},$$

and $\mathcal{R}$ indexes RCT individuals. The final ATE estimate is the mean:

$$\hat{\tau}_{PL} = \frac{1}{10000} \sum_{i=1}^{10000} \hat{\tau}_{PL}^{(i)}$$

with 95% credible intervals derived from the empirical quantiles of the posterior samples.

This method is applicable to both Design 1 and Design 2, and we adopt the same parameterization in both settings. Under Design 1, however, there are no observed semaglutide users in the RWD. As a result, the variance–covariance components involving the treatment variable A are undefined and therefore contribute no information toward the estimation of coefficients associated with A. Nevertheless, the model continues to leverage covariate distributions and predictive structure from the RWD.

C.2.2.7 ES-CVTMLE

We implement the ES-CVTMLE estimator using the `EScvtmle` (Dang et al., 2023b) R package. We use all selected covariates as inputs and apply a discrete `SuperLearner` (van der Laan et al., 2007) ensemble, with the same candidate learners used in AIPW estimation (GLM, GAM, and random forest). Ten-fold cross-validation is used throughout.

Due to the stochastic nature of the method, we repeat the full procedure five times and average the resulting estimates. The variance estimates output by `ES.cvtmle`

are used to construct 95% confidence intervals. This method is compatible with both Design 1 and Design 2.

Appendix D

Appendix for Chapter 4: Exact simulation of longitudinal data from marginal structural models

Contents

D.1 Extension to include multiple time-varying variables 154
D.2 Proof of Lemma 4.1 156
D.3 Proof of Proposition 4.1 157
D.4 Additional details of the illustration in Section 4.5.2 158
D.5 Seaman and Keogh (2024) simulations 160

D.1 Extension to include multiple time-varying variables

We outline how the proposed algorithm in Section 4.4 can be extended to include multiple time-varying covariates. From a simulation perspective, there is no distinction between whether such a covariate is observed or unobserved; for illustrative purposes, we assume that the additional time-varying variable H follows a latent process $\{H_k\}_{k\in\mathcal{K}}$. The steps to simulate data from the model depicted in Figure D.1 are as follows:

1. **Baseline simulation**: The directed acyclic graph in Figure D.1 indicates that the latent variable H_0 precedes Z in topological order, potentially indicating an initial selection bias. Accordingly, we simulate baseline values (z, h_0, l_0, a_0)

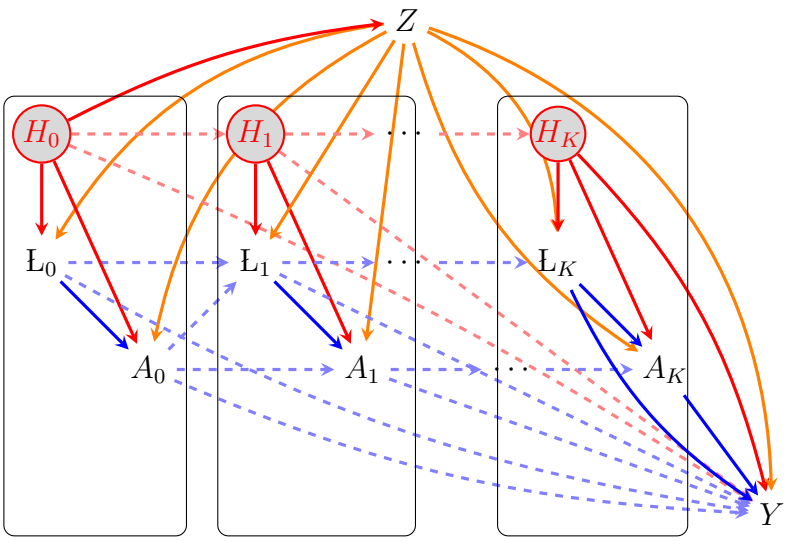


Figure D.1: Extension of the model in Figure 4.1 which allows for an unobserved process H_k.

sequentially:

$$\begin{aligned} h_0 &= F^{-1}_{H_0}(w_{0,0}) \\ z &= F^{-1}_{Z|H_0}(w_Z \,|\, h_0) \\ l_0 &= F^{-1}_{L_0|ZH_0}(w_{0,1} \,|\, z, h_0) \\ a_0 &= F^{-1}_{A_0|ZH_0L_0}(w_{0,2} \,|\, z, h_0, l_0), \end{aligned}$$

where $w_Z, w_{0,0}, w_{0,1}, w_{0,2}$ are independent standard uniform variables.

2. **Time-step simulation**: For each time step $k = 1, \ldots, K$, simulate:

$$\begin{aligned} h_k &= F^{-1}_{H_k|\bar{H}_{k-1}}(w_{k,1} \,|\, \bar{h}_{k-1}) \\ l_k &= F^{-1}_{L_k|Z\bar{H}_k\bar{L}_{k-1}\bar{A}_{k-1}}(w_{k,2}, \,|\, z, \bar{h}_k, \bar{l}_{k-1}, \bar{a}_{k-1}) \\ a_k &= F^{-1}_{A_k|Z\bar{H}_k\bar{L}_k\bar{A}_{k-1}}(w_{k,3}, \,|\, z, \bar{h}_k, \bar{l}_k, \bar{a}_{k-1}) \end{aligned}$$

where $w_{k,1}, w_{k,2}, w_{k,3}$ are again independent $U(0,1)$ variables.

3. **Outcome simulation**: Once all time-varying covariates and treatment have been simulated, we apply the same derivation following the frugal parameterization as in Section 4.4.2 and simulate Y in two steps:

   (a) Compute the marginal rank:

$$u^*_Y = h^{-1}\big[\ldots h^{-1}\big[h^{-1}\big\{h^{-1}\big(w_Y, F(l_K \mid \bar{l}_{K-1}, \bar{h}_K, z, \bar{a}_{K-1}); \Theta_{K,1}\big),$$

$$F(h_K \mid \bar{l}_{K-1}, \bar{h}_{K-1}, z, \bar{a}_{K-1}); \Theta_{K,2}\big\},$$
$$F(l_{K-1} \mid \bar{l}_{K-2}, \bar{h}_{K-1}, z, \bar{a}_{K-2}); \Theta_{K-1,1}\big] \ldots; \Theta_{0,2}\big]$$

where $w_Y \sim \mathrm{U}(0,1)$;

(b) Obtain the outcome: $y = F^{*-1}_{Y|Z,\bar{A}_K}(u^*_Y \mid z, do(\bar{a}_K))$.

Compared to the case without latent variable, the number of applications of inverse h-functions in Step 3(a) doubles. These transformations systematically disentangle the dependencies between Y and H_k, as well as between Y and L_k. In addition to the information required in the case without latent confounders, the user also specifies:

- the development of the latent confounder H_k as influenced by Z, the other time-varying confounder $\bar{L}_{k-1}$, and treatment history $\bar{A}_{k-1}$;
- $K+1$ pair-copula densities describing dependencies between the outcome and each member of the latent process: $c_{YH_k|Z,\bar{H}_{k-1},\bar{L}_{k-1},do(\bar{A}_{k-1})}$ for $k = 0, 1, \ldots, K$.

Extending the framework to include additional time-varying covariates follows a similar approach. In essence, simulating a latent process is technically equivalent to simulating any other time-varying confounder; whether a confounder is observed or unobserved does not alter the fundamental simulation methodology. With common choices of parametric bivariate copulas, evaluating the h-functions is fast, making the simulation efficient. As the dimensionality of time-varying confounders increases, either observed or unobserved, the computational complexity grows linearly.

D.2 Proof of Lemma 4.1

Proof. By definition:

$$\begin{aligned}
&F(l_j \mid \bar{l}_{j-1}, do(\bar{a}_k), y_k = 0) \\
&= \frac{\int_1^\infty F(l_j \mid \bar{l}_{j-1}, do(\bar{a}_k), \tilde{y}_k, y_{k-1} = 0) \cdot p(\tilde{y}_k \mid \bar{l}_{j-1}, do(\bar{a}_k), y_{k-1} = 0)\, d\tilde{y}_k}{\int_1^\infty p(\tilde{y}_k \mid \bar{l}_{j-1}, do(\bar{a}_k), y_{k-1} = 0)\, d\tilde{y}_k} \\
&= \frac{\int_{q_{k-1,j}}^1 F(l_j \mid \bar{l}_{j-1}, do(\bar{a}_k), \tilde{y}_k, y_{k-1} = 0)\, dF(\tilde{y}_k \mid \bar{l}_{j-1}, do(\bar{a}_k), y_{k-1} = 0)}{1 - q_{k-1,j}},
\end{aligned} \tag{D.1}$$

where $q_{k-1,j} = F(\tilde{y}_k = 1 \mid \bar{l}_{j-1}, do(\bar{a}_k), y_{k-1} = 0)$. Recall the result from Joe (1996) in (4.3), let $v_{k-1,j} = F(l_j \mid \bar{l}_{j-1}, do(\bar{a}_k), y_{k-1} = 0)$ and $u = F(\tilde{y}_k \mid \bar{l}_{j-1}, do(\bar{a}_k), y_{k-1} = 0)$, then the numerator in (D.1) can be written as:

$$\int_{q_{k-1,j}}^1 C(v_{k-1,j} \mid u)\, du = \int_{q_{k-1,j}}^1 \frac{\partial}{\partial u} C(v_{k-1j}, u)\, du$$

$$= C(v_{k-1j}, 1) - C(v_{k-1j}, q_{k-1,j})$$
$$= v_{k-1j} - C(v_{k-1j}, q_{k-1,j})$$

where the copula C is $C_{L_j\tilde{Y}_k|\bar{L}_{j-1},do(\bar{A}_k),Y_{k-1}=0}(\cdot;\Theta_{k-1,j})$, specifically. □

D.3 Proof of Proposition 4.1

Proof. We prove that the algorithm correctly simulates data for all $k \geq 0$.

Base case ($k = 0$):

Initialize the algorithm by specifying the cumulative distribution functions F_{L_0}, $F_{A_0|L_0}$, $F_{\tilde{Y}_1|do(A_0)}$, and the copula $C_{\tilde{Y}_1L_0|do(A_0)}$, parameterized by $\Theta_{0,0}$. Draw $w_{0,1}$, $w_{0,2}$, $w_{0,3} \sim U(0,1)$ and compute:

$$l_0 = F^{-1}_{L_0}(w_{0,1}) \qquad\qquad a_0 = F^{-1}_{A_0|L_0}(w_{0,2} \,|\, l_0),$$

to obtain a sample (l_0, a_0). Using the transformation:

$$F_{\tilde{Y}_1 \,|\, do(A_0)}(\tilde{y}_1 \,|\, do(a_0)) = h^{-1}\big\{F_{\tilde{Y}_1 \,|\, L_0,do(A_0)}(\tilde{y}_1 \,|\, l_0, do(a_0)), F_{L_0}(l_0); \Theta_{0,0}\big\},$$

set $w_{0,3} = F_{\tilde{Y}_1 \,|\, L_0,do(A_0)}(\tilde{y}_1 \,|\, l_0, do(a_0))$ and substitute $F_{L_0}(l_0)$ with $w_{0,1}$, yielding:

$$u_1 = h^{-1}\left(w_{0,3}, w_{0,1}; \Theta_{0,0}\right).$$

By inversion, compute $\tilde{y}_1 = F^{-1}_{Y_1|do(A_0)}(u_1)$, and the corresponding binary outcome $y_1 = I(\tilde{y}_1 < 1)$. This completes the first iteration, yielding a sample $(l_0, a_0, \tilde{y}_1, y_1)$. If $y_1 = 1$, stop; otherwise, before proceeding to $k = 1$, we store the following values:

$$u_{0,0} = F_{L_0}(l_0) \qquad\qquad q_{0,0} = F_{\tilde{Y}_1|do(A_0)}(\tilde{y}_1 = 1 \,|\, do(a_0)).$$

For $k = 1$, define the necessary cumulative distribution functions $F_{L_1 \,|\, L_0,A_0}$, $F_{A_1|\bar{L}_1,A_0}$, $F_{\tilde{Y}_2|do(\bar{A}_1)}$, and copulas $C_{\tilde{Y}_2L_1|L_0do(\bar{A}_1)}$ and $C_{\tilde{Y}_2L_0|do(A_0)}$, parameterized by $\Theta_{1,1}$ and $\Theta_{1,0}$, respectively. These two copulas associate $\tilde{Y}_2$ with previous time-varying confounders. Draw $w_{1,1}$, $w_{1,2}$, $w_{1,3}$ from $U(0,1)$, and obtain (l_1, a_1) as before.

We now need to obtain a sample of $\tilde{y}_2$. Calculate:

$$F(l_0 \,|\, do(a_0), y_1 = 0) = \frac{u_{0,0} - C^*_{\tilde{Y}_1L_0|A_0}(u_{0,0}, q_{0,0}; \Theta_{0,0})}{1 - q_{0,0}} \tag{D.2}$$
$$F(\tilde{y}_2 \,|\, l_0, do(\bar{a}_1), y_1 = 0) = h^{-1}\big\{F(\tilde{y}_2 \,|\, \bar{l}_1, do(\bar{a}_1), y_1 = 0), F(l_1 \,|\, l_0, do(\bar{a}_1), y_1 = 0); \Theta_{1,1}\big\}$$
$$F(\tilde{y}_2 \,|\, do(a_1), y_1 = 0) = h^{-1}\big\{F(\tilde{y}_2 \,|\, l_0, do(\bar{a}_1), y_1 = 0), F(l_0 \,|\, do(\bar{a}_1), y_1 = 0); \Theta_{1,0}\big\}.$$

Note that we omit some subscripts for brevity. In (D.2) we derive the rank of L_0 given a_0 among survivors at the start of the $k = 1$ interval, where $u_{0,0}$ and $q_{0,0}$

are already computed in step $k = 0$. All other necessary quantities are specified at $k = 1$. These computations yield a sample $(l_1, a_1, \tilde{y}_2, y_2)$. If $y_2 = 1$, the simulation stops; otherwise, we store and compute the following values:

$$\begin{aligned}
u_{1,0} &= F_{L_0|Y_1}(l_0 \,|\, y_1 = 0) \\
q_{1,0} &= F_{\tilde{Y}_2|do(\bar{A}_1)Y_1}(\tilde{y}_2 = 1 \,|\, do(\bar{a}_1), y_1 = 0) \\
u_{1,1} &= F_{L_1|L_0Y_1}(l_0 \,|\, l_0, y_1 = 0) \\
q_{1,1} &= F_{\tilde{Y}_2|L_0do(\bar{A}_1)Y_1}(\tilde{y}_2 = 1 \,|\, l_0, do(\bar{a}_1), y_1 = 0) \\
&= h\big\{q_{1,0}; u_{1,0}; \Theta_{1,0}\big\},
\end{aligned}$$

where we simply store the first three quantities, which are directly available from the current step, while only the last one requires computation. Therefore, the base case is satisfied since the algorithm simulates the first time step and provides the required inputs for the next.

Inductive step:

Assume the algorithm correctly simulates up to the kth iteration. We have $F(l_k \mid \bar{l}_{k-1}, do(\bar{a}_{k-1}), y_k = 0)$ and $F(y_{k+1} \mid \bar{l}_k, do(\bar{a}_k), y_k = 0)$, as well as a collection of pair copulas between Y_{k+1} and each L_j, $j \leq k$, given $(\bar{l}_{j-1}, \bar{a}_{j-1})$ with parameters $\Theta_{k,j}$. We need to compute the updated ranks $F(l_j \mid \bar{l}_{j-1}, do(\bar{a}_k), y_k = 0)$ for each $j < k$, given from the previous iteration:

$$u_{k-1,j} = F(l_j \mid \bar{l}_{j-1}, do(\bar{a}_{k-1}), y_{k-1} = 0), \quad q_{k-1,j} = F(\tilde{y}_k \mid \bar{l}_{j-1}, do(\bar{a}_{k-1}), y_{k-1} = 0).$$

This can be achieved by noting that

$$F(l_j \mid \bar{l}_{j-1}, do(\bar{a}_{k-1}), y_{k-1} = 0) = F(l_j \mid \bar{l}_{j-1}, do(\bar{a}_k), y_{k-1} = 0),$$

and

$$F(\tilde{y}_k \mid \bar{l}_{j-1}, do(\bar{a}_{k-1}), y_{k-1} = 0) = F(\tilde{y}_k \mid \bar{l}_{j-1}, do(\bar{a}_k), y_{k-1} = 0),$$

which allows us to apply (4.14) to compute the required ranks.

We must then obtain $F(y_{k+1} \mid \bar{l}_j, do(\bar{a}_k), y_k = 0)$ for each j to use in the next iteration. This can be done using (4.15), and hence, we can move to the $(k+1)$th iteration. By the principle of induction, the statement is true for all $k \geq 0$. □

D.4 Additional details of the illustration in Section 4.5.2

In our streamlined setup, we consider baseline variables $B = \{$Presence of CV disease$\}$ and time-varying covariate $L = \{$HbA1c$\}$. We consider $K = 6$ visits in

6-monthly intervals leading to a 3.5-year follow-up periods. A indicates the primary treatment, semaglutide and we assume perfect adherence so it is not time-varying. S_k indicates the use of SGLT2-i at follow-up k. The time to first occurrence of MACE to be T and $Y_k = I(T > k)$ being the binary survival status observed at each visit. Specifically we choose that at baseline:

$$B \sim \text{Bernoulli}(0.7) \qquad L_0 \sim \text{Gamma}(k = 25 + 0.8B,\ \theta = 2.5)$$
$$A \sim \text{Bernoulli}(0.5) \qquad S_0 \sim \text{Bernoulli}(0.1).$$

The choice of these parameter values are loosely based on the baseline characteristics of SOUL as reported in McGuire et al. (2023). For subsequent follow-up visits $k > 0$:

$$S_k \sim \text{Bernoulli}(\text{expit}\{\alpha_0 + \alpha_1 B + \alpha_2 L_k + \alpha_3 A\})$$
$$L_k \sim \text{Gamma}\left(k = 400(L_{k-1} + 0.2B - 0.5 S_{k-1} - A_{k-1}),\ \theta = \frac{1}{400}\right).$$

We target a Cox marginal structural model given by:

$$\lambda(\tilde{y}_{k+1} \mid do(A, \bar{S}_k), Y_k = 0) = \lambda_0 \exp(\beta_A A + \beta_S S_k + \beta_{AS} A S_k) = g_k(A, S_k).$$

This parameterization assumes a constant hazard rate between follow-up visits, influenced by the uses of semaglutide and SGLT2-i at the most recent visit. This implies that given survival at follow-up time k, the incremental survival time $\tilde{Y}_{k+1}$ follows an exponential distribution with the rate equal to the hazard above, $g_k(A, S_k)$. If $\tilde{Y}_{k+1} < 1$ then $Y_{k+1} = 1$, and 0 otherwise.

To understand the impact of differential use of SGLT2-i on the estimation of treatment effect of the primary treatment, semaglutide, we create several scenarios with different prevalence of SGLT2-i in the placebo arm. We achieve this by varying the coefficients (α_0,α_1,α_2 and α_3) of the model for S_k. We also explored varying the efficacy of SGLT2-i. Table D.1 below describes 4 levels of prevalence and 3 levels of efficacy of SGLT2-i. In total, we simulate 10 scenarios.

The values of β_S are based on the findings of McGuire et al. (2021), who reported that SGLT2 inhibitors were associated with a reduced risk of MACE, with a hazard ratio of 0.90 (95% confidence interval: 0.85–0.95) from a meta-analysis of six clinical trials. The interaction term represents the expectation that the combined effect of using both treatments is less than the total of their individual effects.

For each simulated dataset, we fit the above Cox marginal structural model $\lambda(\tilde{y}_{k+1} \mid do(A, \bar{S}_k), Y_k = 0) = \lambda_0 \exp(\beta_A A + \beta_S S_k + \beta_{AS} A S_k)$ using stabilized inverse-probability-of-treatment weights (Hernán et al., 2001), calculated as

$$W_k = \prod_{j=0}^{k} \frac{\hat{p}(S_j \mid A, \bar{S}_{j-1}, Y_j = 0)}{\hat{p}(S_j \mid B, A, \bar{L}_j, \bar{S}_{j-1}, Y_j = 0)}. \tag{D.3}$$

SGLT2-i Prevalence Scenarios	
Baseline	$S_k \sim \text{Bernoulli}(\text{expit}(-9.5 + 0.5B + 0.1L_k - 1.5A))$. Annual take-up: 10% in placebo, 2.5% in semaglutide group.
High drop-in	$S_k \sim \text{Bernoulli}(\text{expit}(-9 + 0.5B + 0.1L_k - 2A))$. Annual take-up: 15% in placebo, 2.5% in semaglutide group.
Low drop-in	$S_k \sim \text{Bernoulli}(\text{expit}(-10 + 0.5B + 0.1L_k - A))$. Annual take-up: 5% in placebo, 2.5% in semaglutide group.
No drop-in	$S_k = 0$
Effect Scenarios	
Moderate effect	$\beta_S = \log(0.9)$, $\beta_{AS} = 0.05$.
Strong effect	$\beta_S = \log(0.85)$, $\beta_{AS} = 0.1$.
Weak effect	$\beta_S = \log(0.95)$, $\beta_{AS} = 0.01$.

Table D.1: Descriptions of SGLT2-i prevalence and efficacy scenarios. The prevalence scenarios vary the annual uptake of SGLT2-i in the placebo and semaglutide groups, while efficacy scenarios are defined by combinations of β_S and β_{AS}, representing the isolated and interaction effects, respectively.

Remark D.1. The denominator in the weight formula represents the correctly specified propensity score model under the assumed data-generating process, while the numerator stabilizes the weights. To avoid unintended confounding, the baseline covariates included in the numerator must also be adjusted for in the marginal structural model (Cole and Hernán, 2008). In accordance with the model in (D.4), we only include A in the numerator for stabilization.

D.5 Seaman and Keogh (2024) simulations

We present an example that broadly replicates the setting described in Seaman and Keogh (2024). We consider baseline variables B and C, two time-varying covariates L_k and W_k, and treatment A_k at follow-up k. The time-to-event is T, with $Y_{k+1} = I(T < k+1)$ indicating failure status observed at the end of the kth interval. The auxiliary continuous variable $\tilde{Y}_{k+1}$ is defined as previously. Specifically we choose that at baseline:

$$B \sim \text{Exp}(2) \qquad L_0 \sim \text{Gamma}(k = 1 + 0.5\,B + 0.5\,C,\ \theta = 1)$$
$$C \sim \text{Bernoulli}(0.7) \qquad W_0 \sim \text{Bernoulli}(\text{expit}\{-0.2 + 0.5\,B + 0.5\,C\}).$$

We specify that throughout the treatment assignment is influenced by B, C, L_k and W_k as:

$$A_k \sim \text{Bernoulli}(\text{expit}\{-1.5 + 0.5\,B + 0.5\,C + 0.5\,L_k + 0.5\,W_k\}).$$

For subsequent follow-up visits $k > 0$:

$$L_k \sim \text{Gamma}(k = 1 + 0.5\,B + 0.5\,C + 0.1\,L_{k-1} - 0.5\,A_{k-1},\ \theta = 1)$$
$$W_k \sim \text{Bernoulli}(\text{expit}\{-0.2 + 0.5\,B + 0.5\,C + W_{k-1} - 0.6\,A_{k-1}\}).$$

We target a Cox marginal structural model given by:

$$\lambda(\tilde{y}_{k+1} \,|\, do(\bar{a}_k), c, y_k = 0) = \lambda_0 \exp\left(\beta_C\, c + \beta_A\, a_k + \beta_{CA}\, c\, a_k\right), \tag{D.4}$$

where $\lambda_0 = \exp(-2)$ is the baseline hazard. We set $\beta_C = -0.5$ and vary the values of (β_A, β_{CA}) to construct scenarios with different strengths of treatment: a strong effect scenario $(-0.5, 0.3)$ and a weak effect scenario $(0.2, 0.1)$. This parameterization assumes a constant hazard rate between follow-up visits, influenced by the treatment status at the most recent visit. This implies that given survival at follow-up time k, the incremental survival time $\tilde{Y}_{k+1}$ follows an exponential distribution with the rate equal to the hazard above:

$$\tilde{Y}_{k+1} \sim \text{Exp}(\lambda_0 \exp\left(\beta_C\, c + \beta_A\, a_k + \beta_{CA}\, c\, a_k\right)).$$

If $\tilde{Y}_{k+1} < 1$ then $Y_{k+1} = 1$, and 0 otherwise. We set the total number of follow-ups K to 6. We set the copulas describing the association between L_j $(j = 0, \ldots, k)$ and Y_{k+1} to be Student t with parameter ρ_L and 5 degrees of freedom, although each pairwise dependence could, in principle, be distinct. Similarly, we model the copulas capturing the dependence between Y_{k+1} and each W_j $(j = 0, \ldots, k)$, given L_k, as Gaussian with correlation ρ_W, though again we note that different copulas could be specified for each time point. We vary the values of ρ_L and ρ_W to construct scenarios with different strengths of confounding. Table D.2 summarizes the combinations of treatment effect size and confounding levels considered.

Effect size	
Large	$(\beta_A, \beta_{CA}) = (-0.5, 0.3)$
Small	$(\beta_A, \beta_{CA}) = (0.2, 0.1)$
Level of confounding	
High	$(\rho_L, \rho_W) = (0.9, 0.7)$
Medium	$(\rho_L, \rho_W) = (-0.5, 0.4)$
Low	$(\rho_L, \rho_W) = (-0.1, -0.2)$

Table D.2: Parameter values of treatment effect size and confounding scenarios.

We simulate data for sample size $n = \{1000, 2000, 5000\}$.

We contrast two approaches for fitting the Cox marginal structural model specified in (D.4): stabilized inverse-probability-weighting and unweighted estimation. The first employs stabilized inverse probability of treatment weighting, with weights for subject i at time step k calculated as:

$$\prod_{j=0}^{k} \frac{\hat{p}_j(A_j \mid C, \bar{A}_{j-1}, Y_j = 0)}{\hat{p}_j(A_j \mid B, C, \bar{L}_j, \bar{W}_j, \bar{A}_{j-1}, Y_j = 0)},$$

where $\hat{p}_j(A_j \mid C, \bar{A}_{j-1}, Y_j = 0)$ and $\hat{p}_j(A_j \mid B, C, \bar{L}_j, \bar{W}_j, \bar{A}_{j-1}, Y_j = 0)$ are estimated using logistic regression models fitted to the survivors at time step j.

Table D.3 summarizes the bias of estimates from 500 replications, compared to corresponding Monte Carlo standard errors as a measure of uncertainty. As the sample size increases, generally, the biases of estimates for β_C, β_A and β_{AC} obtained with inverse probability weights converge to zero, with their absolute values substantially smaller than the corresponding Monte Carlo standard errors. In contrast, the unweighted estimator exhibits considerable bias, particularly for the estimate of the treatment effect, β_A, which persists even at $n = 5000$. The magnitude of this bias far exceeds the Monte Carlo standard errors. The unweighted estimates are most biased under the high level of confounding by L and W, since this is not adjusted for.

Level of confounding	**Effect size**	n	β_C		β_A		β_{AC}	
			IPTW	Unweighted	IPTW	Unweighted	IPTW	Unweighted
High	$\beta_A = -0.5$	1000	0.00(2.18)	−0.74(1.46)	−0.16(2.26)	2.75(1.88)	0.21(2.99)	0.22(2.20)
		2000	−0.07(1.50)	−0.74(1.00)	−0.13(1.56)	2.71(1.25)	0.24(2.06)	0.24(1.51)
		5000	−0.04(1.13)	−0.78(0.67)	−0.02(1.03)	2.89(0.80)	0.04(1.46)	0.11(1.02)
	$\beta_A = 0.2$	1000	0.09(2.26)	−0.51(1.60)	0.27(1.92)	3.00(1.56)	−0.16(2.76)	0.07(2.02)
		2000	0.04(1.64)	−0.63(1.15)	0.07(1.54)	2.84(1.15)	−0.02(2.04)	0.21(1.47)
		5000	−0.04(1.07)	−0.73(0.70)	0.03(0.91)	2.81(0.68)	0.03(1.26)	0.31(0.88)
Medium	$\beta_A = -0.5$	1000	−0.13(2.12)	−0.77(1.48)	0.08(2.40)	2.66(1.78)	0.00(3.12)	0.20(2.23)
		2000	−0.01(1.60)	−0.79(0.97)	0.03(1.72)	2.53(1.25)	0.10(2.29)	0.39(1.52)
		5000	0.00(1.04)	−0.74(0.60)	−0.04(1.09)	2.58(0.79)	0.10(1.52)	0.30(0.96)
	$\beta_A = 0.2$	1000	0.01(2.34)	−0.70(1.56)	0.28(1.97)	2.63(1.54)	−0.11(2.90)	0.30(2.05)
		2000	0.01(1.73)	−0.67(1.04)	0.06(1.38)	2.50(1.08)	0.01(2.02)	0.35(1.37)
		5000	−0.02(1.12)	−0.73(0.68)	0.05(1.03)	2.46(0.66)	0.05(1.41)	0.41(0.86)
Low	$\beta_A = -0.5$	1000	−0.09(2.04)	−0.11(1.41)	−0.05(2.40)	0.64(1.75)	0.12(3.12)	−0.12(2.25)
		2000	0.02(1.57)	−0.12(0.97)	0.02(1.70)	0.63(1.22)	0.09(2.26)	0.05(1.50)
		5000	−0.06(1.07)	−0.12(0.65)	−0.01(1.11)	0.70(0.77)	0.08(1.47)	−0.07(0.94)
	$\beta_A = 0.2$	1000	0.14(2.21)	0.04(1.44)	0.27(2.09)	0.72(1.54)	−0.16(2.85)	−0.14(1.92)
		2000	−0.02(1.75)	−0.08(1.05)	0.01(1.62)	0.58(1.11)	0.12(2.23)	−0.01(1.41)
		5000	0.11(1.03)	−0.07(0.60)	0.10(0.89)	0.58(0.64)	−0.11(1.32)	−0.02(0.81)

Table D.3: Bias and Monte Carlo standard errors (×10) of estimates produced by inverse probability of treatment weighted ('IPTW') and unweighted Cox regressions, under different strengths of confounding, effect sizes, and sample sizes.

Table D.4 presents coverage rates of the 95% confidence intervals. Confidence intervals for the weighted estimator, computed using both sandwich variance esti-

mation and bootstrap resampling, achieve coverage close to the nominal level, with slight undercoverage. In contrast, the sandwich-based confidence intervals for the unweighted estimator perform poorly, especially for the treatment effect estimate, β_A.

Level of confounding	**Effect Size**	n	β_C			β_A			β_{AC}		
			W-S	W-B	UW	W-S	W-B	UW	W-S	W-B	UW
High		1000	93.5%	94.3%	92.5%	93.8%	93.5%	60.3%	94.3%	93.8%	95.0 %
	$\beta_A = -0.5$	2000	94.8%	95.3%	89.8%	95.3%	93.3%	39.0%	95.3%	93.8%	95.5 %
		5000	92.3%	92.0%	75.5%	93.8%	94.5%	5.8%	94.0%	92.8%	93.5 %
		1000	93.5%	94.3%	93.0%	95.5%	95.5%	49.3%	94.5%	94.3%	93.0 %
	$\beta_A = 0.2$	2000	92.5%	93.0%	90.0%	93.8%	92.5%	26.5%	94.0%	93.5%	92.8 %
		5000	93.5%	94.0%	82.0%	94.8%	93.5%	1.0%	94.8%	94.3%	94.0 %
Medium		1000	94.8%	94.8%	91.3%	94.3%	95.0%	64.0%	93.5%	93.3%	95.8 %
	$\beta_A = -0.5$	2000	94.3%	94.3%	88.8%	93.8%	93.8%	44.3%	95.0%	94.0%	95.8 %
		5000	97.5%	96.8%	83.0%	95.0%	94.8%	8.8%	94.3%	93.5%	92.8 %
		1000	93.3%	93.3%	91.8%	93.8%	93.5%	58.3%	92.0%	93.8%	94.8 %
	$\beta_A = 0.2$	2000	93.5%	94.8%	91.0%	96.5%	96.8%	36.3%	94.8%	95.8%	93.5 %
		5000	96.0%	94.5%	79.8%	94.5%	93.8%	6.5%	96.0%	95.0%	91.8 %
Low		1000	93.5%	93.8%	94.0%	93.5%	93.3%	94.0%	93.5%	94.0%	94.5 %
	$\beta_A = -0.5$	2000	94.8%	95.3%	94.0%	95.8%	95.3%	92.5%	94.0%	94.5%	95.8 %
		5000	93.8%	94.3%	92.3%	95.3%	95.0%	86.0%	95.3%	93.8%	95.0 %
		1000	93.0%	94.8%	94.8%	93.3%	94.5%	92.8%	95.3%	95.8%	94.5 %
	$\beta_A = 0.2$	2000	92.0%	92.3%	94.0%	92.5%	93.8%	90.8%	92.3%	94.0%	94.0 %
		5000	96.5%	96.2%	96.0%	95.2%	95.2%	88.5%	95.0%	94.5%	96.5 %

Table D.4: Coverage rates for IPTW-Sandwich (W-S), IPTW-Bootstrap (W-B), and unweighted (UW) estimators of β_C, β_A, and β_{AC} under different strengths of confounding, effect sizes, and sample sizes.

In summary, the asymptotically unbiased nature of the inverse probability weighted estimates, compared to the biased unweighted estimates, further confirms that the simulated data were generated in accordance with the Cox model in (D.4), and that confounding by L and W was appropriately introduced.

Bibliography

Odd O. Aalen. A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8):907–925, 1989.

Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198, 2009.

Alberto Abadie and Guido W. Imbens. Matching on the estimated propensity score. *Econometrica*, 84:781–807, 2016.

Marcia Angell. The ethics of clinical research in the third world. *New England Journal of Medicine*, 337(12):847–849, 1997.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):pp. 1148–1178, 2019. ISSN 00905364, 21688966. URL `https://www.jstor.org/stable/26581894`.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–973, 2005.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Tim Bedford and Roger M Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32:245–268, 2001.

Tim Bedford and Roger M Cooke. Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068, 2002.

David Benkeser and Mark van Der Laan. The highly adaptive lasso estimator. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 689–696. IEEE, 2016.

Lisa M Bodnar, Marie Davidian, Anna Maria Siega-Riz, and Anastasios A Tsiatis. Marginal structural models for analyzing causal effects of time-dependent treatments: an application in perinatal epidemiology. *American Journal of Epidemiology*, 159(10):926–934, 2004.

George EP Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

Carly Lupton Brantner, Ting-Hsuan Chang, Trang Quynh Nguyen, Hwanhee Hong, Leon Di Stefano, and Elizabeth A Stuart. Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity. *arXiv preprint arXiv:2302.13428*, 2023.

Weihua Cao, Anastasios A. Tsiatis, and Marie Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96:723–734, 9 2009.

Christian Carmona and Geoff Nicholls. Semi-modular inference: enhanced learning in multi-modular models by tempering the influence of components. pages 4226–4235, 2020.

Shuxiao Chen, Bo Zhang, and Ting Ye. Minimax rates and adaptivity in combining experimental and observational data. 9 2021.

David Cheng and Tianxi Cai. Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*, 2021.

Philip J Clare, Timothy A Dobbins, and Richard P Mattick. Causal models adjusting for time-varying confounding—a systematic review of the literature. *International Journal of Epidemiology*, 48(1):254–265, 2019.

Stephen R Cole and Miguel A. Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of epidemiology*, 168(6):656–664, 2008.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal Inference Methods for Combining Randomized Trials and Observational Studies: A Review. *Statistical Science*, 39(1):165 – 191, 2024. doi: 10.1214/23-STS889. URL `https://doi.org/10.1214/23-STS889`.

Marie Therese Cooney, Alexandra L. Dudina, and Ian M. Graham. Value and limitations of existing scores for the assessment of cardiovascular risk: A review for clinicians. *Journal of the American College of Cardiology*, 54(14):1209–1227, 2009. ISSN 0735-1097. doi: https://doi.org/10.1016/j.jacc.2009.07.020. URL `https://www.sciencedirect.com/science/article/pii/S0735109709025029`.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology)*, 34(2):187–220, 1972.

Elvira D'Andrea, Aaron S Kesselheim, Jessica M Franklin, Emily H Jung, Spencer Phillips Hey, and Elisabetta Patorno. Heterogeneity of antidiabetic treatment effect on the risk of major adverse cardiovascular events in type 2 diabetes: A systematic review and meta-analysis. *Cardiovascular Diabetology*, 19(1):1–11, 2020. doi: 10.1186/s12933-020-01040-5.

Lauren E Dang, Edwin Fong, Jens Magelund Tarp, Kim Katrine Bjerring Clemmensen, Henrik Ravn, Kajsa Kvist, et al. Case study of semaglutide and cardiovascular outcomes: An application of the causal roadmap to a hybrid design for augmenting an rct control arm with real-world data. *Journal of Clinical and Translational Science*, 7(1):e231, 2023a.

Lauren E Dang, Maya Peterson, and Mark van der Laan. EScvtmle. `https://github.com/Lauren-EylerDang/EScvtmle`, 2023b.

Lauren Eyler Dang, Jens Magelund Tarp, Trine Julie Abrahamsen, Kajsa Kvist, John B Buse, Maya Petersen, and Mark van der Laan. A cross-validated targeted maximum likelihood estimator for data-adaptive experiment selection applied to the augmentation of RCT control arms with external data. *arXiv preprint arXiv:2210.05802*, 2022.

A. Philip Dawid. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1):39–77, 2021.

Vanessa Didelez. Causal concepts and graphical models. In *Handbook of graphical models*, pages 353–380. CRC Press, 2018.

Yuyan Duan, Keying Ye, and Eric P Smith. Evaluating water quality using power priors to incorporate historical information. *Environmetrics: The Official Journal of the International Environmetrics Society*, 17(1):95–106, 2006.

Ian R Dunsmore. Asymptotic prediction analysis. *Biometrika*, 63(3):627–630, 1976.

Ralph B D'Agostino Sr, Ramachandran S Vasan, Michael J Pencina, Philip A Wolf, Mark Cobain, Joseph M Massaro, and William B Kannel. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*, 117(6):743–753, 2008.

Marco Enea et al. Fitting linear models and generalized linear models with large data sets in R. *Statistical Methods for the Analysis of Large Datasets: book of short papers*, pages 411–414, 2009.

European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. `https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-s en.pdf`, 2020. Accessed: 2024-10-29.

European Medicines Agency. Regulatory science to 2025: Strategic reflection. `https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/ema-regulatory-science-2025-strategic-reflection_en.pdf5`, 2021. Accessed: 2025-05-13.

European Medicines Agency. Qualification opinion for prognostic covariate adjustment (PROCOVA™). `https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/qualification-opinion-prognostic-covariate-adjustment-procovatm_en.pdf`, 2022. Accessed: 2023-03-13.

Robin J Evans. *causl: Methods for Specifying, Simulating from and Fitting Causal Models*, 2021. R package version 0.4.2.9001.

Robin J Evans and Vanessa Didelez. Parameterizing and simulating from causal models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):535–568, 2024.

Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2):405–421, 2017.

Jessica M Franklin, Elisabetta Patorno, Rishi J Desai, Robert J Glynn, David Martin, Kenneth Quinto, et al. Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE initiative. *Circulation*, 143(10):1002–1013, 2021.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 1995.

Alan S Gerber, Donald P Green, and Christopher W Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102(1):33–48, 2008.

Ronald Bertus Geskus. *Data analysis with competing risks and intermediate states*, volume 10. CRC press Boca Raton, 2016.

Arthur S Goldberger. Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001, 1972.

Edwin J Green and William E Strawderman. A James-Stein type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Association*, 86:1001–1006, 1991.

Edwin J Green, William E Strawderman, Ralph L Amateis, and Gregory A Reams. Improved estimation for multiple means with heterogeneous variances. *Forest Science*, 51(1):1–6, 2005.

Jennifer B Green, M Angelyn Bethel, Paul W Armstrong, John B Buse, Samuel S Engel, Jyotsna Garg, Robert Josse, Keith D Kaufman, Joerg Koglin, Scott Korn, et al. Effect of sitagliptin on cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine*, 373(3):232–242, 2015.

Susan Gruber and Mark J. van der Laan. tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, 51(13):1–35, 2012. URL `https://www.jstatsoft.org/v51/i13/`. doi:10.18637/jss.v051.i13.

Peter Grünwald and Thijs Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12 (4):1069–1103, 2017.

Ben B Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2): 481–488, 2008.

Christian Holm Hansen, Daniela Michlmayr, Sophie Madeleine Gubbels, Kåre Mølbak, and Steen Ethelberg. Assessment of protection against reinfection with SARS-CoV-2 among 4 million PCR-tested individuals in Denmark in 2020: a population-level observational study. *The Lancet*, 397(10280):1204–1212, 2021.

Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

William G Havercroft and Vanessa Didelez. Simulating from marginal structural models with time-dependent confounding. *Statistics in Medicine*, 31(30):4190–4206, 2012.

Miguel A Hernán and James M. Robins. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183 (8):758–764, 2016.

Miguel A Hernán, Stephen R Cole, Joseph Margolick, Mardge Cohen, and James M. Robins. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety*, 14(7): 477–491, 2005.

Miguel A Hernán, Wei Wang, and David E Leaf. Target trial emulation: a framework for causal inference from observational data. *JAMA*, 328(24):2446–2447, 2022.

Miguel Ángel Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men, 2000.

Miguel A. Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96(454):440–448, 2001.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Brian P Hobbs, Bradley P Carlin, Sumithra J Mandrekar, and Daniel J Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056, 2011.

Brian P Hobbs, Daniel J Sargent, and Bradley P Carlin. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis (Online)*, 7(3):639, 2012.

Chris C Holmes and Stephen G Walker. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503, 2017.

David Hume. *Enquiries Concerning the Human Understanding, and Concerning the Principles of Morals*. Clarendon Press, Oxford, second edition, 1902 edition, 1748. Reprinted from the posthumous edition of 1777, edited with introduction, comparative tables of contents, and analytical index.

Mansoor Husain, Andreas L Birkenfeld, Morten Donsmark, Kathleen Dungan, Freddy G Eliaschewitz, Denise R Franco, et al. Oral semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *New England Journal of Medicine*, 381(9):841–851, 2019.

Joseph G Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, pages 46–60, 2000.

ICH. ICH reflection paper on pursuing opportunities for harmonisation in using real-world data to generate realworld evidence, with a focus on effectiveness of medicines. `https://admin.ich.org/sites/default/files/2024-06/ICH_Reflection%20Paper_RWE_2024_0521_EndorsedbyICHAssembly_2024_0604.pdf`, 2024. Accessed: 2024-07-26.

W James and C Stein. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:361–380, 1961.

H. Joe. Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In L. Rüschendorf, B. Schweizer, and M. D. Taylor, editors, *Distributions with Fixed Marginals and Related Topics*, volume 28, pages 120–141, Hayward, CA, 1996.

Harry Joe. *Multivariate models and multivariate dependence concepts*. CRC press, 1997.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *NeurIPS*, 2018.

Immanuel Kant. *Critique of Pure Reason*. Cambridge University Press, 2nd edition, 1781. Translated by Paul Guyer and Allen Wood.

Ruth H Keogh, Shaun R Seaman, Jon Michael Gran, and Stijn Vansteelandt. Simulating longitudinal data from marginal structural models using the additive hazard model. *Biometrical Journal*, 63(7):1526–1541, 2021.

Alan B Krueger and Diane M Whitmore. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. *The Economic Journal*, 111(468):1–28, 2001.

David Lewis. Causation. *The Journal of Philosophy*, 70(17):556–567, 1973.

Sijia Li and Alex Luedtke. Efficient estimation under data fusion. *Biometrika*, 110 (4):1041–1054, 2023.

Wen Li, Frank Liu, and Duane Snavely. Revisit of test-then-pool methods and some practical considerations. *Pharmaceutical Statistics*, 19(5):498–517, 2020.

Lauren D Liao, Emilie Højbjerre-Frandsen, Alan E Hubbard, and Alejandro Schuler. Prognostic adjustment with efficient estimators to unbiasedly leverage historical data in randomized trials. *The International Journal of Biostatistics*, 2025.

Xi Lin, Jens Magelund Tarp, and Robin J Evans. Combining experimental and observational data through a power likelihood. *Biometrics*, 81(1):ujaf008, 2025.

Peter Lurie and Sidney M. Wolfe. Unethical trials of interventions to reduce perinatal transmission of the human immunodeficiency virus in developing countries. *New England Journal of Medicine*, 337(12):853–856, 1997.

Simon P Lyddon, Chris C Holmes, and Stephen G Walker. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 2019.

Darren K McGuire, Weichung J Shih, Francesco Cosentino, Bernard Charbonnel, David ZI Cherney, Samuel Dagogo-Jack, Richard Pratley, Michelle Greenberg, Shuai Wang, Susan Huyck, et al. Association of sglt2 inhibitors with cardiovascular and kidney outcomes in patients with type 2 diabetes: a meta-analysis. *JAMA Cardiology*, 6(2):148–158, 2021.

Darren K McGuire, Rodica P Busui, John Deanfield, Silvio E Inzucchi, Johannes FE Mann, Nikolaus Marx, Sharon L Mulvagh, Neil Poulter, Mads DM Engelmann, G Kees Hovingh, et al. Effects of oral semaglutide on cardiovascular outcomes in individuals with type 2 diabetes and established atherosclerotic cardiovascular disease and/or chronic kidney disease: Design and baseline characteristics of soul, a randomized trial. *Diabetes, Obesity and Metabolism*, 25(7):1932–1941, 2023.

Darren K. McGuire, Nikolaus Marx, Sharon L. Mulvagh, John E. Deanfield, Silvio E. Inzucchi, Rodica Pop-Busui, Johannes F.E. Mann, Scott S. Emerson, Neil R. Poulter, Mads D.M. Engelmann, Maria Sejersten Ripa, G. Kees Hovingh, Kirstine Brown-Frandsen, Stephen C. Bain, Matthew A. Cavender, Mette Gislum, Jens-Peter David, and John B. Buse. Oral semaglutide and cardiovascular outcomes in high-risk type 2 diabetes. *New England Journal of Medicine*, 0(0), 2025.

Carl N Morris. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983.

Beat Neuenschwander, Michael Branson, and David J Spiegelhalter. A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566, 2009.

Romain Neugebauer and Mark van der Laan. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137 (2):419–434, 2007.

Jerzy Neyman. Sur les applications de la théorie des probabilités aux expériences agricoles: Essai de principe. *Statistical Science*, 5(4):465–472, 1923.

Michael Oberst, Alexander D'Amour, Minmin Chen, Yuyan Wang, David Sontag, and Steve Yadlowsky. Understanding the risks and rewards of combining unbiased and possibly biased estimators, with applications to causal inference. *arXiv preprint arXiv:2205.10467*, 2023.

Harsh Parikh, Marco Morucci, Vittorio Orlandi, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. A double machine learning approach to combining experimental and observational data. *arXiv preprint arXiv:2307.01449*, 2023.

Judea Pearl. Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

Sally Picciotto, Miguel A Hernán, John H Page, Jessica G Young, and James M. Robins. Structural nested cumulative failure time models to estimate the effects of interventions. *Journal of the American Statistical Association*, 107(499):886–900, 2012.

Melania Pintilie. Analysing and interpreting competing risk data. *Statistics in Medicine*, 26(6):1360–1367, 2007.

Anton Pottegård, Xavier Kurz, Nicholas Moore, Christian F Christiansen, and Olaf Klungel. Considerations for pharmacoepidemiological analyses in the SARS-CoV-2 pandemic. *Pharmacoepidemiology and Drug Safety*, 29(8):825–831, 2020.

Elpiniki Rentzeperi, Stavroula Pegiou, Theocharis Koufakis, Maria Grammatiki, and Kalliopi Kotsa. Sex differences in response to treatment with glucagon-like peptide 1 receptor agonists: Opportunities for a tailored approach to diabetes and obesity care. *Journal of Personalized Medicine*, 12(3):454, 2022. doi: 10.3390/jpm12030454.

Thomas S Richardson and James M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30), 2013.

Charlotte Rietbergen, Irene Klugkist, Kristel J.M. Janssen, Karel G.M. Moons, and Herbert J.A. Hoijtink. Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemporary Clinical Trials*, 32(6):848–855, 2011.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

James M. Robins. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pages 113–159, 1989.

James M. Robins. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79(2):321–334, 1992.

James M. Robins. Marginal structural models. In *Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, pages 1–10, Alexandria, VA, 1997. American Statistical Association.

James M. Robins. Association, causation, and marginal structural models. *Synthese*, 121(1/2):151–179, 1999.

James M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 95–133. Springer, 2000.

James M. Robins and Anastasios A. Tsiatis. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics-Theory and Methods*, 20(8):2609–2631, 1991.

James M. Robins and Larry Wasserman. Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 409–420, 1997.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

James M. Robins, Miguel A. Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 45(2):212–218, 1983.

Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

Evan TR Rosenman, Guillaume Basse, Art B Owen, and Mike Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 79(4):2961–2973, 2023a.

Evan TR Rosenman, Francesca Dominici, and Luke Miratrix. Empirical Bayes double shrinkage for combining biased and unbiased causal estimates. *arXiv preprint arXiv:2309.06727*, 2023b.

Andrea Rotnitzky, James M. Robins, and Daniel O Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339, 1998.

Christian Röver and Tim Friede. Dynamically borrowing strength from another study through shrinkage estimation. *Statistical Methods in Medical Research*, 29 (1):293–308, 2020.

Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.

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

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58, 1978.

Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014.

Alejandro Schuler, David Walsh, Diana Hall, Jon Walsh, Charles Fisher, et al. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *The International Journal of Biostatistics*, 18(2):329–356, 2021.

Benjamin M Scirica, Deepak L Bhatt, Eugene Braunwald, P Gabriel Steg, Jaime Davidson, Boaz Hirshberg, Peter Ohman, Robert Frederich, Stephen D Wiviott, Elaine B Hoffman, et al. Saxagliptin and cardiovascular outcomes in patients with type 2 diabetes mellitus. *New England Journal of Medicine*, 369(14):1317–1326, 2013.

Shaun R Seaman and Ruth H Keogh. Simulating data from marginal structural models for a survival time outcome. *Biometrical Journal*, 66(8):e70010, 2024.

Xu Shi, Ziyang Pan, and Wang Miao. Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1581, 2023.

A Sklar. Fonctions de répartition á N dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pages 229–231, 1959.

Abe Sklar. Random variables, joint distribution functions, and copulas. *Kybernetika*, 9(6):449–460, 1973.

Oleg Sofrygin, Mark J van der Laan, and Romain Neugebauer. `simcausal` R package: conducting transparent and reproducible simulation studies of causal effect estimation with complex longitudinal data. *Journal of Statistical Software*, 81, 2017.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proceedings of the Third Berkeley Symposium*, pages 197–206, 1956.

Nicholas Syring and Ryan Martin. Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486, 2019.

Mark W Tenforde, H Keipp Talbot, Christopher H Trabue, Manjusha Gaglani, Tresa M McNeal, Arnold S Monto, Emily T Martin, Richard K Zimmerman, Fernanda P Silveira, Donald B Middleton, et al. Influenza vaccine effectiveness against hospitalization in the United States, 2019–2020. *The Journal of Infectious Diseases*, 224(5):813–820, 2021.

Julie Tibshirani, Susan Athey, Erik Sverdrup, and Stefan Wager. grf: Generalized random forests. *CRAN: Contributed Packages*, 07 2017. doi: 10.32614/cran.package.grf. URL `https://cir.nii.ac.jp/crid/1360020310293803392`.

U.S. Food and Drug Administration. Framework for FDA's real-world evidence program. `https://www.fda.gov/media/120060/download`, 2018. Accessed: 2023-03-13.

U.S. Food and Drug Administration. Drug trial snapshot: Ozempic. `https://www.fda.gov/drugs/drug-approvals-and-databases/drug-trial-snapshot-ozempic`, 2020. Accessed: 2024-02-27.

U.S. Food and Drug Administration. Use of real-world evidence to support regulatory decision-making for medical devices: Guidance for industry and food and drug administration staff. `https://www.fda.gov/media/154449/download`, 2021. Accessed: 2025-05-13.

U.S. Food and Drug Administration. Considerations for the use of real-world data and real-world evidence to support regulatory decision-making for drug and biological products: Guidance for industry. `https://www.fda.gov/media/171667/download`, 2023. Accessed: 2024-07-25.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

Mark J van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

Mark J van der Laan, Sky Qiu, and Lars van der Laan. Adaptive-TMLE for the average treatment effect based on randomized controlled trial augmented with real-world data. *arXiv preprint arXiv:2405.07186*, 2024.

Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27 (5):1413–1432, 2017.

Aki Vehtari, J Gabry, M Magnusson, Y Yao, P Bürkner, T Paananen, and A Gelman. loo: Efficient leave-one-out cross-validation and waic for Bayesian models. R package version 2.4. 1, 2020.

Kert Viele, Scott Berry, Beat Neuenschwander, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1): 41–54, 2014.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113 (523):1228–1242, 2018.

Stephen Walker and Nils Lid Hjort. On Bayesian consistency. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(4):811–821, 2001.

Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12), 2010.

Suparna Wedam, Lola Fashoyin-Aje, Erik Bloomquist, Shenghui Tang, Rajeshwari Sridhara, Kirsten B Goldberg, Marc R Theoret, Laleh Amiri-Kordestani, Richard Pazdur, and Julia A Beaver. FDA approval summary: Palbociclib for male patients with metastatic breast cancer. *Clinical Cancer Research*, 26(6):1208–1212, 2020.

Nicky J Welton, Alexander J Sutton, Nicola Cooper, Keith R Abrams, and AE Ades. *Evidence synthesis for decision making in healthcare*, volume 132. John Wiley & Sons, 2012.

Elizabeth J Williamson, Andrew Forbes, and Ian R White. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine*, 33(5):721–737, 2014.

Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C and R. *Journal of Statistical Software*, 77 (i01), 2017.

Lili Wu and Shu Yang. Integrative R -learner of heterogeneous treatment effects combining experimental and observational studies. In *Conference on Causal Learning and Reasoning*, pages 904–926. PMLR, 2022.

Yongling Xiao, Michal Abrahamowicz, and Erica EM Moodie. Accuracy of conventional and marginal structural cox model estimators: a simulation study. *The International Journal of Biostatistics*, 6(2), 2010.

Yan Xie, Evan Xu, Benjamin Bowe, and Ziyad Al-Aly. Long-term cardiovascular outcomes of COVID-19. *Nature Medicine*, 28(3):583–590, 2022.

Li Yang and Anastasios A Tsiatis. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.

Shu Yang. IntegrativeHTEcf. `https://github.com/shuyang-stat/IntegrativeHTEcf`, 2020.

Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115:1540–1554, 2020.

Shu Yang and Chenyin Gao. ElasticIntegrative. `https://github.com/Gaochenyin/ElasticIntegrative`, 2023.

Shu Yang, Donglin Zeng, and Xiaofei Wang. Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. *arXiv preprint arXiv:2007.12922*, 2020.

Shu Yang, Chenyin Gao, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3): 575–596, 2023.

Jessica G Young and Eric J Tchetgen Tchetgen. Simulation from a known cox msm using standard parametric models for the g-formula. *Statistics in Medicine*, 33 (6):1001–1014, 2014.

Jessica G Young, Mats J Stensrud, Eric J Tchetgen Tchetgen, and Miguel A Hernán. A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine*, 39(8):1199–1236, 2020.

Yongqi Zhong, Edward H. Kennedy, Lisa M. Bodnar, and Ashley I. Naimi. AIPW: An R package for augmented inverse probability weighted estimation of average causal effects. *American Journal of Epidemiology*, 190:2690–2699, 12 2021.