UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title

Applications of the Causal Roadmap to Danish registry data: case studies of the effects of second-line diabetes medications on dementia and adverse cardiac events

Permalink

https://escholarship.org/uc/item/43z810gp

Author

Nance, Nerissa

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Applications of the Causal Roadmap to Danish registry data: case studies of the effects of second-line diabetes medications on dementia and adverse cardiac events

by

Nerissa Nance

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Epidemiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Maya Petersen, Chair
Professor Mark van der Laan
Professor Jennifer Ahern
Dr. Andrew Mertens

Summer 2024

Abstract

Applications of the Causal Roadmap to Danish registry data: case studies of the effects of second-line diabetes medications on dementia and adverse cardiac events

by

Nerissa Nance

Doctor of Philosophy in Epidemiology

University of California, Berkeley

Professor Maya Petersen, Chair

The Causal Roadmap is a formal framework for causal and statistical inference that supports clear specification of the causal question, interpretable and transparent statement of required causal assumptions, robust inference, and optimal precision. The Roadmap is thus particularly well-suited to evaluating longitudinal causal effects using large scale registries; however, application of the Roadmap to registry data also introduces particular challenges. In this work, Ioutline the use of simulations for estimation specification in real-world data problems. I further motivate this through two detailed case studies of the Causal Roadmap applied to longitudinal Danish National Registry data to evaluate the comparative effectiveness of second-line diabetes drugs on dementia and cardiovascular risk, respectively. I provide practical guidance on the implementation of the Roadmap using registry data, and highlight how rare exposures and outcomes over long-term follow up can raise challenges for flexible and robust estimators, even in the context of the large sample sizes provided by the registry. Moreover, I demonstrate how simulations can be used to help address these challenges by supporting careful estimator pre-specification. To protect research integrity, it is essential that the algorithm for statistical estimation and inference be pre-specified prior to conducting any effectiveness analyses. However, it is often unclear which algorithm will perform optimally for the real-data application. Here, I call for the use of simulations that realistically reflect the application, including key characteristics such as strong confounding and dependent or missing outcomes, to objectively compare candidate estimators and facilitate flexible yet full specification of the statistical analysis plan. Such simulations are informed by the Causal Roadmap and conducted after data collection but prior to effect estimation.

In Chapter 1, I outline the benefits simulations for pre-specification through two worked examples. In Chapter 2, I examine the effect of glucagon-like receptor-1 agonists (GLP1) on dementia onset, and utilize simulations to rigorously pre-specify our estimation selection. In Chapter 3, I walk through a guided simulation study using the example of GLP1 and adverse cardiac events, highlighting both strengths of this approach and areas of further research. Overall, this work shows how realistic simulations empower us to flexibly pre-specify an estimation approach and thus to improve the rigor and reproducibility of our research.

Contents

Contents **i**

1 The Causal Roadmap and simulations to improve the rigor and reproducibility of real-data applications **5**
1.1 Introduction 6
1.2 On simulations in epidemiology 12
1.3 Defining the estimation problem with the Roadmap 13
1.4 Simulations to inform the real-data analysis 14
1.5 Fostering transparent and reproducible research 23
1.6 Discussion 24
1.7 Supplemental materials 44

2 Applying the Causal Roadmap to longitudinal national Danish registry data: a case study of second-line diabetes medication and dementia **44**
2.1 Introduction 45
2.2 Case study of the longitudinal Causal Roadmap: evaluating the long-term effect of GLP-1RA use on incident dementia risk using the Danish National Registry 48
2.3 Statistical model and estimator 55
2.4 Results 59
2.5 Discussion 66
2.6 Supplemental materials 68

3 Simulations to pre-specify estimators in causal analyses: an illustration of considerations using a diabetes medication and cardiovascular outcomes from registry data **86**
3.1 Introduction 87
3.2 Case study 88

3.3 Designing an applied simulation study 89
3.4 Discussion 102
3.5 Supplemental materials 104

Bibliography **110**

Acknowledgments

So many people to thank, it is hard to know where to begin. Thank you of course to Maya Petersen, for the guidance and mentorship throughout the PhD. I feel incredibly grateful to have had such an opportunity to learn from you. Thank you to Laura Balzer, for all the support and mentorship (particularly with the first chapter of this work, but also generally in research and life). Thank you to my committee members, Mark, Jen and Andrew, for all their helpful comments and encouragement. A big thank you also to my mentors and colleagues at the University of Copenhagen—particularly Drs. Thomas Gerds and Christian Torp-Pedersen. Thanks also to Dr. Kajsa Kvist at Novo Nordisk for her support of this work.

I'm also grateful to the mentors I met through Kaiser Permanente Northern California Division of Research. Thank you to Romain Neugebauer and the analyst team for being part of my academic and personal journey. Thank you to Alyce Adams for her support with the qualifying exam and for fostering my interest in diabetes translational research. Thank you to Lyndsay Avalos for supporting my application to the PhD.

I am extremely grateful to my family, extended and chosen—you know who you are, and thank you. Thanks especially to Abigail Gutmann-Gonzalez and Caitlin Turner for their diligent copyediting and cheerleading.

Preface

Pharmacoepidemiology applies epidemiologic analytic methods to the study of drug usage patterns, effects, and interactions in a "real world" setting.[1] Randomized controlled trials (RCTs), while necessary for establishing efficacy, are conducted in highly controlled environments that do not reflect real-world use (in terms of polypharmacy, adherence, etc.). They are often also conducted in populations that are not representative of the general population (in terms of risk profile, demographics, etc.). These facets affect the generalizability of results to practical settings. Moreover, RCTs can sometimes be cost-prohibitive and challenging to power in the cases of rare disease and/or disease with long-term follow-up. They may also be ethically infeasible when there is no equipoise (i.e., benefit or harm has already been shown). All these shortcomings necessitate the use of observational data. The pharmaceutical industry recognizes the value of observational research, and most large pharmaceutical companies dedicate significant resources to real-world evidence (RWE) research. Regulatory agencies like the Federal Drug Administration (FDA) and the European Medicines Agency (EMA) recognize both the value of and the methodological challenges inherent to RWE.[2, 3] Once a drug is released on the market, vast amounts of data are passively collected about its use through medical records, registries, and claims. This observational data is often studied to determine potential unintended benefits and harms—both health-related and economic—associated with the drug.[1]

Within the field of pharmacoepidemiologic research with RWE, there is a push to improve the methodological rigor with which studies are conducted. For example, the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance released updated methodological guidance in 2023 explicitly calling for the use of causal inference to increase the validity and transparency of pharmacoepidemiologic studies.[4] Machine learning is also of increasing interest in the RWE community. These advanced analytical techniques can handle large datasets and complex variable interactions more effectively than traditional statistical methods, and can produce highly accurate and efficient predictions. However, these machine learning algorithms optimize precision for a given prediction problem. High prediction accuracy, while

important, is only part of the puzzle in causal inference. We want to not only produce accurate estimates, but also inference about our target estimand of interest including counterfactual quantities of interest. Targeted learning allows researchers to utilize the power of machine learning while maintaining statistically valid inference of target causal quantities.[5, 6, 7]

While rigorous pharmacoepidemiologic methods can be applied to many types of problems, chronic disease treatments serve as particularly interesting case studies due to the need for prolonged intervention and monitoring. Type 2 diabetes mellitus, a cardio-metabolic disease that affects almost a half a billion people globally, is one example of this.[8] Many treated patients will use second-line therapies to replace or supplement first-line therapies as some point in their lives due to poor blood glucose regulation from first-line therapy alone. Adequate glucose regulation can help reduce the risk of developing other comorbidities, such as dementia and cardiovascular disease (CVD).

Diabetes is a key risk factor for vascular dementia and Alzheimer's disease,[9] and a growing body of evidence suggests that some antidiabetic therapies may reduce cognitive decline in diabetic patients. The neuroprotective effects of the drug class glucagon-like peptide-1 receptor agonists (GLP1) are an area of active research[10, 11], including mechanisms of action that improve feelings of satiety and subsequent weight reduction, as well as a direct effect on neuroplasticity.[12, 13] Secondary analyses of randomized placebo-controlled trial data in diabetic patients also support these findings.[14, 15] However, while there is continued interest in research studying the relationship between GLP1 and cognitive function, these findings have not yet shifted clinical recommendations for the prescription of GLP1s among older adults at risk of cognitive decline.

Diabetes is also a key risk factor for CVD, and several randomized trials have shown strong evidence that GLP1 can lower the risk of major adverse cardiovascular events.[16, 17, 18] Evidence is so strong, in fact, that the American Diabetes Association has included GLP1s in their treatment recommendations for patients with existing CVD or for those at high risk of CVD.[19] Likewise, the European Society of Cardiology recommends the prescription of GLP1s to reduce the risk of heart attack and stroke in diabetes patients with existing CVD, irrespective of glucose control.[20] Notably, however, those who have less than high risk of CVD are excluded from these recommendations.

While we know that some trials have shown promising effects of second-line therapies—specifically GLP1—on cardiovascular and dementia outcomes among diabetic patients, trials are not without their limitations. In the case of dementia, trials rely on proximal biomarkers rather than dementia onset. For both dementia and CVD, recruited patients are higher risk than the general population. In addi-

tion, most of these trials have been placebo-controlled studies of single drugs rather than comparisons of different treatment options. Moreover, these trials have not been conducted under realistic, real-world conditions with a broad population. More evidence is needed, but prospective trials with rarer outcomes requiring longer-term follow up can be challenging to adequately power. Observational studies circumvent these shortcomings by allowing access to expansive datasets with sufficient sample sizes to investigate relatively rare medications and outcomes, such as in the case of dementia. Several studies have examined these relationships using observational data[21, 22], but no study has yet done so utilizing causal inference methods, and in particular the Causal Roadmap, to circumvent the biases unique to longitudinal observational data.

Observational studies can suffer from confounding bias that is particularly challenging in longitudinal settings. One example is time-dependent confounding, whereby confounders affected by prior exposure can be on the causal pathway between exposure and outcome[23], as is often the case in medication treatment. Modern causal inference methods have the potential to address these shortcomings by allowing researchers to adjust estimates for these longitudinal covariates that are affected by prior exposure without adjusting for covariates that are potential mediators or colliders between prior treatment and the outcome.

Complex longitudinal estimates present their own analytic problems. Sustained exposure over time can reduce the data support for the target comparison of interest. One way to help ensure not only that data support exists but also that you are achieving desired metrics is through simulation. Simulations can be used in RWE to improve the rigor and transparency of real-data applications.

This body of work aims to illustrate the use of the Causal Roadmap and simulations to improve the generation of RWE in complex settings. It offers several key contributions to the literature of applied pharmacoepidemiology and diabetes research. It produces guidance, through example, of applied simulations for estimation specification. These case studies provide guidance to applied epidemiologists and applied biostatisticians in conducting longitudinal causal inference analyses. It is the first to isolate and compare second-line regimens—specifically, GLP1 receptor agonists and sodium-glucose cotransporter-2 (SGLT2) inhibitors—head-to-head to evaluate effects dementia outcomes using causal methodology. It is also among the first to harness the Danish health registries to answer causal questions using the Causal Roadmap.

This work is organized as follows. In Chapter 1, I provide motivation for implementing simulations to pre-specify estimator selection through statistical analysis plans using two motivating data examples. In Chapter 2, I outline an application of the Causal Roadmap to Danish registry data to examine the relationship between

GLP1 use and dementia using simulations to pre-specify estimation selection. In Chapter 3, I provide a step-by-step implementation of a simulation, using a case study of GLP1 and cardiovascular disease. Finally, I conclude with lessons learned and future directions for this work.

Chapter 1

The Causal Roadmap and simulations to improve the rigor and reproducibility of real-data applications[1]

[1]This article has been accepted in a similar form for publication in the journal *Epidemiology*, with coauthors Maya Petersen, Mark van der Laan, and Laura Balzer. Link pending.

1.1 Introduction

Formal frameworks for causal and statistical inference can help researchers to clearly structure and understand the links between their research question, causal model, data, statistical estimation, and results interpretation. Examples of such frameworks include the Causal Roadmap and target trial emulation.[24, 6, 25] Recent commentaries on epidemiologic training have highlighted the role of such frameworks in asking thoughtful and feasible study questions, particularly amid a proliferation of novel analytic methods which may aid or distract from answering that question.[26, 27] Even after we have specified a well-defined and relevant question, there are many steps to setting up the analysis to answer it. For example, the remaining steps of the Causal Roadmap (hereafter, "the Roadmap") are to: (2) specify a causal model reflecting background knowledge and uncertainties; (3) define the causal effect of interest; (4) describe the data available to answer the question; (5) assess identifiability; (6) select a statistical model and estimand; (7) estimate and obtain inference; and (8) interpret results.

Roadmap steps 1-6 set up a statistical estimation problem, reflecting our research question and the real-world challenges of the data. Specifically, the Roadmap leads us to a well-defined statistical estimand, which is a function of the observed data distribution, and a realistic statistical model. (Formally, the statistical model is the set of all possible observed data distributions.[6]) However, the Roadmap does *not* tell us which algorithm to apply for estimation and inference. While an algorithm's theoretical properties can narrow the scope of possibilities, it is often unclear *a priori* which approach will perform best in the real-data application. Instead, there is a tendency to simply apply one's preferred algorithm, using the default settings of a computing package or recycling prior code. Likewise, there is a temptation to try several implementations and pick the implementation yielding the most favorable or logical result. As detailed below, we advocate for the use of realistic simulations to objectively select the algorithm for estimation and inference and pre-specify the statistical analysis plan.

The statistical analysis plan delineates key features of the real-data analysis, including the target population, primary outcome, exposure conditions, causal effect of interest, approaches to handling potential inferential threats (e.g., confounding and missing data), statistical estimand, primary/secondary analyses, and sensitivity analyses. While many of these features follow from earlier steps of the Roadmap, the statistical analysis plan requires us to state the precise implementation of the estimator, including approaches for estimating nuisance parameters and approaches for obtaining inference. (Nuisance parameters are quantities needed to evaluate the statistical estimand, but are not the estimand itself.) Full pre-specification requires more

than simply stating the statistical estimand (e.g., the longitudinal G-computation formula) and the general class of estimators (e.g., targeted minimum loss-based estimation [TMLE]). Indeed, the process of pre-specifying a statistical analysis plan requires us to think critically about different estimation and inferential strategies as well as their expected performance *before* running any analyses to assess causality. As a result, pre-specification of the statistical analysis plan helps improve transparency and protect against *ad hoc* analyses, which can lead to a "fishing expedition" to find the most promising results and inflated Type-I error rates.

Regulatory and funding agencies typically require the statistical analysis plan to be pre-specified prior to conducting effectiveness analyses in randomized trials.[28, 29, 30, 31] There is also a growing movement to improve reproducibility and transparency of observational studies through rigorous planning and reporting (e.g., [32, 33, 34, 35]). Mathur and Fox provide an excellent review on the principles and practices to improve open and reproducible research in epidemiology; in particular, they highlight pre-registration of statistical analysis plans for observational studies and code sharing.[36] Notably, Gruber et al. discuss how the Roadmap for Targeted Learning can inform the development of statistical analysis plans.[37] Here, we build on this work by providing guidance and context on *how* to select the approach for statistical estimation and inference for the real-data application.

Our goal is to describe how finite sample simulations, informed by the Roadmap and reflecting the real-data application, can be used to objectively compare estimation strategies and develop a completely pre-specified statistical analysis plan. For demonstration, we provide two worked examples: (1) an observational study with a time-varying exposure and censoring, and (2) a randomized trial with missing and dependent outcomes. We also highlight how this approach naturally leads to fully pre-specified and quality-checked computing code. Thus, our approach has the potential to improve the transparency, reproducibility, and rigor of our analyses aiming to evaluate causal effects.

Our presentation assumes familiarity with foundational concepts in causal and statistical inference (e.g., causal models, identifiability assumptions, and the G-computation formula). For a review of these concepts and an introduction to the Causal Roadmap, we refer to Petersen and van der Laan[24] and Dang et al.[38]. An overview of the Roadmap for the running examples is provided in Table 1.1. Indeed, our worked examples are inspired by real studies and are inherently complex, highlighting the real-world challenges that commonly arise when aiming to infer causality. The remainder of the article is organized as follows. First, we outline how simulations are used in epidemiology. Then, we demonstrate the utility of the Roadmap in setting up the statistical estimation problem and designing the simulation study. Next, we describe how to conduct simulations for estimator selection in

real-data applications; specifically, we discuss pre-specification of the candidate estimators, data generating process, performance metrics, and selection scheme. Finally, we describe the consequences of our approach to improve research transparency and reproducibility.

Table 1.1: Overview of the Causal Roadmap for the two running examples

Causal Roadmap steps	**Observational study example**	**Randomized trial example**
(1) Research question • Specify the primary exposure(s), outcome, and target population	What is the effect of SGLT2[a] inhibitor use on the risk of renal disease onset among patients with diabetes in an integrated healthcare system in the United States?	What is the effect of the multicomponent, SEARCH-Youth intervention on viral suppression among 15-24 year-olds with HIV in rural Kenya and Uganda?[39]
(2) Causal model • Describe the confounding structure, missing data mechanism, (in)dependence structure, ...	Longitudinal model including time-varying confounders and factors influencing censoring	Multi-level model reflecting that the intervention is randomized at the clinic-level, but outcomes are at the individual-level and subject to missingness
(3) Causal effect • Using counterfactuals, specify the target effect and scale of interest	Causal risk difference: $\mathbb{E}[Y_{\bar{a}=1,\bar{c}=0}(t)] - \mathbb{E}[Y_{\bar{a}=0,\bar{c}=0}(t)]$ with $Y_{\bar{a},\bar{c}}(t)$ as the counterfactual outcome at time t under SLGT2 use $\bar{A} = \bar{a}$ and no censoring $\bar{C} = 0$ throughout follow-up	Sample prevalence ratio: $\frac{1/N\sum_i Y_i^c(1)}{1/N\sum_i Y_i^c(0)}$ with $Y_i^c(a)$ as the counterfactual % with viral suppression in clinic $i = \{1, \ldots, N\}$ under treatment-level $A^c = a^c$ and complete outcome measurement

(4) Observed data • Describe the data that have been or will be observed, include the exposure(s), outcome(s), ...	Longitudinal data structure: history of covariates $\bar{L}(t)$, exposure $\bar{A}(t)$, censoring $\bar{C}(t)$, and outcomes $\bar{Y}(t)$ through time t	Multi-level data structure: cluster- and individual-level baseline covariates (E^c, W), cluster-level exposure A^c, and individual-level, post-baseline covariates M, measurement indicator Δ, and outcome Y
(5) Identifiability • Evaluate the assumptions needed to express the causal effect as statistical estimand	No unmeasured confounding, sufficient data support (i.e. no practical positivity violations), and censoring-at-random	Within each clinic and values of adjustment variables, individual-level outcomes are missing-at-random and there is a positive probability of measurement
(6) Define the statistical estimation problem • Specify the statistical estimand and model	The statistical estimand is the iterated conditional expectation expression of the longitudinal g-formula.[23, 40] The statistical model is non-parametric.[b]	Within each clinic, the cluster-level endpoint, accounting for missingness, is $Y^c = \mathbb{E}[\mathbb{E}(Y\|\Delta = 1, W, M)]$. The statistical estimand to evaluate the intervention effect is $\Psi = \frac{1/N \sum_i \mathbb{E}(Y^c\|A^c=1,E_i^c)}{1/N \sum_i \mathbb{E}(Y^c\|A^c=0,E_i^c)}$.[41] The statistical model is semi-parametric.[b]
(7) Estimate and obtain inference • In a pre-specified way, chose and implement an estimator	Outcome-blind simulations to objectively select between alternative implementations of longitudinal TMLE[c]; see Table 1.3.[7, 42]	Treatment-blind simulations to objectively select between alternative implementations of Two-Stage TMLE[c]; see Table 1.3.[41]

(8) Interpret • State the results in light of the causal and statistical assumptions.	After flexibly accounting for measured time-varying confounders, two years of continuous use of SGLT2[a] inhibitors was associated with a 5% (95%CI: 2.75-7.25%) decrease in the risk of renal disease onset.	After flexibly accounting for missing outcomes and clustering, SEARCH-Youth increased viral suppression among adolescents and young adults with HIV by 10% (risk ratio=1.10, 95%CI: 1.03-1.16).[39]

[a]Sodium-glucose cotransporter-2 inhibitors
[b]See the Statistical Analysis Plan in the eAppendices for additional details.
[c]Targeted minimum loss-based estimation

1.2 On simulations in epidemiology

Simulation studies are widely applied in methodologic research to evaluate the finite sample properties of existing and recently developed estimators.[43] Since the true value of the target parameter is known, simulations enable us to calculate performance metrics, such as bias and confidence interval coverage . For example, the well-known Kang and Schafer censored data simulations revealed the instability of estimating equation-based methods under data sparsity and inspired suspicion of doubly robust approaches.[44] Subsequent replication of these simulations have highlighted the potential for doubly robust, substitution estimators (e.g., TMLE and collaborative TMLE) to overcome these challenges.[45] More recently, simulations have been used to illustrate the potential advantages and perils of using machine learning in analyses seeking to infer causality.[46, 47, 48]

Beyond methods evaluation, epidemiology uses simulations in several other settings. Examples include teaching epidemiologic concepts, evaluating study designs, forecasting disease trajectories, agent-based modeling, addressing transportability, and data pooling (e.g., [49, 50, 51, 52, 53, 54, 55]). Prior to data collection, simulation studies are commonly implemented to inform the design randomized trials, including power calculations.[56, 57, 58]) Following data collection and after effect estimation, simulations are also applied in sensitivity analyses, including quantitative bias analysis (e.g., [59, 60]). However, to the best of our knowledge, there are few published examples of using simulations in the principled comparison and selection of estimators for a real-data analysis *after* data have been collected *but* before any effectiveness analyses are conducted. Some exceptions include the use of outcome-blind simulations to select the primary analysis in a SMART trial, to evaluate propensity score estimators within TMLE in a drug safety monitoring study, and to compare estimators for longitudinal effects with registry data.[61, 62, 63]

A pertinent commentary in the *British Medical Journal* called for the broader use of simulations to inform applied data analyses, but also recognized that the implementation and reporting of such studies is the subject of continued debate.[64] We aim to help address these and other issues by guiding researchers on the use simulations, informed by the Roadmap and reflecting the real-data application, to aid in the development and full pre-specification of the statistical analysis plan and corresponding computing code.

1.3 Defining the estimation problem with the Roadmap

As illustrated in Table 1.1, the first six steps of the Roadmap setup the statistical estimation problem, which is defined by the statistical estimand and the statistical model.[24] Of course, one could specify these elements without the Roadmap. In our experience, however, applying the Roadmap has several strengths relative to other frameworks and the following benefits.[27] Among others, the Roadmap helps clarify the research goals, highlight potential inferential threats, specify the handling of events occurring after the initial exposure or treatment, and facilitate transparent discussions about the plausibility of assumptions. Perhaps most crucially, the Roadmap leads to a statistical estimand reflecting our original research question as well as the real-world challenges in the data. In other words, even if the identifiability assumptions do not hold, the Roadmap guides us to statistical estimand coming as close as possible to the wished-for effect. (The size of the "causal gap" can be formally explored in sensitivity analyses and is taken into account during interpretation.[27]) In most cases, our statistical estimand is a complicated function of the observed data distribution and not equal to a single coefficient in a parametric regression. This complexity is needed to generate the most appropriate answer to our research question and often precludes the use of more traditional statistical approaches. Equally important, the Roadmap highlights that we rarely have the knowledge to support functional form assumptions — beyond treatment randomization in a trial. Instead, our statistical model is often non-parametric or semi-parametric, and we need to harness machine learning during estimation to avoid unsubstantiated assumptions.

As a concrete example, consider a study aiming to evaluate the effect of a time-varying and non-randomized exposure: sustained use of sodium-glucose cotransporter 2 (SGLT2) inhibitors on the onset of renal disease among patients with diabetes. As shown in Table 1.1, application of the Roadmap highlights the potential for bias and misleading inference due to confounding, censoring, and practical violations of the positivity assumption, occurring when there is insufficient variability in the exposure within confounder strata.[65, 66] These inferential threats can be particularly fraught in settings with longitudinal exposures; the longer follow-up time, the more potential there is for time-dependent confounding, right-censoring, and lower support for the longitudinal exposures of interest. Given these challenges, the Roadmap leads to a complex statistical estimand: a contrast of the iterated conditional expectation expression of the longitudinal G-computation formula (Supplementary Material A).[23, 40] Importantly, the Roadmap also leads to a non-parametric statistical model without functional form assumptions. Altogether, the

Roadmap narrows the scope of possible estimators to algorithms that can handle time-dependent confounding, right-censoring, and poor data support, as well as harness machine learning to avoid unsubstantiated modeling assumptions. For this setting, common approaches include singly robust estimators, such as inverse probability weighting and G-computation, as well as doubly robust alternatives, such as augmented inverse probability weighting and TMLE.[67, 68, 23, 69, 6] Each has statistical properties that may lend themselves (or not) to a specific analysis. As described below, we can use simulations, informed by the Roadmap, to chose the estimator expected to perform best in the actual analysis.

As a second example, consider the SEARCH-Youth study, a cluster randomized trial to evaluate the effect of a multi-component intervention on viral suppression among youth with HIV in East Africa.[39] As shown in Table 1.1, the Roadmap highlights the impacts of randomizing the treatment to health clinics (instead of individuals) and missing data. Specifically, each Roadmap step reflects the dependence between participants within clinics as well as the potential biases from the missing data–equivalent to time-dependent confounding. Again, the Roadmap leads to a complex statistical estimand: a contrast of clinical-level summary measures, each accounting for baseline and post-baseline causes of measurement and outcomes (Supplementary Material B).[41, 70, 71] The Roadmap also leads us to a semi-parametric statistical model, only reflecting our knowledge of treatment randomization. Here, the Roadmap narrows the set of possible estimators to those that flexibly handle dependent and missing data — specifically, approaches allowing the missingness mechanism to vary by cluster.[41] In Two-Stage TMLE, for example, we first estimate a summary measure accounting for missing data in cluster separately and then evaluate the intervention effect on those cluster-level summaries. Additionally, as common in cluster randomized trials,[72] few clinics were randomized, specifically 28, in SEARCH-Youth. Therefore, we also need an estimation and inferential approach that performs well with few independent units. Again, simulations can aid in the formal evaluation of alternatives and pre-specification of the primary analysis.

1.4 Simulations to inform the real-data analysis

We now detail how simulations, informed by the Roadmap and reflecting the real-data application, can aid in objectively selecting and appropriately implementing the estimation and inferential approach expected to perform best in the real-data analysis. To do so, we need to pre-specify the candidate estimators, data generating process for the simulations, performance metrics and selection process.

Choosing the candidate estimators

It is essential to choose candidates targeting the statistical estimand of interest. This may seem obvious, but without careful consideration, we could end up comparing estimators of marginal versus conditional effects, especially in hierarchical data settings.[73, 74, 42, 70] As previously discussed, following the Roadmap narrows the set of candidate algorithms to those targeting the statistical estimand. This set can further be narrowed by considering the asymptotic properties of the estimators (e.g., efficiency, double robustness). Even if we settle on a single class of estimators, such as TMLE, there are still many decisions before the statistical analysis plan is fully specified.

We must decide *how* to estimate nuisance parameters. In doubly robust estimators, for example, nuisance parameters typically include the outcome regressions (i.e., the conditional expectation of the outcome given past exposure/measurement and covariates) and propensity scores (i.e., the conditional probability of exposure/measurement given the past). To respect our statistical model, machine learning is often required for flexible, data-adaptive estimation of nuisance parameters. However, application of machine learning requires additional choices. For example, in the ensemble algorithm Super Learner, we need to specify the candidate learners (including their tuning parameters), the cross-validation scheme, and the loss function.[75, 76] After obtaining initial estimates of the nuisance parameters, there may be additional decisions. For example, with practical positivity violations, we can decide to truncate the estimated propensity scores at various levels.[65, 77] Finally, there are a variety of options for statistical inference. For TMLE, for example, some approaches for variance estimation are the non-parametric bootstrap, standard or cross-validated estimates of influence curve, plug-in estimation of the variance, or other doubly robust options.[6, 78, 79, 80, 7]

For our running examples, Table 1.3 provides an overview of the candidate approaches that were pre-specified for objective comparison in simulations. In the observational study, the candidate algorithms were longitudinal TMLE with various implementations. For nuisance parameter estimation, Super Learner with and without covariate screening as well as bounded or unbounded estimates of the propensity score were considered. For statistical inference, candidates included Wald-Type 95% confidence intervals with variance estimated by the influence curve or the non-parametric bootstrap. For the cluster randomized trial, we limited the candidate algorithms to Two-Stage TMLE with the following specifications. For estimation of the cluster-level endpoints accounting for missing outcomes, candidates were TMLE using Super Learner, TMLE using parametric regressions, and the empirical mean among those measured. For estimation of the intervention effect, candidates were

TMLE with various approaches to covariate adjustment for precision gains.[78, 81]. Finally, candidates for variance estimation included standard or cross-validated estimates of the influence curve. We now discuss how to define the data generating process for the simulation to formally evaluate the performance of these candidates.

Table 1.3: Overview of simulation setup and results for the two running examples.

Simulation specifications & results	**Observational study example**	**Randomized trial example**
Statistical estimation problem[a]	The longitudinal G-computation formula under practical positivity violations with a long-term rare exposure, right-censoring, and rare outcome + non-parametric statistical model	Two-Stage estimand with (1) differential missingness of individual-level outcomes; (2) few clusters randomized; and (3) the cluster as the independent unit + semi-parametric statistical model
Estimators compared[b]	Longitudinal TMLE with alternative approaches for • Nuisance parameter estimation using Super Learner with/without covariate pre-screening and with/without truncation of the estimated propensity scores • Variance estimation: influence curve vs. non-parametric bootstrap	Two-Stage TMLE with alternative approaches to (1) account for missing individual-level outcomes: the empirical mean, TMLE with main terms regression, or TMLE with Super Learner (2) adaptively adjust to improve precision: Adaptive Pre-specification with a limited vs. expanded adjustment set (3) obtain inference with standard or cross-validated estimates of the influence curve [78, 41, 81]

Data generation process	Outcome-blind simulations preserving the baseline covariate structure, while simulating exposure-censoring variables with positivity challenges and a synthetic outcome with a similar marginal distribution as the real-data	Treatment-blind simulations preserving the covariate-outcome data structure but randomly permuting the treatment indicator + generation of outcome measurement indicators by an independent statistician
Performance metrics and selection approach[b]	Over 1000 iterations, (i) Select the approach for nuisance parameter estimation that minimizes empirical variance and preserves Oracle coverage (ii) Given (i), select the approach for variance estimation that minimizes the estimated variance and preserves 95% confidence interval coverage.	Over 1000 iterations, (i) Select the approach for the cluster-level endpoints resulting in nominal confidence interval coverage for those endpoints and Type-I error control for the overall effect. (ii) Given (i), select the approach for effect estimation and variance estimation resulting in optimal Type-I error control
Resulting primary analysis[b]	Longitudinal TMLE with Super Learner with algorithm prescreening and without propensity score truncation, with influence curve-based variance estimation.[7, 42, 82]	(1) TMLE with Super Learner to flexibly estimate the cluster-level endpoints, (2) TMLE with Adaptive Pre-specification with limited candidates to efficiently estimate the intervention effect; (3) variance estimation with the cluster-level, influence curve

[a]See Table 1.1 for details. [b]See the SAPs in the eAppendices for details, including the precise descriptions of the candidate estimators (e.g., the library, loss-function, and cross-validation scheme for Super Learner), performance metrics, and the primary analysis. [c]Targeted minimum loss-based estimation

Defining the data generation process for the simulation

Thus far, the Roadmap has aided in defining the statistical estimation problem and specifying the set of candidate algorithms for estimation and inference. Simulations reflecting the real-data application can facilitate objective comparison and selection between these candidates *if* we choose a data generating process that is close to the real one. Concretely, application of the Roadmap highlighted several potential biases and inferential threats in the running examples (Table 1.1). For the observational study, we need to design a simulation with, at minimum, the same exposure/confounder/censoring structure and, therefore, the same practical positivity challenges as the real-data. For the randomized trial, we need to design a simulation with, at minimum, the same number of clusters, a similar distribution of participants per cluster, and plausible missing data mechanism as the real-data. Given these specifications, several options exist.

Monte Carlo simulations, where we repeatedly sample from a known data generating process, are common and traditionally employ parametric models for data generation.[83] Such parametric models often fail to reflect the complexities of the real-data, especially in longitudinal or clustered data settings. Considering the limitations of fully parametric simulations, plasmode simulations have gained popularity and may be particularly useful for our focus: estimator selection *after* data have been collected *but before* effect estimation. Plasmode simulations, as defined here, encompass a range of semi-parametric methods that sample partially from the empirical data distribution, while allowing for some user-specification.[65, 84]

There are various types of plasmode simulations. In "outcome-blind" plasmode simulations, we preserve the relationships between the baseline covariates, while simulating the outcome (and other variables) through parametric or semi-parametric methods (e.g., [61, 85, 62, 86]). In these simulations, the value of the (simulated) effect is known, but we remain blinded to the true exposure-outcome relationship. As described in Table 1.3 and Supplementary Material A, outcome-blind simulations were conducted in the observational study by resampling the baseline covariates from the empirical distribution and then applying highly adaptive LASSO to simulate the longitudinal exposures, censoring, time-varying covariates, and outcome.[87] This approach preserves the complex relationships between baseline covariates, while generating the remaining variables to reflect challenges in the real-data application (e.g., poor data support due to the rare exposure, long-term follow-up, and strong confounding).

"Treatment-blind" simulations are another plasmode simulation technique where the covariate-outcome data are preserved but the treatment indicator is randomly permuted.[81] As detailed below, treatment-blind simulations are particularly rele-

vant for evaluating Type-I error control, because the null hypothesis is true by design. As outlined in Table 1.3 and Supplementary Material B, such simulations were implemented in the trial example by randomly shuffling the treatment indicator and imposing missingness on outcomes through a measurement indicator, which was generated by an independent statistician and as a function of the baseline cluster-level and individual-level covariates, the permuted treatment indicator, and time-varying covariates. This simulation approach preserves the covariates and underlying outcome, while facilitating a rigorous comparison of alternative approaches and their potential to reduce bias due to differential outcome measurement and improve efficiency through covariate adjustment, as described next.

Specifying the performance metrics and selection approach

Once we have the set of candidate estimators and data generating process for the simulations, we need to pre-specify the performance metrics and process to objectively compare the candidates. In Table 1.5, we review some common metrics, such as the bias and variance of the point estimates as well as 95% confidence interval coverage (i.e., the proportion of calculated confidence intervals that contain the true effect). To compare estimators in a way that is agnostic to the variance estimator and evaluate the extent to which an estimator's bias is negligible, we can use "Oracle coverage", where the 95% confidence intervals are calculated using the variance of the point estimates across the simulation iterations, instead of the estimated variance. In simulations to inform randomized trials, common metrics include statistical power (i.e., the proportion of times the false null hypothesis is rejected) and Type-I error control (i.e., the proportion of times the true null hypothesis is rejected). We may additionally be interested in estimating the potential savings in sample size to achieve the same power.[88, 85, 81] Finally, we pre-specify the selection process for objectively choosing the best-performing candidate and, thereby, the primary analytic approach.

Table 1.5: Examples of performance metrics to use in simulations to inform real-data analyses.

Performance metric	**Calculation**	**Relevance**
Bias (point estimates)	The average difference between the point estimate and the target effect across the simulation iterations	What is the accuracy of the estimator?
Variance (point estimates)	The variance of the point estimates across the simulation iterations	How precise is the estimator?
Mean-squared error	The average of the squared differences between the point estimate and target effect (equivalent to Bias^2+variance)	What is the variability of the estimator around the target effect?[59] (Akin to asking how is the estimator balancing bias and variance?)
Bias-variance ratio	Ratio between the bias and variance	Is the estimator's bias disappearing at a fast enough rate relative to its variance?
Variance to estimated variance ratio	Ratio between the variance of the point estimates and the average variance estimate	Is the variance being over- or under-estimated?
Oracle coverage	The proportion of 95% confidence intervals, calculated with the variance of the point estimates, containing the target effect	Is the bias in the point estimates impacting the estimator's confidence interval coverage?
Confidence interval coverage	The proportion of 95% confidence intervals, each calculated with the estimated variance, containing the target effect	Does the approach for estimation and inference result in valid inference?
Power	The proportion of simulation iterations where the true null hypothesis is rejected	Will the approach for estimation and inference identify an effect when it exists?
Type I error	The proportion of simulation iterations where the true null hypothesis is rejected	Will the approach for estimation and inference lead to the incorrect conclusion of an effect when none exists?

For the running examples, Table 1.3 provides the performance metrics, selection process, and final estimator. In both studies, selection was a two-step process, implemented in `R`, and with 1000 simulation iterations. In the observational study, the optimal approach for nuisance parameter estimation was first selected to minimize the empirical variance but preserve Oracle coverage. Then given this choice, the optimal approach for inference was selected to minimize the variance estimate but preserve 95% confidence interval coverage. In the randomized trial, the optimal approach for estimating the cluster-level endpoints accounting for missing outcomes was based on attaining nominal confidence interval coverage for those endpoints and Type-I error control for the intervention effect. Then given this choice, the optimal approach for estimation and inference for the intervention effect was selected to maximize precision without sacrificing Type-I error. Estimation approaches with good, but not optimal, performance were then pre-specified as sensitivity analyses.

1.5 Fostering transparent and reproducible research

Informed by the Roadmap, we have conducted a simulation study, reflecting the real-data application and facilitating objective comparison of various approaches for estimation and inference. Specifically, we pre-specified our candidates, the data generating process, the performance measures, and the selection scheme. With this simulation study, we have a responsible and "hands-off" approach to selecting the best estimator and, thus, the primary analytic approach for our real-data application. The corresponding statistical analysis plans for the running examples are given in the Appendices and include the design and results of the simulation study. As illustrated in Table 1.1, interpretation of the study results must account for the statistical assumptions of the selected estimator as well as the plausibility of the identifiability assumptions. Altogether, our approach facilitates objective selection and implementation of the best analysis to answer our research question, while protecting research integrity by ensuring we remain blinded to the true causal effect. Importantly, our approach is in line with regulatory guidelines to update the statistical analysis plan based on a blinded review of the data.[28]

Our proposed process has several consequences for improving research transparency and reproducibility. First, the simulation leads to a fully pre-specified statistical analysis plan, where analytic decisions are clearly stated and can be critically evaluated. Second, conducting the simulation requires implementing all candidate

estimators in computing code. Thus, these simulations serve as an invaluable tool for debugging code and identifying potential issues (e.g., lack of convergence due to rare outcomes) that may arise in the real-data application. Uploading the computing code and results from both the simulation study and real-data analysis to an online repository, such as GitHub, and including detailed explanation through a markup language further improves reproducibility, trust, and open science.

1.6 Discussion

For real-data analyses, we have outlined how simulations can guide the objective selection of the optimal approach for estimation and inference and, thus, full pre-specification of the statistical analysis plan. Anchored on the Roadmap, these simulations are designed with our research question at the forefront and to explore the primary concerns of the real-data application. The results of the simulation may ultimately reveal that it is not feasible to reliably estimate the statistical estimand of interest. In such cases, we may need to return to early steps of the Roadmap and modify the research question and causal estimand to accommodate the limitations in the real-data.[65]

Further guidance is needed, and our presentation has limitations. First, we focused on plasmode simulations and did not cover alternative approaches, which might be needed if the real-data are not available or only partially available. Additionally, we emphasized the need to emulate the real-data closely, but did not discuss how to assess the quality of the emulation. It is worth noting that creating simulated data that are "too close" to the true distribution can inspire fears of data "snooping";[89] pre-specification and code-sharing can help alleviate these fears. Third, our presentation did not cover practical implementation, such as how to vary simulation parameters, determine the number of iterations, and parallelize; we refer to Morris et al.[43] for an excellent overview of these and other considerations. Additional practical details on the projects inspiring our running examples are available in Nance et al.[63] and Balzer et al.[90]. Fourth, while our worked examples incorporated common challenges, including confounding, dependence, and missing data, they did not cover other concerns like generalizability, transportability, and partial identification in detail.[91, 92, 93, 94, 95] Finally, we have presented a two-step process for estimator selection and implementation: (1) conduct a realistic simulation study to objectively compare pre-specified estimators according pre-specified metrics, and (2) implement the optimal estimator (as defined by the simulation study) for the real-data analysis. Alternative approaches, such as auto-TMLE, are being developed to dynamically evaluate estimators in simulations and implement the optimal estimator

in a single step.[96, 97]

Altogether, we believe simulations, anchored on the Roadmap, are an indispensable and under-utilized tool for the objective comparison of approaches for estimation and inference in real-data applications. They are a crucial alternative to the status quo: naively applying a preferred algorithm or trying several algorithms and selecting the "best" in an *ad hoc* manner. Instead, our approach provides a formal framework for comparative assessment of alternative strategies for estimation and inference, pre-specification of the corresponding statistical analysis plan, and generating quality-controlled computing code. Our approach strives to improve the transparency, rigor, and reproducibility of real-data analyses in epidemiology and beyond.

Supplemental Material A

Causal effects of SGLT2 inhibitor use on renal disease in a population of adults with diabetes

EXAMPLE STATISTICAL ANALYSIS PLAN
v1.1

This analysis plan is inspired by work from the Joint Initiative for Causal Inference (JICI) and follows the Roadmap for Causal and Statistical Inference.[1] The analysis that inspired this work can be found on arxiv (2310.03235).[2] Part of the example analytic approach also inspired by work with Dr. Romain Neugebauer at Kaiser Permanente Northern California Division of Research.

Research question

What is the effect of sustained exposure to SGLT2 inhibitors (a class of diabetes drug) over time on chronic kidney disease (CKD) among a cohort of patients with diabetes in an integrated health system in the US?

Causal model

We define the following endogenous nodes for time t=1,...,K+1:
X = {L(t), A(t)}
where:
K+1 is the maximum follow up time of two years. (We explicitly define the time discretization in the Observed Data section below.)

A(t)=(A_1(t), C(t))

- A_1(t) is SGLT2i use at time t
- C(t) is censoring indicator by time t

L(t)=(L_1(t),Y(t))

- L_1(0) defines the baseline covariates are at t=0 and includes age, sex, education, income, neighborhood deprivation index
- L_1(t) defines the time-varying covariates at time t and includes insulin use, cardiovascular disease (e.g., ischemic heart disease, myocardial infarction), hypertension
- Y(t) is the outcome, chronic kidney disease, by time t

And exogenous error:
$U = (U_{A(t)}, U_{L(t)})$ for all t

Structural equations are as follows:

$$L(t) = f_{L(t)}(\bar{A}(t-1), \bar{L}(t-1), U_{L(t)}),\ t = 1,\ldots, K+1$$

$A(t) = f_{A(t)}(\bar{A}(t-1), \bar{L}(t), U_{A(t)}),\ t = 1,\ldots, K$
where the overbar denotes the history of the variable

Causal effect of interest

The counterfactual difference in the risk of CKD between exposure to SGLT2 vs no exposure to SGLT2 over time

(a) Treatment regimes of interest

Our treatment regimes of interest correspond to sustained SGLT2 use ($\bar{a}_1 = 1$) throughout follow-up, compared to sustained non-use ($\bar{a}_1 = 0$) throughout follow-up. In both regimes of interest, we consider a hypothetical intervention to prevent censoring $\bar{c} = 0$.

(b) Counterfactual outcomes of interest

Our counterfactual outcomes of interest are the time to CKD under each possible longitudinal intervention to set SGLT2 use (and prevent censoring):

$$Y_{\bar{a}_1=a,\bar{c}=0}(t),\ t = 1,\ldots, K+1$$

Where $a \in A = \{0, 1\}$

(c) Target causal parameter

The primary causal estimand is the difference in the two-year, counterfactual risk of renal disease under sustained use of SGLT2 versus no use throughout follow up:

$$E[Y_{\bar{a}_1=1,\bar{c}=0}(t)] - E[Y_{\bar{a}_1=0,\bar{c}=0}(t)]$$

at time t=(K+1) corresponding to the maximum follow-up time of two years.

We are also interested in contrasting the two-year counterfactual survival curves over time. Therefore, .
we will visualize the counterfactual curves by plotting the counterfactual risk at each time point t, over time for each level of the exposure

$$E[Y_{\bar{a}_1=1,\bar{c}=0}(t)]\ , E[Y_{\bar{a}_1=0,\bar{c}=0}(t)],\ t = 1,\ldots, K+1$$

We will then contrast the areas under these curves by summing the causal risk differences across time:

$$\sum_{t=1}^{K+1} [E[Y_{\bar{a}_1=1,\bar{c}=0}(t)] - E[Y_{\bar{a}_1=0,\bar{c}=0}(t)]]$$

We will test this against the hypothesis that the differences are equal to zero using a variance estimate derived by the delta method.

Observed Data

Study population

We will collect data on a retrospective cohort of adult (age 18 years or older) patients with type II diabetes and who have initiated any second-line regimen by 2014 or later. Specifically, we will extract patient data from electronic medical records from a large integrated healthcare system in the United States, spanning across 4 sites in three different states. Patients will be excluded if they have evidence of prior CKD diagnosis. Index date for cohort entry will be defined as the date of switch to second-line regimen, among persons meeting all other inclusion criteria.

Observed Data structure

Individual level data from this retrospective cohort can be described by a general longitudinal data structure, denoted as:

$$O = (L(0),\ A(0),\ \ldots,\ L(K),\ A(K),\ L(K+1)) \sim P_0 \in M$$

where L(0) and A(0) are baseline covariate and exposure measurements, respectively (defined explicitly below). Time is discretized into 60-day intervals, and data are collected until a maximum time of t=K+1, where K+1 is the maximum follow up time of two years or twelve 60-day intervals [i.e., K=11].

A(t): Time-varying intervention variables, t=0,...,K

Our intervention variables will include the time-varying exposure of use of SGLT2 inhibitors and the time-varying censoring variable:

$$A(t) = (A_1(t),\ C(t))$$

$A_1(t)$ takes the value of 1 when on SGLT2, and 0 when not on SGLT2. This will be operationalized as any use (from one day up to 60 days) of the corresponding drug during the interval.

$C(t)$ denotes administrative end of follow up and death. This is a time-varying variable and will take the value of 0 prior to censoring, and 1 at and after the time of censoring. Administrative censoring date will be defined as December 31, 2023. Death will also be considered a censoring event, and is, thus, also included in C(t). All-cause mortality will be measured through state death certificates.

$L_1(t)$: Time-varying covariates of interest, t=0,...,K+1

Time-varying covariates will be measured at the start of each interval and defined as an indicator of presence at start of time t; each covariate takes the value of zero until evidence of the condition or medication is present in the time interval, at which point the covariate takes the value of one for the remainder of follow-up. $L_1(t)$ variables will include diagnoses and medication covariates. These variables include: insulin use (medication NDC codes), cardiovascular disease (ICD-10 diagnosis codes, medication NDC codes; ischemic heart disease, myocardial infarction), hypertension (ICD-10 diagnosis codes, medication NDC codes).

In addition, for the baseline time point t=0, $L_1(0)$ includes the following measured baseline covariates: age, sex, education, income, neighborhood deprivation index.

For t=1,...,K, L(t) also includes the time-varying outcome Y(t), defined below. i.e.,
$L(t) = (L_1(t), Y(t))$

Y(t): Outcomes of interest (included in L(t))

The outcome Y will take the value of 0 prior to onset, and jump to 1 in the time bin immediately following that which they develop the outcome, and and remain at 1 deterministically for the remainder of follow up. We define Y(t) to be CKD diagnosis by time t, defined by <u>either</u> of the following criteria:

1) CKD diagnosis, measured through inpatient diagnosis code; or
2) glomerular filtration rate (GFR) lab results strictly less than 60 mL/min.

Identifiability

Sequential Randomization Assumption

$$Y_{\bar{a}=a,\bar{c}=0}(t) \perp\!\!\!\perp A(t)|\bar{L}(t), \bar{A}(t-1), t = 0, ..., K$$

Holds if for each intervention node A(t), the measured past is sufficient to block all back door paths from A(t) to Y(k) for t<k< K+1. In particular, for this application, we are particularly concerned with time invariant and time varying variables that affect the decision to modify SGLT2 and affect (or predict) CKD, as well as censoring events (death, administrative end of study).

Positivity assumption

Need some positive probability of continuing to follow the regime of interest at each time point, given covariates:

$$P(A(t) = a(t)|\bar{A}(t-1) = \bar{a}(t-1), \bar{L}(t)) > 0, \ t = 1, \ldots, K \quad a.e.$$

Statistical model and estimand

Here, our statistical model M is nonparametric (making no assumptions set of possible observed data distributions). For the final time point (*t=K+1)* and a given exposure/censoring level $\bar{a}$, *o*ur statistical estimand is the iterated conditional expectation implementation of the g-computation formula[3–5]:

$$E(Y_{\bar{a}}(K+1)) = E[E[\ldots E[E[Y_{\bar{a}}(K+1)|\bar{L}(K), \bar{A}(K) = \bar{a}(K)]|\bar{L}(K-1), \bar{A}(K-1) = \bar{a}(K-1)]\ldots|L_0]]$$

Where $\bar{a}$, used for notational convenience, represents our regimes of interest and contains both treatment $\bar{a}_1$ and censoring $\bar{c}$.

Taking the difference of $E(Y_{\bar{a}}(K+1))$ with and without sustained SGLT2 use (and intervening to prevent censoring) provides our primary statistical estimand, corresponding to the two-year, counterfactual risk difference under the identifiability assumptions. Then summing the difference between the counterfactuals of interest (see section “Causal effect of interest”, subsections b and c) for all time points provides our contrast of the estimated survival curves.

Estimation

We will utilize simulations (detailed below) to select the optimal estimation method, including nuisance parameter estimation and variance estimation. To estimate our parameter of interest, we will use longitudinal TMLE[6] with Superlearner[7] as implemented via the “ltmle” R package[8].

Simulation study implementation

Prior to analyzing the registry data and finalizing this statistical analysis plan, we conducted a simulation study to assess estimator performance, benchmark computation time, and finalize the coding of the TMLE implementation for both point estimation and inference.

We created longitudinal exposure, censoring, outcome and covariate data guided by the Observed Data description above. To generate realistic data, we sampled from the empirical data distribution of the following baseline covariates: age, sex, education, income, neighborhood deprivation index. We additionally simulated longitudinal exposure, censoring, as well as the

time-varying covariates and outcome. Time-varying covariates included insulin use, cardiovascular disease, (ischemic heart disease, myocardial infarction), and hypertension. For each time point, the past history was used to predict the current time point, consistent with our structural causal model. We utilized highly adaptive LASSO (HAL) models for the prediction.[9] For this reason, HAL models were removed from the candidate algorithm set.

We simulated 70k patients per each dataset, and created 1,000 datasets. With each simulated dataset, we used longitudinal TMLE to estimate the statistical parameter (using the longitudinal g-formula), comparing sustained SGLT2 usage to sustained non-usage over time, using 60-day time discretization and a 2-year follow-up. We simulated the truth via large-sample (n=1000000) Monte Carlo Method simulation, setting the exposure to the desired counterfactual levels of interest.

Within each estimator, we considered a broad selection of algorithms to be chosen by a discrete SuperLearner algorithm for nuisance parameter estimation: GLM, penalized regression, generalized additive models, gradient boosting models, and neural networks.

With the simulation, we compared the following specifications:

Point estimation:

- We evaluated the performance of the longitudinal TMLE estimator
- We evaluated nuisance parameter estimation with discrete SuperLearner with and without pre-screening for covariates
    - For algorithms without built-in covariate selection (e.g., logistic regressions and generalized additive models), we included versions adjusting for all covariates as well as versions adjusting for covariates prescreened via likelihood ratio tests to avoid overadjustment with covariates not related to the outcome.
    - By default the *ltmle* package truncates the estimated propensity scores at the bounds: [0.01,1.0]. We compared TMLEs with and without this bounding.
- We compared strategies using the negative log-likelihood loss function in SuperLearner

Variance estimation:

- We chose between the empirical variance of the influence curve and nonparametric bootstrap. In all cases, we will use the resulting variance estimate to create Wald-Type 95% confidence intervals.

We compared possible strategies for estimation and inference with the empirical variance across the simulations, oracle coverage (i.e. coverage obtained with the empirical variance) , estimated variance, and 95% confidence coverage (i.e., coverage obtained with the estimated variance). First, we chose the strategy for point estimation that resulted in the lowest empirical variance, while maintaining nominal oracle coverage (i.e., close to 95%). Then, we selected the strategy for variance estimation that resulted in the most precise confidence intervals (i.e., smallest variance estimate), while maintaining nominal confidence interval coverage.

Simulation study results & Pre-specified estimation

Longitudinal TMLE with SuperLearner and covariate screeners had better oracle coverage and lower empirical variance than longitudinal TMLE without covariate screening. There was no meaningful difference in the performance of the TMLEs with and without truncation of the estimated propensity scores; so we chose untruncated results. For estimating the variance of the selected TMLE (with screeners and without propensity score truncation), the empirical influence curve (IC) and nonparametric bootstrap had similar 95% confidence interval coverage; so we chose IC for computational ease.

The final estimation selection is longitudinal TMLE with a discrete SuperLearner (with candidate algorithms GLM, penalized regression, generalized additive models, gradient boosting models, and neural networks) with algorithm pre-screening and no propensity score truncation, with empirical IC for variance estimation.

Interpretation

If identification assumptions are met, we will interpret the causal parameter both over time and as a cumulative measure (i.e. at time t=12):

- *After accounting for measured time-varying confounders, the counterfactual survival curves comparing sustained SGLT2 use vs no SGLT2 use over two years were/were not significantly different*
- *After accounting for measured time-varying confounders, two years of continuous use of SGLT2i was shown to decrease (or not) the risk of renal disease onset by X%*

Version history

- Version 1.0: We pre-specified the simulation study to objectively select the primary analytic approach (see section, “Simulation study implementation”.
- Version 1.1 (this version): We included the section, “Simulation study results & pre-specified estimation”, containing results from the simulation.

References

1. Petersen, M. L. & van der Laan, M. J. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiol. Camb. Mass* **25**, 418–426 (2014).
2. Nance, N. *et al.* Applying the causal roadmap to longitudinal national Danish registry data: a case study of second-line diabetes medication and dementia. Preprint at http://arxiv.org/abs/2310.03235 (2023).
3. Robins, J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **7**, 1393–1512 (1986).
4. Bang, H. & Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973 (2005).
5. van der Laan, M. J. & Gruber, S. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *Int. J. Biostat.* **8**, (2012).
6. Petersen, M. *et al.* Targeted Maximum Likelihood Estimation for Dynamic and Static Longitudinal Marginal Structural Working Models. *J. Causal Inference* **2**, 147–185 (2014).
7. Polley, E. SuperLearner: Prediction model ensembling method. (2023).
8. Lendle, S. D., Schwab, J., Petersen, M. L. & Laan, M. J. van der. ltmle: An R Package Implementing Targeted Minimum Loss-Based Estimation for Longitudinal Data. *J. Stat. Softw.* **81**, 1–21 (2017).
9. Benkeser, D. & van der Laan, M. The Highly Adaptive Lasso Estimator. *Proc. Int. Conf. Data Sci. Adv. Anal. IEEE Int. Conf. Data Sci. Adv. Anal.* **2016**, 689–696 (2016).

Supplementary Material B

Example Statistical Analysis Plan (SAP) inspired by the SEARCH-Youth Study

Laura B. Balzer
With the SEARCH-Youth Study Team

SEARCH-Youth was a cluster (clinic) randomized trial evaluating the effect of a multicomponent intervention on HIV viral suppression among adolescents and young adults (15-24 years) with HIV in rural Uganda and Kenya (Clinicaltrials.gov: NCT03848728). The results of the trial can be found at Ruel et al.[1] with corresponding Statistical Analysis Plan (SAP) in Balzer et al.[2] In the following, **we present a simplified SAP**, anchored on the Causal Roadmap and inspired by the primary endpoint in the SEARCH-Youth study. We again refer the reader to [1,2] for more information about the study and the actual SAP, including additional details for the primary analysis, secondary endpoints compared between arms, implementation outcomes, descriptive statistics, and power calculations. We gratefully acknowledge the SEARCH-Youth study team.

Roadmap Step 1: Research question

The SEARCH-Youth trial is designed to evaluate the effect of a multi-level, combination intervention on viral suppression (HIV RNA<400 c/mL) at two-years among adolescents and young adults (15-24 years) with HIV and in care at government-sponsored health clinics in rural western Kenya and southwestern Uganda. Since the intervention involves changes in clinical practice, the unit of randomization is clinics: 14 intervention and 14 control, balanced on country. Participants randomized to intervention clinics are offered life-stage assessments, flexible care access, and rapid feedback of viral suppression status. Providers at intervention clinics have a secure WhatsApp platform for inter-provider consultation. Participants and providers at control clinics follow the standard-of-care. Our primary research question is "what is the effect of the SEARCH-Youth intervention on viral suppression (HIV RNA<400 c/mL) at two-years of follow-up?"

Roadmap Step 2: Causal model

At enrollment, we measured the following baseline characteristics: age, gender, education attained, marital/partnership status, alcohol use, mobility, antiretroviral therapy (ART) combination, viral suppression status, and HIV care status (recently engaged, previously engaged, or re-engaging in care). Let W denote these characteristics for a given participant. Additionally, let E^c denote clinic-level, baseline characteristics, such as country, the number of persons 15-24 years in HIV care, and the proportion with viral suppression among those engaged at baseline. Let A^c be a clinic-level indicator of being randomized to the intervention. Throughout superscript c denotes cluster-level variables.

During endpoint data collection, we will ascertain whether the participant has outmigrated M, defined as moving ≥3 hours traveling distance from the study clinic. As detailed in the Study Protocol, we will attempt viral load measurement — via clinic-based ascertainment or community-based tracking — regardless of outmigration status. Let Δ be an indicator of HIV RNA (i.e., viral load) measurement at endline for a given participant. Finally, Y be an indicator of viral suppression at two-years. Specifically, Y is

- 1 if the participant's endline viral load is <400 c/mL
- 0 if their endline viral load is ≥400 c/mL
- otherwise missing.

Using the following hierarchical, non-parametric structural equation model,[3,4] we represent the relationships between these variables. Throughout, bold denotes matrices or vectors of the individual-level variables across participants in a given cluster.

$$E^c = f_{E^c}(U_{E^c})$$
$$\boldsymbol{W} = f_{\boldsymbol{W}}(E^c, \boldsymbol{U_W})$$
$$A^c = f_{A^c}(U_{A^c})$$
$$\boldsymbol{M} = f_{\boldsymbol{M}}(E^c, \boldsymbol{W}, A^c, \boldsymbol{U_M})$$
$$\boldsymbol{\Delta} = f_{\boldsymbol{\Delta}}(E^c, \boldsymbol{W}, A^c, \boldsymbol{U_\Delta})$$
$$\boldsymbol{Y} = f_{\boldsymbol{Y}}(E^c, \boldsymbol{W}, A^c, \boldsymbol{M}, \boldsymbol{\Delta}, \boldsymbol{U_Y})$$

By design the unmeasured factors contributing the intervention assignment U_{A^c} are independent of the others. This causal model represents the data generating process for a clinic. Importantly, this causal model allows for arbitrary dependence within a cluster.

Roadmap Step 3: Causal effect

Let $Y_i{}^c(a^c)$ denote the counterfactual proportion of participants with viral suppression at two-years if possibly contrary-to-fact clinic i were randomized to arm $A^c = a^c$. To increase interpretability and precision, the primary effect measure will be for the $N = 28$ clinics in the trial (i.e., the sample effect).[5,6] Specifically, we focus on the sample prevalence ratio

$$\Phi(1) \div \Phi(0) \; = \frac{1}{N}\sum_{i=1}^{N} Y_i{}^c(1) \div \frac{1}{N}\sum_{i=1}^{N} Y_i{}^c(0)$$

In secondary analyses, we will evaluate the sample average treatment effect (SATE): $\Phi(1) - \Phi(0)$.[5,6] In additional secondary analyses, we will include weights in these empirical means to weight individuals (instead of clinics) equally, as detailed in Benitez et al.[7]

Roadmap Step 4: Observed data and statistical model

The observed data for each clinic are denoted $\boldsymbol{O} = (E^c, \boldsymbol{W}, A^c, \boldsymbol{M}, \boldsymbol{\Delta}, Y)$ and are assumed to arise from a data generating process compatible with the prior causal model (Roadmap Step 2). The resulting statistical model for the set of possible observed data distributions is semi-parametric. We know that with probability 0.5 a given clinic is either intervention or control: $\mathbb{P}(A^c = 1) = 0.5$.

Roadmap Steps 5 & 6: Identifiability, estimation, and inference

The primary approach will be Two-Stage TMLE, which is appropriate for cluster randomized trials with missing outcomes.[8,9] In Stage-I, we first identify and estimate a cluster-level endpoint that accounts for missingness/censoring at the individual-level. Then in Stage-II, we use the endpoint estimates to evaluate the cluster-level effect with maximum precision, while accounting for limited numbers of clusters (i.e., small sample size). As detailed in Balzer et al.,[8] the Two-Stage approach lets the measurement mechanism vary by cluster and circumvents complicated estimands from the missing data equivalent of time-dependent confounding. Furthermore, the Two-Stage approach enables us to focus our Stage-II efforts on empirical efficiency maximization.

Stage-I: Identification & Estimation

In the primary analysis, we will include all participants and adjust for potentially differential measurement of viral suppression status.[i] To identify the cluster-level endpoint under a hypothetical intervention to ensure complete outcome measurement, we need the following missing-at-random assumption to hold within each cluster. Specifically, within values of

[i] Note in the actual SEARCH-Youth study, the primary endpoint considered missing endpoints to be failures (i.e., as unsuppressed); this is discussed in sensitivity analysis #2, described below.

adjustment variables, we will assume that participants with measured viral loads are representative of participants with missing viral loads *within* that cluster. Our adjustment set will consist of age, sex, baseline suppression status, and the time-varying indicator of outmigration. Therefore, our statistical estimand is

$$Y^c = \mathbb{E}[\mathbb{E}(Y|\Delta = 1, W, M)]$$

The cluster-level covariates E^c and exposure A^c are constant in each cluster and, thus, omitted. For the cluster-level endpoint Y^c to be well-defined, we need an additional assumption on data support:

$$\mathbb{P}(\Delta = 1|W, M) > 0, \text{ a.e.}$$

In words, we need a positive probability of viral load measurement within all possible values of adjustment variables (in each cluster).

For estimation of the cluster-level endpoint Y^c, we will implement targeted minimum loss-based estimation (TMLE).[8,10] To reduce the risk of bias due to parametric modeling assumptions and to improve efficiency, we will use Super Learner, an ensemble machine learning method, to combine predictions from generalized linear models, generalized additive models, and the mean through 10-fold cross-validation and the non-negative least squares loss.[11] We will implement TMLE with Super Learner in each cluster separately and, thus, N=28 times to obtain cluster-level endpoints $\hat{Y}_i^c$ for each clinic $i = 1, \ldots, N$.

We will conduct two sensitivity analyses to examine the robustness of our findings:

1. Excluding participants without an endline viral load measure
2. Including all participants, but considering missing viral loads as non-suppressed

The first sensitivity analysis is sometimes called a "complete-case analysis" and relies on the missing-completely-at-random (MCAR) assumption. Under the MCAR assumption, we can estimate the clinic-level endpoint Y^c with the raw proportion among those measured. The second sensitivity analysis considers a composite outcome and is not subject to missingness. Therefore, we can again estimate the clinic-level endpoint Y^c with the proportion with viral suppression (among all participants).

Stage-II: Identification and Estimation

Given the estimated cluster-level endpoints $\hat{Y}_i^c$ for $i = 1, \ldots, N$ from Stage-I, we will then evaluate the intervention effect with a cluster-level TMLE. At the cluster-level, there is no confounding or missingness. Therefore, we can easily identify and estimate the causal effect of interest (Roadmap Step 3) by contrasting the average of the cluster-level endpoints between arms. This unadjusted prevalence ratio will be used in a sensitivity analysis.

In the primary analysis, we will adjust for baseline covariates to improve efficiency, as repeatedly demonstrated.[12–20] In other words, our statistical estimand for Stage-II is given by

$$\Psi(1) \div \Psi(0) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}(Y^c|A^c = 1, {E^c}_i, {W^c}_i) \div \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}(Y^c|A^c = 0, {E^c}_i, {W^c}_i)$$

where ${W^c}_i$ denote aggregates of individual-level baseline covariates. We will use TMLE with Adaptive Pre-specification (APS) for estimation and inference. APS is a formal procedure to data-adaptively choose from a pre-specified set the adjustment variables that maximize

empirical efficiency, while protecting Type-I error control.[7–9,21,22] Our candidate adjustment variables will be the clinic-specific number of adults 15-24 years in HIV care at baseline, the clinic-specific proportion with viral suppression among those engaged at baseline, or nothing (unadjusted). Our candidate estimation algorithms for the outcome regression $\mathbb{E}(Y^c|A^c, E^c, W^c)$ will be limited to working GLMs[16] adjusting for at most of covariate. Our candidate estimation algorithms for the known propensity score $\mathbb{P}(A^c = 1|E^c, W^c)$ will also be limited to working GLMs adjusting for at most of covariate. Using leave-one-out cross-validation, we will then select the combination of estimators and, thus, the TMLE that minimizes the cross-validated variance estimate.

Statistical Inference

Under the conditions detailed in Balzer et al.,[8] Two-Stage TMLE will be an asymptotically linear estimator, meaning that it will be normally distributed in the large data limit. Informally, these conditions require that the cluster-level endpoints are estimated well in Stage-I. For variance estimation, we use the sample variance of the estimated, cluster-level, influence curve. As a finite sample approximation to the standard normal distribution, we will use the Student's t-distribution with 28-2=26 degrees of freedom.[12]

Using a two-sided test at the 5% significance level,[ii] we will formally test the *null hypothesis* that the SEARCH-Youth intervention did not impact viral suppression, as compared to the standard-of-care. We will report point estimates and 95% confidence intervals for the arm-specific endpoints as well as the relative and absolute effects.

Roadmap Step 7: Interpretation

We will interpret the effect analyses in light of our missing data assumptions and the randomization of the intervention. In the actual SEARCH-Youth Study, Ruel et al. concluded, "The mean proportion of participants with virological suppression at 2 years was 88% (95% CI 85–92) for participants in intervention clinics and 80% (77–84) for participants in control clinics, equivalent to a 10% beneficial effect of the intervention (risk ratio [RR] 1.10, 95% CI 1.03–1.16; p=0.0019)."[1]

[ii] Note in the actual SEARCH-Youth study, the pre-specified hypothesis test was one-sided, corresponding to the null hypothesis of no improvement in viral suppression due to the intervention.

How simulations informed this SAP

Prior to unblinding the data, we pre-specified and conducted the following treatment-blind simulations to select the primary analytic approach that strictly controlled Type-I error. As a secondary criterion, we evaluated 95% confidence interval coverage for the cluster-level endpoints Y^c.

Treatment-blind simulations with dependent and missing outcomes

Recall the individual-level outcome Y is subject to missingness but the true measurement mechanism is unknown. To evaluate estimator performance with simulations, we re-defined the outcome Y to be 1 if the participant's endline viral load was <400 c/mL and 0 otherwise (including endline viral load ≥400 c/mL or missing). Then we generated 1000 plasmode datasets as follows. We preserved the real covariate and (redefined) outcome data, but randomly permuted the treatment indicator A^c across clusters. Then an independent statistician simulated individual-level, measurement indicators Δ as a function of the baseline cluster-level and individual-level covariates, the permuted exposure, and time-varying covariates. The primary statistician (LBB) was blinded to the functional form of the measurement mechanism. In the simulation, the (redefined) outcomes were set to missing if the simulated measurement indicator Δ was 0. Using these simulated datasets, we compared various implementations of Two-Stage TMLE, as follows.

First, we focused on Stage-I estimation of the cluster-level endpoint $Y^c = \mathbb{E}[\mathbb{E}(Y|\Delta = 1, W, M)$ and compared

- The empirical mean (i.e., the unadjusted proportion with the outcome among those measured)
- TMLE with main terms regression for nuisance parameter estimation
- TMLE with Super Learner (combining predictions from generalized linear models, generalized additive models, and the mean with 10-fold cross-validation and non-negative least squares loss) for nuisance parameter estimation

For all, we fixed our approaches to Stage-2 estimation to the unadjusted effect estimator and for variance estimation to the standard, cluster-level influence curve (IC). This allowed us to evaluate the potential for alternative Stage-I approaches (both identifiability assumptions and estimators) to reduce bias due to differential outcome measurement.

To evaluate these approaches, we calculated the proportion of simulated datasets where the 95% confidence interval excluded the null (equivalent to the Type I error rate). As expected, only flexible adjustment for missing outcomes using TMLE with Super Learner in Stage-I preserved the Type-I error rate at $\leq 5\%$ (Table 1). Additionally, in these simulations the true values of the cluster-level endpoints Y^c in the absence of missingness were known; therefore, we also verified this approach resulted in nominal-conservative 95% confidence interval coverage for the cluster-level endpoints.

Second, we focused on Stage-II estimation of the intervention effect $\Psi(1) \div \Psi(0)$ and compared

- TMLE with “limited APS” where candidate estimators were limited to working GLMs adjusting for at most a single covariate[21]
- TMLE with “expanded APS” where candidates including worked GLMs adjusting for a single covariate, working GLMs adjusting for multiple covariates, penalized regression, and the unadjusted estimator[22]
- Standard versus cross-validated estimates of the cluster-level IC for variance estimation[21]

For all, we fixed our approach to Stage-I estimation of the cluster-level endpoints as TMLE with Super Learner. This allowed us to evaluate the potential efficiency gains through data-adaptive adjustment for baseline covariates in Stage-II, after appropriately addressing missing data in Stage-I. As shown in Table 2, the limited APS implementation in Stage-II and the standard IC approach to variance estimation resulted in the best Type-I error control.

Thus, we settled on the following implementation of Two-Stage TMLE as the primary analytic approach: Stage-I estimation using TMLE with Super Learner, Stage-II estimation with TMLE using limited APS, and variance estimation with the standard IC approach. Two-Stage TMLE with the unadjusted effect estimator in Stage-II was a pre-specified secondary analysis.

Stage-I estimation	**Stage-II estimation**	**Variance estimation**	**Type-I error rate for $\Psi(1) \div \Psi(0)$**	**95% CI coverage for Y^c**
Simple mean	Unadjusted	Standard IC	11%	80%
TMLE with main terms regression	Unadjusted	Standard IC	8%	83%
TMLE with Super Learner	Unadjusted	Standard IC	3%	96%

Table 1: Hypothetical simulation results when comparing Stage-I estimators of the cluster-level endpoint Y^c. All approaches used the unadjusted effect estimator in Stage-II and the standard influence curve (IC)-based inference.

Stage-I estimation	**Stage-II estimation**	**Variance estimation**	**Type-I error rate for $\Psi(1) \div \Psi(0)$**
TMLE with Super Learner	Limited APS	Standard IC	5%
TMLE with Super Learner	Limited APS	Cross-validated IC	2%
TMLE with Super Learner	Expanded APS	Standard IC	10%
TMLE with Super Learner	Expanded APS	Cross-validated IC	7%

Table 2: Hypothetical simulation results when comparing Stage-II estimators of the intervention effect $\Psi(1) \div \Psi(0)$ and approaches to variance estimation. All approaches used TMLE with Super Learner in Stage-I.

History of changes

- Version 1.0: We pre-specified the simulation study to objectively select the primary analytic approach: data generating processes, candidate estimators, and estimation metrics.
- Version 2.0 (this version): We included the results of the simulation study and pre-specified the primary and secondary/sensitivity analyses.

Acknowledgements

We gratefully acknowledge the SEARCH-Youth study team, clinic staff, and study participants. SEARCH-Youth was supported by a grant from Eunice Kennedy Shriver National Institute of Child Health and Human Development at the National Institutes of Health (UH3HD096915). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Ruel T, Mwangwa F, Balzer LB, et al. A multilevel health system intervention for virological suppression in adolescents and young adults living with HIV in rural Kenya and Uganda (SEARCH-Youth): a cluster randomised trial. *The Lancet HIV*. 2023;10(8):e518-e527. doi:10.1016/S2352-3018(23)00118-2
2. Balzer LB, Ruel T, Havlir DV, the SEARCH-Youth Study Team. Statistical Analysis Plan for Primary and Selected Secondary Health Endpoints of the SEARCH-Youth Study. *arXiv*. 2022;https://arxiv.org/abs/2211.02771.
3. Pearl J. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge University Press; 2009.
4. Balzer LB, Zheng W, van der Laan MJ, Petersen ML, the SEARCH Collaboration. A new approach to hierarchical data analysis: Targeted maximum likelihood estimation for the causal effect of a cluster-level exposure. *Stat Meth Med Res*. 2019;28(6):1761-1780.
5. Neyman J. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (In Polish). English translation by D.M. Dabrowska and T.P. Speed (1990). *Statistical Science*. 1923;5:465-480.
6. Balzer LB, Petersen ML, van der Laan MJ. Targeted estimation and inference of the sample average treatment effect in trials with and without pair-matching. *Statistics in Medicine*. 2016;35(21):3717-3732. doi:10.1002/sim.6965
7. Benitez A, Petersen ML, van der Laan MJ, et al. Defining and estimating effects in cluster randomized trials: A methods comparison. *Stat Med*. 2023;42(19):3443-3466. doi:10.1002/sim.9813
8. Balzer LB, van der Laan M, Ayieko J, et al. Two-Stage TMLE to reduce bias and improve efficiency in cluster randomized trials. *Biostatistics*. 2023;24(2):502-517. doi:10.1093/biostatistics/kxab043
9. Nugent JR, Marquez C, Charlebois ED, Abbott R, Balzer LB. Blurring cluster randomized trials and observational studies using Two-Stage TMLE to address sub-sampling, missingness, and few independent units. *Biostatistics*. 2023;In Press.
10. van der Laan M, Rose S. *Targeted Learning: Causal Inference for Observational and*

Experimental Data. Springer; 2011.
11. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Statistical Applications in Genetics and Molecular Biology*. 2007;6(1):Article 25. doi:10.2202/1544-6115.1309
12. Hayes RJ, Moulton LH. *Cluster Randomised Trials*. Chapman & Hall/CRC; 2009.
13. Fisher RA. *Statistical Methods for Research Workers*. 4th ed. Oliver and Boyd Ltd.; 1932.
14. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. *Stat Med*. 1996;15:1069-1092.
15. Moore KL, van der Laan MJ. Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation. *Statistics in Medicine*. 2009;28(1):39-64. doi:10.1002/sim.3445
16. Rosenblum M, Laan MJ van der. Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables. *The International Journal of Biostatistics*. 2010;6(1):Article 13. doi:10.2202/1557-4679.1138
17. Colantuoni E, Rosenblum M. Leveraging Prognostic Baseline Variables to Gain Precision in Randomized Trials. *Stat Med*. 2015;34(18):2602-2617.
18. Turner EL, Prague M, Gallis JA, Li F, Murray DM. Review of Recent Methodological Developments in Group-Randomized Trials: Part 2-Analysis. *Am J Public Health*. 2017;107(7):1078-1086. doi:10.2105/AJPH.2017.303707
19. Murray DM, Taljaard M, Turner EL, George SM. Essential Ingredients and Innovations in the Design and Analysis of Group-Randomized Trials. *Annu Rev Public Health*. 2020;41:1-19.
20. Benkeser D, Díaz I, Luedtke A, Segal J, Scharfstein D, Rosenblum M. Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics*. 2021;77(4):1467-1481. doi:10.1111/biom.13377
21. Balzer L, van der Laan MJ, Petersen M, SEARCH Collaboration. Adaptive Pre-specification in Randomized Trials With and Without Pair-Matching. *Statistics in Medicine*. 2016;35(10):4528-4545.
22. Balzer LB, Cai E, Garraza LG, Amaranath P. Adaptive Selection of the Optimal Strategy to Improve Precision and Power in Randomized Trials. *Biometrics*. 2023;In Press.

Chapter 2

Applying the Causal Roadmap to longitudinal national Danish registry data: a case study of second-line diabetes medication and dementia[1]

[1]At the time of dissertation submission, this article was being resubmitted with revisions per request of the Journal for Causal Inference. Coauthors of the article include: Andrew Mertens, Thomas Gerds, Zeyi Wang, Christian Torp-Pedersen, Mark van der Laan, Kajsa Kvist, Theis Lange, Bochra Zareini, Maya Petersen.

2.1 Introduction

There has been a significant increase in the amount of existing observational or "real world" health data both collected by and aggregated for researchers over the past decades. This type of data has become increasingly important in medical decision-making, as it allows researchers to answer questions that would not be feasible to address in a randomized trial due to financial constraints, ethical considerations, or logistical challenges. Electronic health record databases, with their large sample sizes and long follow-up times, enable the study of rare outcomes, and realistic treatment usage. However, because observational data are particularly susceptible to confounding, real-world evidence is often used only for associational analyses and causal claims are discouraged. [98, 99, 100] Further, while measured confounders can be adjusted for, standard statistical estimation approaches that rely on pre-specified parametric models may fail to fully adjust for measured confounders (introducing bias that does not decrease with greater sample size), while *ad hoc* model adaptation based on data exploration undermines the basis of robust statistical inference.[6]

The Causal Roadmap provides a structured framework for navigating these challenges and generating robust evidence for causal questions, when appropriate assumptions are met.[24] Specifically, the Causal Roadmap provides guidance for the following steps: 1) translation of a causal question, such as the comparative effectiveness of two longitudinal treatment protocols, into a formally defined *causal estimand*; 2) explicit statement of the *observed data* and knowledge about the processes that gave rise to it (*causal model*); 3) *identification*, or the formal translation of the causal estimand (which describes an ideal hypothetical experiment) into a *statistical estimand* (a function of the observed data distribution that can be estimated) under explicit causal assumptions; 4) pre-specification of a *statistical model* that avoids any unsubstantiated statistical assumptions, and a *statistical estimator* selected to provide the best expected performance (e.g., nominal 95% CI coverage, minimum variance); and 5) support for *interpretation* in the context of these assumptions, including sensitivity analyses.

Large, comprehensive, and representative longitudinal datasets like national registries offer advantages for evaluating causal effects by allowing for generalizable study populations and investigation of the real-world effectiveness of long-term treatments on rare outcomes. Their use also introduces challenges, making implementation of the Casual Roadmap simultaneously more complex and arguably even more essential to ensure robust inferences.

The first challenge arises because the full longitudinal counterfactual interventions of interest must be specified, including not only initial treatment assignments, but also any post-baseline restrictions on treatment modification or drop-in, as well

as any hypothetical interventions on treatment compliance, censoring or measurement.[101] Specification of a "target trial"[25] is one commonly used device to support specification of a causal estimand. Of note, however, specification of post-baseline "interventions" or protocols, including those to enforce compliance or to prevent censoring, need not correspond to any realistic intervention that might be accomplished in an actual trial.

Secondly, both the observed data and confounding structures in longitudinal registry data are complex, and thus the potential identification of the target causal estimand in such studies requires a more complex statistical estimand, such as the longitudinal g-formula.[7] This, in turn, necessitates more complex statistical estimation procedures capable of adjusting for time-varying confounders. Finally, many of the advantages of using registry data—including the ability to study the effects of prolonged treatments on rare outcomes and to capture of a wide range of potential and time-dependent confounders—also introduce additional challenges to estimation. Flexible machine learning approaches to covariate adjustment are paramount to making optimal use of complex longitudinal data captured in a registry, but must be implemented such that statistical inference is preserved.[7] Further, even with initially large sample sizes and common treatments, the number of patients who comply with long-term protocols can dwindle quickly, resulting in practical positivity violations; examples of this include longitidinal inverse probability of treatment weighting (IPTW), parametric g computation, and longitudinal maximum likelihood estimation.[65] Finally, the unique ability of registries to evaluate effects on rare outcomes can introduce new challenges to estimator performance. Careful estimator specification and benchmarking can address many of these challenges; however, in order to provide a firm basis for statistical inference, the full estimation procedure must be pre-specified.

Motivated by both the utility of the Causal Roadmap for longitudinal registry analyses and the practical challenges that arise in its application, in this paper we present a detailed case study. Through it, we provide practical guidance on implementation of the longitudinal Causal Roadmap, with a particular emphasis on the use of simulations to guide pre-specification of a statistical estimator that integrates machine learning. Specifically, we utilize the Roadmap to evaluate the long-term cumulative causal effect of second-line diabetes medications on dementia risk among a cohort of diabetes patients using the Danish National Registry. The Danish National Registry is one of the longest-standing national registry databases, and has expansive longitudinal data dating back several decades.[102] Due to its large size, rich comorbidity and concurrent medication measurements, and extended follow-up time, the Danish registry is well-suited for examining rare and longer-term exposures and outcomes. It also has a high degree of diagnostic reliability for some (but not all)

diseases, with a positive predictive value of 0.96 and 0.98 for diabetes and dementia, respectively.[103] The Danish registry has high-quality data for studying longer-term diabetes care; the study, prognosis, and adequate treatment of diabetes is a national priority reflected in practice and national research interests. Dementia affects around three percent of the older Danish population, but several studies suggest that the true prevalence is higher.[104] Dementia is particularly burdensome to the diabetic population.[105] Continually emerging evidence suggests newer second-line diabetes medications may hold promise in preventing neurodegeneration(eg. [21, 14, 22]); however, some newer medications are less historically common relative to more established therapies.[106] Due to the relatively uncommon nature of exposure and outcome of interest and the need for prolonged follow up, examining this relationship without the use of a large data source like the Danish registry would be infeasible.

With this motivation, we evaluate the difference in the counterfactual five-year cumulative risk of dementia if a target population of adults with type 2 diabetes had initiated and remained on GLP-1 receptor agonists (GLP-1RAs), a relatively new class of drugs, compared to a range of active comparator protocols. We utilize a non-parametric statistical model that avoids unsubstantiated assumptions, and is thus known to contain the true observed data distribution; this contrasts with commonly employed approaches to statistical estimation in which a more restrictive statistical model (such as a main term logistic model on the propensity score) is adopted from convenience or habit. The use of a large semi- or non-parametric statistical model, as dictated by the Roadmap, provides crucial protection against bias due to both model misspecification and post-hoc data exploration. It does, however, set a high bar for choice of a statistical estimator. Specifically, the pre-specified statistical estimator must be able to provide precise and robust estimates of the treatment effect without reliance on simplified lower-dimensional parametric models.

Time-dependent confounding is accounted for through use of the iterated conditional expectation expression of the longitudinal g-formula as the statistical estimand.[40, 107] Longitudinal targeted maximum likelihood estimation (L-TMLE)[101, 7] provides a general estimation approach for these causally derived estimands, while integrating "double" machine learning (in both a series of outcome regressions and in estimates of propensity score), in order to flexibly incorporate covariate adjustment.[6] However, full pre-specification of a TMLE for this estimand requires several key decisions. These decisions can have profound impacts on finite sample estimator performance, particularly in the challenging setting of large scale registry analysis of long-term treatment effects on rare outcomes. Further, given the challenges of rare outcomes in our case study, an investigator may also wish to verify whether the use of a TMLE is warranted, or whether a simpler estimator such as in inverse probability of treatment (IPTW) estimator, despite having theoretical

shortcomings, might provide to be an equivalent or superior finite sample estimate.

Our case study provides practical guidance on the implementation of the Roadmap in large-scale registry data. In particular, we highlight how rare outcomes can raise challenges for flexible and robust estimators such as L-TMLE, even in the context of the large sample sizes provided by the registry, and discuss estimation approaches to addressing these challenges, including careful specification of the machine learning library and selection of an appropriate variance estimator. We show how simulations can be used to help pre-specify the estimation algorithm (a prerequisite for valid statistical inference) in order to optimize the expected performance of the statistical estimator (in terms of confidence interval coverage and precision). As we discuss, simulations can also be useful when working with registry data that can be difficult to access for logistical reasons, prompting a further need for simulated data. There are examples of previous analyses of Nordic registry data using a target trial framework (eg. [108, 109, 110]). However, to the best of our knowledge, this is among the first studies to apply the Causal Roadmap to a national registry of historic data on over 200,000 diabetes patients.[111]

The paper is organized as follows. In Section 2.2, we step through the application of the Causal Roadmap to our case study, evaluating the long-term effect of GLP-1RA use on incident dementia risk using the Danish National Registry, up to the step of statistical estimation. We illustrate opportunities, challenges, and practical responses at each step of the Roadmap. In Section 2.3, we specify the statistical model and detail the estimation procedure, including the simulation study, which was used to pre-specify the analysis. In particular, we provide a detailed overview of key estimation choices in the implementation of our longitudinal targeted maximum likelihood estimator, and the use of simulations to guide these choices. In Section 2.4, we present the results and interpretation of our analysis, including sensitivity analyses. In Section 2.5, we discuss the findings and the implications of this work.

2.2 Case study of the longitudinal Causal Roadmap: evaluating the long-term effect of GLP-1RA use on incident dementia risk using the Danish National Registry

Background and motivation

Type two diabetes mellitus (T2DM) is a cardio-metabolic disease that affected 462 million people globally in 2017 and is increasing in prevalence each year.[112] It is

characterized by elevated blood sugar levels, which over time can strain not only the cardiovascular but also the nervous system.[113] Dementia is a progressive neurodegenerative disease characterized by impaired cognitive functioning that interferes with daily life. T2DM has been shown to accelerate both brain aging and neurodegeneration, which are antecedents to dementia.[105, 114] While the exact mechanisms are an active area of research, hypothesized mechanisms include a complex mixture or pathways—including defective insulin signaling, metabolic/mitochondrial dysfunction, oxidative stress, and vascular damage, among others.[115, 12]

The connection between GLP-1RAs and improved cognition has been well-demonstrated in mouse models. Specifically, liraglutide (a GLP-1RA) reduced neuroinflammation and improved synaptic function (reducing amyloid formation).[116, 117] GLP-1RAs are thought to be particularly protective, both because of their effects on glucose regulation (and subsequently weight reduction) as well as a direct effect on neuroplasticity.[12, 118, 13] Secondary analyses of randomized placebo-controlled trial data in diabetic patients also showed a significant reduction in cognitive decline among those in the GLP-1RA treatment arm compared to the control arm.[21, 14, 119] Due to the short-term follow-up of these randomized studies, the outcome of interest was either cognitive worsening (from validated cognition tests) or dementia-related adverse events. Studying dementia onset requires longer-term follow-up in large cohorts that are seldom practically or financially feasible for most randomized trials. Further, trial participants in these secondary analyses were also higher risk patients and not representative of the general population.

Observational studies circumvent these shortcomings by allowing access to expansive datasets with a sufficient number of events to look at relatively rare exposures or outcomes and evaluate long-term treatment effects. A growing body of observational research studying the effects of GLP-1RAs on dementia has shown a reduction in dementia risk.[21, 22, 120] In the Danish National Registry specifically, one analysis showed a protective effect of GLP-1RA use on dementia for every year increase in exposure.[22] While these results are encouraging, the application of the Causal Roadmap illustrates how they may fall short of estimating the causal effect of greatest interest. (1)*Causal estimand specification*: some used a hazard ratio to measure effect[21, 22]; while commonly used, the hazard ratio is not causally interpretable. A hazard is the instantaneous rate of having an event given that the event has not already occurred, and thus at each time point the hazard ratio compares event rates between study arms among participants who remain event-free. By definition, over time, this measure is susceptible to an inherent selection bias which can make it impossible to recover a counterfactual quantity generally of interest in a causal study.[121] (2)*Identification*: in order to causally interpret results, we must satisfy the sequential randomization assumption (see Section 2.3.1), which states that the

counterfactual outcome must be independent of treatment assignment, conditional on measured past. Adjusting for only baseline covariates may not be sufficient to adjust for confounding, particularly if there are time-dependent confounders that are affected by past treatment usage and affect future treatment usage.(3)*Estimation*: estimation, particularly in the context of data sparsity and time-varying covariate histories, can prove challenging. The Cox model is commonly used in longitudinal registry analyses. Estimators that can incorporate variable selection can optimize precision, and become particularly important in such cases. Integration of machine-learning for flexible adjustment can both reduce bias and improve precision, but must be incorporated into appropriate estimators.

Through the subsequent analysis, we apply the longitudinal Causal Roadmap to answer the question: what is the effect of sustained, cumulative exposure to GLP-1RAs vs i) any active comparator or, ii) specific active comparators (sodium-glucose cotransporter-2 or SGLT2 inhibitors, dipeptidyl peptidase 4 or DPP-4 inhibitors) on the cumulative risk of dementia by five years among patients with T2DM initiating second-line antihyperglycemic therapy for the first time? For each of these causal questions, we outline the causal estimand, specify the observed data and a causal model describing what is known about the real-world processes that generated it, and finally identify the causal estimand as a function of the observed data distribution that is equivalent to the causal estimand under explicit causal assumptions. This sets the stage for defining the statistical estimation problem and specifying an estimator, as we do in Section 3.

Target causal estimand

Our target population consists of insulin-naive Danish diabetes patients who initiate second-line diabetes medication for the first time. We follow these patients from index date $t = 1$, corresponding to the date of first use of secondline diabetes medication, until a maximum of five year follow-up period, $t = 1, ..., K + 1$, where each unit of time is six months in length. The time $K + 1$ is defined as the time at which the outcome (dementia) status is evaluated, in main analyses $K + 1$ is set to five years, or ten 6-month long time intervals. In this population, we are interested in two general causal questions, corresponding to two types of hypothetical interventions (also known as treatment strategies or treatment regimes). We are firstly interested in a hypothetical intervention only on GLP-1RA use throughout follow-up ($\bar{\tilde{A}}_1 \equiv \tilde{A}_1(1), ..., \tilde{A}_1(K + 1)$), where $\tilde{A}_1(t)$ is an indicator that denotes use of GLP-1RA during time t, and where, here and throughout, an overbar denotes the the longitudinal history of a variable). Our hypothetical regimes of interest are sustained GLP-1RA use throughout follow-up (denoted by $\bar{\tilde{a}}_1 = 1$), and sustained

GLP-1RA non-use throughout follow-up (denoted by $\bar{\tilde{a}}_1 = 0$); in both regimes we also hypothetically intervene to prevent censoring ($\bar{c} = 0$, where $C(t)$ denotes an indicator of right censoring by time t). Let Y(t) denote an indicator that a participant is diagnosed with dementia by time t; for notational convenience, let $Y \equiv Y(K+1)$ denote diagnosis of dementia by the final time point of interest, $K+1$. Our counterfactual outcomes of interest under a hypothetical intervention on GLP-1RA use ($\bar{\tilde{A}}_1$) correspond to the diagnosis of dementia by time $K+1$ that would have been seen under these two hypothetical regimes (Table 2.1; these counterfactual quantities are formally defined using the causal model described Supplemental Materials 1): $Y_{\bar{\tilde{a}}_1=1,\bar{c}=0}, Y_{\bar{\tilde{a}}_1=0,\bar{c}=0}$. The above hypothetical regimes allow us to define our target causal estimand (or "causal contrasts of interest"[25]). Here, we are interested in the cumulative causal risk difference for dementia diagnosis by five years if all patients had complied with the intervention arm vs. the control arm, intervening to prevent administrative censoring (Table 2.1): $E[Y_{\bar{\tilde{a}}_1=1,\bar{c}=0}] - E[Y_{\bar{\tilde{a}}_1=0,\bar{c}=0}]$.

	Causal estimand 1: GLP1 vs. no GLP1	Causal estimand 2: GLP1 vs. SGLT2	Causal estimand 3: GLP1 vs. DPP4
Eligibility criteria	Dementia and insulin-naive diabetes patients initiating second-line therapies between 2009-2021	Dementia and insulin-naive diabetes patients initiating second-line therapies between 2014-2021	Dementia and insulin-naive diabetes patients initiating second-line therapies between 2009-2021
Treatment regime	Sustained GLP1 use ($\bar{\tilde{a}}_1 = 1$) throughout follow up, intervening to prevent censoring	Sustained GLP1 use ($\bar{\tilde{a}}_1 = 1$) and no SGLT2 use ($\bar{\tilde{a}}_2 = 0$) throughout followup, intervening to prevent censoring	Sustained GLP1 use ($\bar{\tilde{a}}_1 = 1$) and no DPP4 use ($\bar{\tilde{a}}_2 = 0$) throughout followup, intervening to prevent censoring
Counterfactual outcome under treatment	$Y_{\bar{\tilde{a}}_1=1,\bar{c}=0}(t)$	$Y_{\bar{\tilde{a}}_1=1,\bar{c}=0}$	
Control regime	No GLP1 use ($\bar{\tilde{a}}_1 = 0$) throughout follow-up, intervening to prevent censoring	Sustained SGLT2 use ($\bar{\tilde{a}}_2 = 1$) and no GLP1 use ($\bar{\tilde{a}}_1 = 0$) throughout follow up, intervening to prevent censoring	Sustained DPP4 use ($\bar{\tilde{a}}_2 = 1$) and no GLP1 use ($\bar{\tilde{a}}_1 = 0$) throughout follow up, intervening to prevent censoring
Counterfactual outcome under control	$Y_{\bar{\tilde{a}}_1=0,\bar{c}=0}(t)$	$Y_{\bar{\tilde{a}}_1=0,\bar{\tilde{a}}_2=1,\bar{c}=0}(t)$	
Target causal parameter	$E[Y_{\bar{\tilde{a}}_1=1,\bar{c}=0}] - E[Y_{\bar{\tilde{a}}_1=0,\bar{c}=0}]$	$E[Y_{\bar{\tilde{a}}_1=1,\bar{\tilde{a}}_2=0,\bar{c}=0}] - E[Y_{\bar{\tilde{a}}_1=0,\bar{\tilde{a}}_2=1,\bar{c}=0}]$	

Table 2.1: Causal estimands for three key target comparisons of second line medication use

We are secondly interested in a hypothetical study intervening on both exposure (GLP-1RA use, $\bar{\tilde{A}}_1$) and active comparator use ($\bar{\tilde{A}}_2$, where $\tilde{A}_2(t)$ denotes use of the active comparator, either SGLT-2 or DPP-4 use at time t). Specifically, we are interested in contrasting a regime of sustained GLP-1RA use throughout follow-up and no active comparator use throughout follow-up ($\bar{\tilde{a}}_1 = 1, \bar{\tilde{a}}_2 = 0$), with a regime of sustained non-use of GLP-1RAs throughout follow-up and sustained active comparator use throughout follow-up ($\bar{\tilde{a}}_1 = 0, \bar{\tilde{a}}_2 = 1$). In both regimes, we hypothetically intervene to prevent censoring ($\bar{c} = 0$). $E[Y_{\bar{\tilde{a}}_1=1,\bar{\tilde{a}}_2=0,\bar{c}=0}] - E[Y_{\bar{\tilde{a}}_1=0,\bar{\tilde{a}}_2=1,\bar{c}=0}]$, sustained GLP-1RA use, with no active comparator use and no administrative censoring throughout follow-up, compared to sustained active comparator use (e.g., SGLT2), with no GLP-1RA use and no administrative censoring throughout follow-up.

Observed data and causal model

The observed data available to estimate this casual estimand included all diabetes patients in the Danish National Registry who met the following inclusion criteria: were at least 50 years of age; had evidence of prior metformin use; initiated a second line medication between 2009 and 2021; were insulin-naive and dementia-free at index date (Supplemental Materials 2).

Data from this cohort can be described by a general longitudinal data structure. For ease of notation, we refer to the treatment and censoring processes together as $A(t) = (\tilde{A}_1(t), \tilde{A}_2(t), C(t))$ and also refer to the time-varying covariate information $\tilde{L}(t)$, incident dementia diagnosis $Y(t)$, and death (the competing risk) $D(t)$ collectively as $L(t) = (\tilde{L}(t), Y(t), D(t))$. We denote the individual-level observed longitudinal data as:

$$O = (W, L(1), A(1), ..., L(K), A(K), L(K+1)) \sim P_0,$$

where W denotes baseline (time-invariant) covariate information. Note that $\tilde{A}(1)$ is exposure status (second-line diabetes medication use) at the beginning of the first node or time interval (i.e. index date), and $c(t = 1) = 0$ by definition. We follow patients through the registers until dementia, death, emigration, or August 2021, whatever comes first.

Index date and baseline measurements are defined at the date of the start of second-line regimen. We use a discrete time scale with 6-month long time intervals. For notational convenience, we define variables after death, dementia diagnosis, or censoring as equal to their last observed value. For further explanation of the data-generating process, see Supplemental Materials 2.

We utilized prescription fill data from the Danish registry to define exposure to diabetes medications. Specifically, a patient was defined as exposed to a medication

at any time point if the patient redeemed at least one prescription of a medication in the corresponding time interval. For a full list of second-line diabetes medications used and relevant ATC codes, see Supplemental Materials 3.

We approximate our outcome, dementia onset, by time t ($Y(t)$), through either of the following criteria:(1) dementia diagnosis, measured through inpatient diagnosis codes; or (2) purchase of dementia medication, measured through registry medication fill table. For a list of ICD and ATC codes used, see Supplemental Materials 3. Death ($D(t)$) is treated as a competing risk. We consider the following baseline covariates: age (years), sex (male/female), education (basic, some college, college or higher), baseline income (tertiles), and time on metformin prior to index date (days). We also consider the following time-varying comorbidities: heart failure, renal disease, chronic pulmonary disease, any malignancy, ischemic heart disease, myocardial infarction, hypertension, and stroke. The time-varying comorbidities remain 0 until evidence of their presence through an inpatient ICD code in the registry is detected, then remain 1 for the remainder of follow up. We also include co-medication use, including: beta-blockers, calcium channel blockers, renin-angiotensin system-acting inhibitors, loop diuretics, mineralocorticoid receptor antagonists and chronic obstructive pulmonary disease (COPD) medications, which take the value of 1 if there is evidence of a fill in the time window. For a more detailed definition of all covariates, see Supplemental Materials 4.

Identification: causal to statistical estimand

The translation of our causal estimand into a statistical estimand (i.e., a function of the distribution of the observed data) requires causal assumptions about the data-generating process beyond simply the time ordering of covariates. The core assumption in the assessment of long-term cumulative treatment effects is the sufficiency of measured variables to adjust for confounding of treatment decisions as well as informative censoring. Such an assumption can be assessed either from the causal graph directly using the sequential back door criteria, or stated in the language of counterfactuals as the sequential randomization assumption. The sequential randomization assumption[7, 122] posits that at each time point, the counterfactual outcome $Y_{\bar{a}}$ is independent of treatment assignment and censoring, conditional on the history of measured covariates and prior treatment:

$$Y_{\bar{a}} \perp\!\!\!\perp A(t) \mid W, \bar{L}(t), \bar{A}(t-1) = \bar{a}(t-1), t = 1, .., K,$$

for every treatment history $\bar{a}$ of interest. Of note, this could equivalently be written to more explicitly incorporate the distinct elements of $L(t)$, as

$$Y_{\bar{a}} \perp\!\!\!\perp A(t) \mid W, Y(t) = 0, D(t) = 0, \bar{\bar{L}}(t), \bar{A}(t-1) = \bar{a}(t-1), t = 1, .., K,$$

for every treatment history $\bar{a}$ of interest (see Table 2.1). These definitions are equivalent because for the other possible values of $\bar{Y}(t)$ and $\bar{D}(t)$ it holds by definition.

For this assumption to hold, it is sufficient to assume in our causal model (Supplemental Materials 1), that we have no unmeasured common causes of our exposure and censoring nodes $A(t)$ at each time point and counterfactual outcome $Y_{\bar{a}}$. For each pairwise comparison, we consider other second-line medications not included in the contrast of interest as observable past.

We must also assume positivity—specifically, that there is a positive probability of continuing to follow the treatment of interest at each time point and no censoring, given previous treatment and covariate history[122, 7]:

$$P(A(t) = a(t) \mid \bar{A}(t-1) = \bar{a}(t-1), \bar{L}, W) > 0, t = 1, ..., K$$

for all possible covariate histories that can occur in the underlying data generating process among those who followed the regime. Threats to inference posed by violations (and near violations) of this assumption can be investigated empirically using the data, and used to inform specification of an estimator, as described further in the following section.

Under these assumptions, the causal estimands corresponding to the counterfactual difference in risk of dementia by five years under sustained GLP-1RA use compared to sustained use of an active comparator, can be translated into a target statistical estimand using Robins' longitudinal g formula.[23] In the present case study, we use the iterated conditional expectation form of this estimand.[40]

2.3 Statistical model and estimator

Prior to specifying a statistical estimator, the Roadmap makes clear that it is crucial to specify a statistical model that incorporates only assumptions on the observed data distribution (P_0) that are known to hold. The existence of death without dementia as a competing risk provides an example of such knowledge in our case study: after death, the probability of incident dementia is deterministically zero. While such a statement may seem obvious, we illustrate below its explicit incorporation into the estimator.

In addition to incorporating model knowledge, there are several decisions made in the estimation of the causal quantity of interest that require careful selection. In the following section, we highlight how simulations can be used to pre-specify an estimator and ensure its expected performance.

Simulation

Simulations are helpful tools that enable the selection of a valid and efficient estimation procedure. Simulations allow for the comparison of distinct estimation procedures. When registry data that are not always easily accessible due to privacy concerns, simulations also allow researchers to work with realistic longitudinal data remotely. We illustrate how simulations were used to pre-specify our statistical estimator, including the approach to inference. This includes the following choices: estimator choice (IPTW—commonly used in the literature, despite challenges—vs L-TMLE); approach to nuisance parameter estimation (specifically the pre-specification of machine learning algorithms to estimate the "treatment mechanism"-—conditional probabilities of medication use and censoring given the past—and iterated outcome regressions); and variance estimation (specifically, use of the empirical variance of the estimated influence function, an alternative "robust" variance estimator[80], or the non-parametric bootstrap).

Data generating process

To simulate data with the complex longitudinal structure found in the Danish registry, we ran logistic regression models on the real registry cohort of interest (Section 5.1.1). We used these models to estimate each node at each time point, conditional on the observable past. We then created a matrix of beta coefficients from each of these regressions and exported them from the Danish server.

Using these coefficients, we simulated n=500 datasets which contain follow-up data of 100,000 simulated patients. We simulated a first scenario (scenario 1) with no relationship between exposure and outcome, and a second scenario (scenario 2) with a significant relationship between exposure and outcome. We simulated the datasets using the lava package.[123] For each node at each time point, we used the vector of beta coefficients to predict the node values. This dataset has a dependent correlation structure between exposure and outcome, reflective of the parametric relationship fitted in the observed data. We then permuted the outcome, competing risk and censoring nodes in the dataset to create the scenario where the true risk difference between the treatment regimens is zero. This allowed us to approximate the complex correlation structure of treatment and covariate measures in the data.

We calculated the true causal risk difference between full GLP-1RA use throughout follow-up and no GLP-1RA use throughout follow-up in the first simulation scenario by simulating large datasets with counterfactual covariates and outcomes setting all exposure nodes ($\tilde{A}_1, \tilde{A}_2$) to 1 or 0, depending on the causal contrast of interest. For further detail, see Supplemental Materials 5.

Estimation procedure

We considered both IPTW estimator and L-TMLE as candidate estimators in the simulation. L-TMLE is a double-robust plug-in estimator that enables to control for time-dependent confounding without adjusting away the effect of the exposure on the outcome through downstream mediators. [6, 7] Nuisance parameters in L-TMLE estimation can incorporate machine learning algorithms while maintaining statistically valid inference under assumptions, including, informally, an assumption that initial estimators of the outcome regressions are not overfit.[6, 7] While the use of internal sample splits and cross-fitting approaches (such as cv-TMLE[7, 124]) can partially address these challenges, the lack of a computationally efficient implementation of the cv-TMLE estimator for the longitudinal treatment effect of interest led us to instead rely on simulations to carefully pre-specify the approach to nuisance parameter estimation.

We considered the following candidate algorithms for treatment propensity and censoring mechanisms and iterated conditional outcome regression estimation: logistic regression with LASSO penalty[125, 126] (one with cross-validated λ hyperparameter—the λ value with minimum mean cross-validated error—and one undersmoothed—choosing the minimum penalization across a range of candidate λ values), logistic regression with ridge and elastic net[125, 126] penalty (one with λ chosen to minimize CV error, and one undersmoothed), unpenalized logistic regression model (identity link, adjusted and unadjusted), and random forest[127]. We applied all candidate estimators and estimation procedures to both scenario 1 and secnario 2 data generating processes. Finally, prior work demonstrates that even when a point estimator performs well (in that its bias to standard error ratio is low enough to allow for valid statistical inference, implying nominal "oracle coverage" i.e. confidence interval coverage treating the true variance as known rather than estimated, calculated using the empirical variance across simulation repetitions), standard approaches to variance estimation based on the empirical variance of the estimated influence function, commonly used for both the IPTW and TMLE estimators, may underestimate the true variance, particularly in challenging settings with rare outcomes and/or practical positivity violations, leading to under-coverage.[80] Alternative approaches to variance estimation include a non-parametric bootstrap, and an approach based on fitting a separate TMLE of the variance itself, implemented as the "robust" option in the ltmle package.[80] As each of these approaches has its own vulnerabilities, we further employed simulations to pre-specify an approach to variance estimation and inference.

As described previously, the incorporation of death as a competing event for dementia onset implies knowledge about the statistical model; namely, that once a

death occurs, the probability of dementia is deterministically zero. By incorporating this knowledge in the estimation of the iterated outcome regression, the model will be correctly specified (i.e. will incorporate essential subject matter knowledge). We therefore examined all candidate estimators and algorithms described above with this knowledge explicitly incorporated (using the deterministic Q function in the *ltmle* R package, setting probability of the outcome deterministically to zero after death has occurred).[128] Code used for the simulation is available in an online repository.[129]

Simulation and estimation selection

We evaluated performance under each combination of the above specifications on 500 simulated datasets generated from our data-generating process described above. To evaluate the performance of the estimation procedures, we compared bias, variance, bias to standard error ratio, 95 % confidence interval coverage, and oracle coverage (Supplemental Materials 6).

In both scenario 1 and 2, we found the TMLE estimator to have nominal oracle coverage (i.e. closer to 95%), compared to the IPTW estimator. Among the TMLE candidate algorithms for nuisance parameter estimation (LASSO, Ridge, Elastic Net, Random Forest) in Scenario 1, we found the undersmoothed penalized regression algorithms with g-truncation performed best—i.e. had nominal oracle coverage and lowest variance (Supplemental Materials 5 and 6). We then compared variance estimation methods; we evaluated variance estimators for Scenario 1 only due to computational expense. The non-parametric bootstrap was the variance estimation option with the coverage closest to 95%; we found TMLE variance estimation was conservative and influence-curve-based estimation was anti-conservative.

Based on simulation results for the risk difference in scenario 1 (the realistic data generating process), we chose the L-TMLE estimator using undersmoothed ridge regression with g-truncation. This estimation strategy had the best oracle coverage, as well as low bias and variance. For more detailed results, see Supplemental Materials 6.

For all analyses, we used R version 4.2[130], and the *ltmle* package[128].

Ethical approval

Registry-based studies do not require informed consent or ethical approval in Denmark. The reference number from the Data Responsibility institute in the Capital Region of Denmark was P-2019-191.

2.4 Results

Cohort demographics

Of the registered diabetes patients using metformin in the Danish registry and initiating second-line treatment between 2009 and 2021 (n=154,128), 104,928 met inclusion criteria (Figure 2.1).

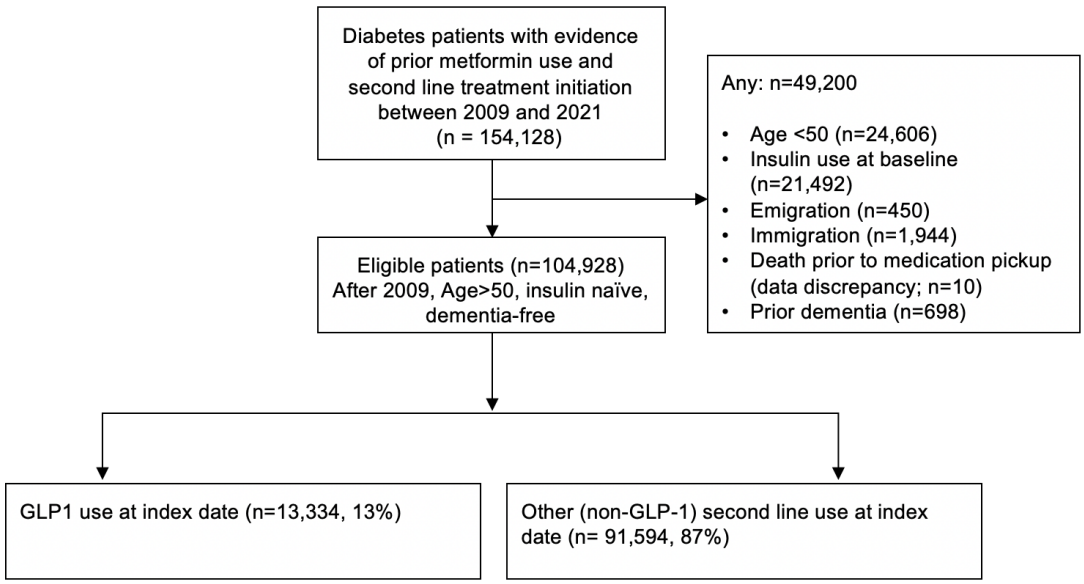


Figure 2.1: Flowchart of selection criteria for a retrospective cohort of diabetes patients initiating second-line therapy for the first time between 2009 and 2021 in Denmark

Among those meeting inclusion criteria for the study, 13% (n=13,334) initiated GLP-1RAs at baseline, and 87% (n=91,594) initiated other second-line medications (for further detail on demographics of other treatment medication initiators, see Supplemental Materials 7). Those who initiated GLP-1RAs at index date were more likely to be below 55, to be more educated, have a higher income, and have a longer duration of diabetes (Table 2.2). Those initiating GLP-1RAs at baseline were more likely to have some conditions (chronic pulmonary disease), and less likely to have others (heart failure, renal disease, ischemic heart disease, myocardial infarction).

Those initiating GLP-1RAs were more likely to be treated with Mineralocorticoid receptor antagonists (MRAs), Calcium channel blockers (CCBs), or Renin–angiotensin system inhibitors (RASIs).

Variable	Level	GLP1-RA (n=13,334)	no GLP1-RA (n=91,594)	Total (n=104,928)
Age category	< 55	2,851 (21.4)	12,106 (13.2)	14,957 (14.3)
	55-60	2,925 (21.9)	14,348 (15.7)	17,273 (16.5)
	60-65	2,583 (19.4)	15,942 (17.4)	18,525 (17.7)
	65-70	2,166 (16.2)	15,834 (17.3)	18,000 (17.2)
	70-75	1,557 (11.7)	14,078 (15.4)	15,635 (14.9)
	75-80	821 (6.2)	9,980 (10.9)	10,801 (10.3)
	80-85	306 (2.3)	5,823 (6.4)	6,129 (5.8)
	85-90	92 (0.7)	2,528 (2.8)	2,620 (2.5)
	> 90	33 (0.2)	955 (1.0)	988 (0.9)
Education	Basic	4,634 (34.8)	38,764 (42.3)	43,398 (41.4)
	Medium	6,471 (48.5)	41,043 (44.8)	47,514 (45.3)
	High	2,229 (16.7)	11,787 (12.9)	14,016 (13.4)
Income (tertile)	1	3,103 (23.3)	31,923 (34.9)	35,026 (33.4)
	2	3,943 (29.6)	31,008 (33.9)	34,951 (33.3)
	3	6,288 (47.2)	28,663 (31.3)	34,951 (33.3)
Diabetes duration (years)	< 1	3,747 (28.1)	31,795 (34.7)	35,542 (33.9)
	1-5	5,379 (40.3)	33,287 (36.3)	38,666 (36.9)
	5-10	3,082 (23.1)	20,932 (22.9)	24,014 (22.9)
	> 10	1,126 (8.4)	5,580 (6.1)	6,706 (6.4)
Diagnoses				
Heart failure		614 (4.6)	5,886 (6.4)	6,500 (6.2)
Renal disease		91 (0.7)	1,134 (1.2)	1,225 (1.2)
Chronic pulmonary disease		906 (6.8)	5,795 (6.3)	6,701 (6.4)
Any malignancy		689 (5.2)	6,366 (7.0)	7,055 (6.7)
Ischemic heart disease		1,416 (10.6)	11,310 (12.3)	12,726 (12.1)
Myocardial infarction		675 (5.1)	5,149 (5.6)	5,824 (5.6)
Hypertension		2,876 (21.6)	20,249 (22.1)	23,125 (22.0)

Stroke	454 (3.4)	4,126 (4.5)	4,580 (4.4)
Medications			
Beta-blockers	3,655 (27.4)	27,185 (29.7)	30,840 (29.4)
Calcium channel blockers	4,107 (30.8)	27,105 (29.6)	31,212 (29.7)
Renin-angiotensin system-acting inhibitors	8,710 (65.3)	57,552 (62.8)	66,262 (63.1)
Loop diuretics	2,112 (15.8)	14,538 (15.9)	16,650 (15.9)
Mineralocorticoid receptor antagonists	897 (6.7)	5,495 (6.0)	6,392 (6.1)
COPD medications	2,108 (15.8)	11,781 (12.9)	13,889 (13.2)

Table 2.2: Baseline demographic characteristics of a cohort of diabetes patients initiating second line treatment for the first time between 2009 and 2021 in the Danish registry, stratified by GLP1-RA initiation at baseline. N(%)

Risk differences

We applied our pre-specified estimator, LTMLE, with propensity scores estimated using undersmoothed ridge regression, and inference based on the non-paramteric bootstrap, to the registry data. The results show a protective effect comparing five-year sustained GLP-1RA use to sustained non-use (risk difference percent RD% -0.48, 95% CI -0.94,-0.01; Table 2.3) and to sustained use of DPP4 inhibitors (RD% -0.48, 95% CI -0.93,-0.02; Table 2.3). No significant differences were found comparing GLP-1RAs to SGLT2 inhibitors (RD% -0.48, 95% CI -0.93,-0.02; Table 2.3). In sensitivity analyses, we varied the time interval length, and found these results to hold irrespective of time discretization (Supplemental Materials 8); shortening the time unit led to a wider confidence interval. We include relative risk results in Supplemental Materials 9.

Comparison	Treatment Risk	Control Risk	Risk Difference % (RD%, 95% CI)
1) GLP1 vs. no GLP1	0.0093	0.0141	-0.47(-0.94,-0.01)
2) GLP1 vs. SGLT2	0.0071	0.0074	-0.03 (-0.27,0.22)
3) GLP1 vs. DPP4	0.0068	0.0115	-0.48 (-0.93,-0.02)

Table 2.3: Five-year risk of dementia under sustained GLP1-RA use vs sustained comparator use in a cohort of Danish diabetes patients initiating second-line therapy (RD represented as %, rescaled x100)

We additionally plotted the marginal risk of dementia at each six-month interval, among those who have survived until that time point and have followed the longitudinal treatment regimes of interest (sustained GLP-1RA use, sustained non-use) up until that time point. Results show a difference in the risk of dementia over time between the longitudinal exposure groups (Figure 2.2; Table 2.4).

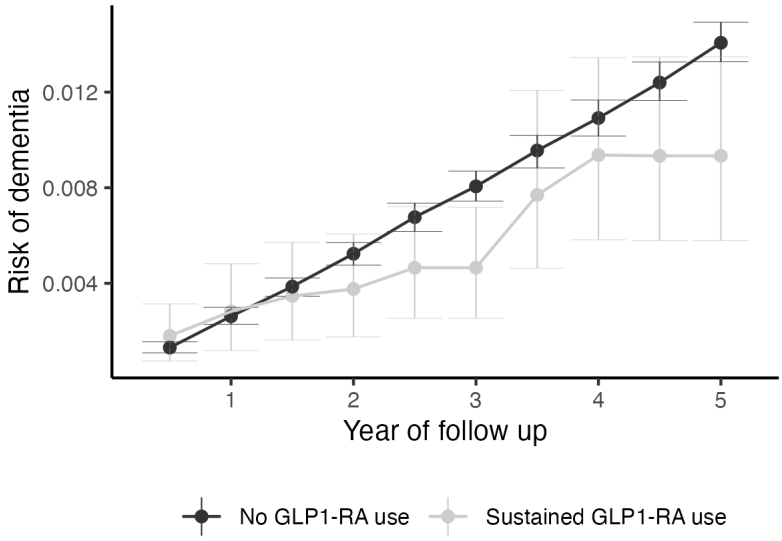


Figure 2.2: The L-TMLE dementia risk estimates comparing sustained use and non-use of GLP-1RAs in a cohort of Danish second-line diabetes drug initiators 2009-2021.

Interval	0	1	2	3	4	5	6	7	8	9	10
GLP1-RA use	13,334	10,419	7,627	6,008	4,983	4,282	3,587	3,084	2,746	2,484	2,216
Death	0	96	82	63	59	46	40	36	47	43	45
Dementia	0	9	<5*	<5*	<5*	<5*	<5*	5	<5*	<5*	<5*
End of´ follow-up	0	1,992	2,052	1,456	995	686	714	468	357	269	290
Non-adherent	0	818	1,473	1,571	1,540	1,507	1,447	1,441	1,372	1,321	1,253
No GLP1-RA use	91,594	82,666	74,352	67,744	61,901	56,768	51,814	47,492	43,221	39,505	35,552
Death	0	1,740	1,361	1,243	1,137	1,094	977	967	925	876	814
Dementia	0	114	111	99	94	107	88	85	78	79	79
End of´ follow-up	0	4,525	5,126	4,039	3,871	3,327	3,688	3,078	3,456	2,911	3,277
Non-adherent	0	2,549	4,265	5,492	6,233	6,838	7,039	7,231	7,043	6,893	6,676

Table 2.4: Table of longitudinal patterns in treatment and control groups in a cohort of diabetes patients initiating second line use between 2009 and 2021 in Denmark
*cells with less than five observations are not shown in the interest health information privacy

2.5 Discussion

In this case study, we apply the Casual Roadmap to a longitudinal analysis of a large national registry database. Through the use of the Roadmap, we explicitly state our causal estimand and the assumptions required to be able to estimate this quantity from the observed data. We found a significant relationship between sustained GLP-1RA use and five-year dementia risk compared to no usage of GLP-1RAs and compared to sustained DPP4 use, after adjusting for a wide range of time invariant and time-varying confounders using a pre-specified L-TMLE estimator that incorporates machine learning. The relationship of GLP-1RAs compared to active comparator SGLT2 inhibitors on dementia diagnosis was not significant; however, confidence intervals were wide due to rare sustained usage of SGLT2 and greater administrative censoring as SGLT2 is a newer drug (usage increasing after 2011). High administrative censoring of the treatment group is likely due to the later introduction and relatively slower adoption of GLP-1RAs relative to other, more established second-line therapies.[131] Our findings agree with prior literature on the benefit of GLP-1RAs on dementia onset in the Danish registry[22], observational studies[13, 120], and some randomized trials[21, 14, 15, 132].

As with all studies, ours is not without limitations. Firstly, due to data unavailability we were not able to adjust for smoking status or body mass index, which may have resulted in residual confounding of our effect estimate—leading to either under or overestimates of the protective association of sustained GLP1-RA use. In addition, while this sample is generalizable nationally to Denmark and to other Nordic countries, the ethnically homogeneous population in the Danish registry does not necessarily generalize to more diverse populations. Finally, the data from our simulation was generated using coefficients from sequential regression models; this imposes a simpler parametric structure on the data generating process than may exist in reality. This simplified structure may lead to some overestimation of model performance, as some of the complexities of the true data distribution are not captured.

We found the implementation of a complex estimator such as L-TMLE with machine learning at the scale national registry to be feasible. However, implementation of such an analysis involves a variety of estimation decisions. Simulation provided a means to pre-specify these decisions and evaluate the performance of the resulting estimator in data generating processes designed to resemble the data, including to evaluate the performance (precision and confidence interval coverage) in the presence of rare events and long-term follow up. The simulations also provided additional evidence that if implemented with care, a TMLE estimator with machine learning can reduce bias and variance and improve inference relative to both IPTW estima-

tors and TMLE estimators implemented without machine learning. Nonetheless, in this setting or rare outcomes and practical postivity violations, variance estimation remained a challenge. A common approach used for IPTW, TMLE, and other estimators, based on the empirical variances of the estimated influence curve, resulted in markedly anti-conservative confidence interval coverage, as has been reported by others.[80] While the non-parametric bootstrap provided coverage closer to nominal it was still slightly anti-conservative. Future work is needed to further improve variance estimation for complex longitudinal estimators in settings such as these. Future work may also consider implementing a continuous-time TMLE, which would circumvent any potential measurement error introduced through time discretization. Software for this kind of analysis was released after the present analysis was conducted.

While simulation results presented are interesting in their own right, the results must be interpreted locally to this particularly analysis. We recommend that investigators construct their own simulations with data similar to their own in key ways, rather than follow the estimation specifications here or use estimation defaults. To our knowledge, this is the first study of its kind of apply L-TMLE using longitudinal Causal Roadmap [24, 101] to a large national registry database.

Supplemental Materials

Contents

1 Structural causal model **69**

2 Cohort selection and data processing **70**

3 Treatment and outcome definitions **72**

4 Covariate definitions **73**

5 Simulation methods **75**

6 Simulation results **77**
6.1 Scenario 1: Realistic simulation, protective effect of GLP1 on dementia 77
6.1.1 Performance of difference point estimators (scenario 1) 78
6.1.2 Performance of difference variance estimators (scenario 1) . . . 80
6.2 Scenario 2: Realistic simulation, no exposure-outcome relationship . . 81
6.2.1 Comparison of different estimators' performance (scenario 2) . 81

7 Baseline demographic characteristics by baseline treatment **83**

8 Supplemental analysis: time discretization **86**

9 Relative risk results **87**

1 Structural causal model

We define the following exogenous nodes for $t = 1...k+1$:

$$X = W, L(t), A(t)$$

where: W includes all time-constant covariates: age (years), sex (male/female), education (basic, some college, college or higher), baseline income (tertiles), time on metformin prior to index date (days).

$$A(t) = (\tilde{A}_1(t), \tilde{A}_2(t), C(t))$$

- $\tilde{A}_1(t)$ is GLP1-RA use at time t
- $\tilde{A}_2(t)$ is active comparator use (however designated) at time t
- $C(t)$ is administrative censoring at time t

$$L(t) = (\tilde{L}(t), Y(t), D(t))$$

- $\tilde{L}(t)$ is the time-varying covariate status at time t, these include diagnoses (heart failure, renal disease, chronic pulmonary disease, any malignancy, ischemic heart disease, myocardial infarction, hypertension, stroke) and medication use (beta-blockers, calcium channel blockers, renin-angiotensin system-acting inhibitors (RASI), loop diuretics, Mineralocorticoid receptor antagonists (MRAs) and COPD medications)
- $Y(t)$ is the outcome, dementia diagnosis, at time t
- $D(t)$ is death at time t

And exogenous error:

$$U = (U_A, U_L)$$

Structural equations are as follows:

$$L(t) = f_{L(t)}(\bar{A}(t-1), \bar{L}(t-1), U_{L(t)}), t = 1, ..., K+1$$

$$A(t) = f_{A(t)}(\bar{A}(t-1), \bar{L}(t-1), U_{A(t)}), t = 1, ..., K$$

2 Cohort selection and data processing

We collected a retrospective cohort of diabetes patients who:

- were at least 50 years of age
- had at least one medication fill for metformin use (Table 1)
- did not immigrate to Denmark in the five years prior to beginning a second line therapy or emigrate before the start of the study
- initiated a second line medication (Table 2) between January 1st, 2009 and July 20th, 2022 (second line start date as the index date)
- were insulin-naive at index date (Table 1)
- were dementia-free at index date (Table 1)

Table 1: List of codes used for inclusion and exclusion criteria in cohort of second line diabetes drug initiators

drug	ATC
Insulin	$A10A$
Metformin	$A10BA02$, $A10BD02$, $A10BD03$, $A10BD05$, $A10BD07$, $A10BD08$, $A10BD10$, $A10BD11$, $A10BD13$, $A10BD14$, $A10BD15$, $A10BD16$, $A10BD17$, $A10BD18$, $A10BD20$, $A10BD22$, $A10BD23$, $A10BD25$, $A10BD26$

Table 2: List of codes used to define second line diabetes medication use, for inclusion and exposure node definition

Node	Drug	ATC Codes
$\tilde{A}_1$	GLP1-RA	A10BJ
$\tilde{A}_2$	SGLT2 Inhibitors	A10BK, A10BD15, A10BD16, A10BD19, A10BD20, A10BD21, A10BD23, A10BD24, A10BD25
	Sulfonylureas	A10BB, A10BD01, A10BD02, A10BD04, A10BD06
	DPP-4 Inhibitors	A10BH, A10BD07, A10BD08, A10BD09, A10BD10, A10BD11, A10BD12, A10BD13, A10BD18, A10BD19, A10BD21, A10BD22, A10BD24, A10BD25
	Thiazolidinediones	A10BG, A10BD03, A10BD04, A10BD05, A10BD06, A10BD09, A10BD12, A10BD26
	Sulfonamides	A10BC
	Alfa-glucosidase Inhibitors	A10BF, A10BD17
	Repaglinides	A10BX, A10BD14
	Biguanides	A10BA01, A10BA03, A10BD01

3 Treatment and outcome definitions

Treatment exposure: Usage of each second-line medication (Table 2) was defined for each time node (6-month bins in main analyses) if the medication coverage time (medication pick up date +6 months) overlapped with the time node. **Outcome:** Dementia onset date was defined as the first occurrence of either (1) a dementia medication pickup, or (2) a dementia diagnosis during follow up. (Table 3)

Table 3: ATC and ICD codes used to define dementia endpoint

Code	Type	Description
N06DA01	ATC	Tacrine
N06DA02	ATC	Donepezil
N06DA03	ATC	Rivastigmine
N06DA04	ATC	Galantamine
N06DA05	ATC	Ipidacrine
N06DA52	ATC	Donepezil and memantine
N06DA53	ATC	Donepezil, memantine and Ginkgo folium
290	ICD-08	Dementias
F00	ICD-10	Dementia in Alzheimer disease
F01	ICD-10	Vascular dementia
F02	ICD-10	Dementia in other diseases classified elsewhere without behavioral disturbance
F03	ICD-10	Unspecified dementia
G30	ICD-10	Alzheimer's disease
G31.1	ICD-10	Senile degeneration of brain, not elsewhere classified
F051	ICD-10	Delirium superimposed on dementia
331	ICD-08	Alzheimer's disease

4 Covariate definitions

Table 4: Covariate names, definitions and relevant codes
for all variables included in main analyses

Variable	Definition and codes
Age category	Age of patient at baseline
Education	Patient education prior to baseline; basic (elementary), medium (trade school; some college) or high (college +)
Income	Income range categorized into tertiles
Diabetes duration	Evidence of diabetes duration through diabetes diagnosis or evidence of metformin use; years from first evidence up until baseline
Diagnoses	Time-varying covariates; first evidence of ICD-08 or ICD-10 diagnosis code (below) prior to baseline or during follow up. Absence of diagnosis coded as 0 and presence as 1.
Heart failure	42709, 42710, 42711, 42719, 42899, 78249, DI099, DI110, DI130, DI132, DI255, DI425, DI426, DI427, DI429, DI428A, DP290, DI43, DI50, DE105, DE115, DE125, DE135, DE145
Renal disease	403, 404, 580, 581, 582, 583, 584, 59009, 59319, 75310, 75311, 75312, 75313, 75314, 75315, 75316, 75317, 75318, 75319, 792, DN032, DN033, DN034, DN035, DN036, DN037, DN052, DN053, DN054, DN055, DN056, DN057, DZ490, DZ491, DZ492, DN18, DN19, DI120, DI131, DI132, DN250, DZ940, DZ992, DN26
Chronic pulmonary disease	515, 516, 517, 518, 490, 491, 492, 493, DJ40, DJ41, DJ42, DJ43, DJ44, DJ45, DJ46, DJ47, DJ60, DJ61, DJ62, DJ63, DJ64, DJ65, DJ66, DJ67, DJ684, DI278, DI279, DJ84, DJ701, DJ703, DJ920, DJ953, DJ961, DJ982, DJ983

Any malignancy	140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 27559, DC0, DC1, DC2, DC3, DC40, DC41, DC42, DC43, DC44, DC45, DC46, DC47, DC48, DC49, DC5, DC6, DC70, DC71, DC72, DC73, DC74, DC75, DC76, DC86, DC97
Ischemic heart disease	411, 412, 413, 414, DI20, DI23, DI24, DI25
Myocardial infarction	410, DI21
Hypertension	41009, 41099, DI10, DI109, DI11, DI110, DI119, DI119A, DI12, DI120, DI129, DI129A, DI13, DI130, DI131, DI132, DI139, DI15, DI150, DI151, DI152, DI158, DI159, 401, 402, 403, 404
Stroke	433.434.435.436.437.438.DI63.DI64
Medications	Time-varying covariates; in each time node, patients are classified as using the medication if a medication pickup with relevant ATC code (below) was made on a date falling in time bin. Absence of medication coded as 0 and presence as 1.
Beta-blockers	C07
Calcium channel blockers	C08
Renin-angiotensin system-acting inhibitors (RASI)	C09
Loop diuretics	C03C, C03EB
Mineralocorticoid receptor antagonists (MRAs)	C03D
COPD medications	R03BA, R03AK06, R03AK07, R03AK08, R03AK09, R03AK10, R03AK11, R03AK12, R03AL08, R03AL09, R03AC, R03AK06, R03AK07, R03AK08, R03AK09, R03AK10, R03AK11, R03AK12, R03AK13, R03AL01, R03AL02, R03AL03, R03AL04, R03AL05, R03AL06, R03AL07, R03AL08, R03AL09, R03BB

5 Simulation methods

We conducted a simulation study to find the best estimator for a longitudinal targeted maximum likelihood estimation (L-TMLE) analysis of the effect of second-line diabetes drugs on dementia risk. The data was simulated to approximate the features and multivariate distribution of the Danish National Registry analysis dataset used for the primary analysis, which has five key features: many time points (10), a relatively rare treatment regime of interest, a rare outcome and exposure , competing risks from death, and a high degree of administrative censoring. Two simulation scenarios were used:

1. A simulation using data with a similar multivariate distribution of treatment regimes A, and covariates L, and outcomes Y to the Danish registry data. The truth was calculated as the counterfactual 5-year risk of dementia prior to death when continuously on GLP-1RA's versus not, with the effect of GLP-1RA's on death removed to eliminate the competing risk.

2. A simulation using data with a similar distribution of treatment regimes A and covariates L to the Danish registry data but with permuted outcomes Y such that there was no association between A and Y and the true treatment effect is null.

For both simulation scenarios, we estimated the effect of 5-year continuous usage of GLP-1RA compared to no GLP-1RA usage, analogous to the target causal parameter 1 (Table 1 of main text, column 1).

Fore each of the two data generating processes outlined above, we examined:

- *Estimator selection*: We considered both inverse-probability of treatment weighting (IPTW) and targeted maximum likelihood estimation (TMLE)

- *Nuisance parameter estimation*:
  We considered in the simulation the following candidate algorithms for g and Q estimation:

  - LASSO (λ (lambda) chosen to minimize cross-validated error, and undersmoothed*)
  - Ridge (λ chosen to minimize cross-validated error, and undersmoothed*)
  - Elastic net ($\alpha = 0.5$ and λ chosen to minimize cross-validated error, and undersmoothed*)
  - Generalized linear models (identity link)
  - Random forest

- For each of the penalized regressions above (LASSO, Ridge, Elastic net) we considered one with a cross-validated tuning parameter selection, and one undersmoothed version which utilized the minimum λ value (i.e. the least penalization) that did not fail across a range of candidate values.

- *Variance estimation*: To estimate variance, we considered empirical influence curve variance, "robust" variance estimator [2], and the non-parametric bootstrap.

We estimated the performance of each combination of the above specifications on 500 datasets generated from our data-generating process described above. To evaluate performance of the estimation procedures, we compared oracle coverage, estimator bias (σ^2), and estimator variance of the risk difference (see Section 7). These quantities were calculated using the true parameter value (Ψ_0), the estimated parameter value ($\hat{\psi}_i$) for each iteration of the simulated data i where $i = 1....500$:

Bias: $E(\hat{\psi}) - \Psi_0$

Variance: $E[\hat{\psi}^2] - [E(\hat{\psi})]^2$

Bias/SE ratio: $(E(\hat{\psi}) - \Psi_0)/(E[\hat{\psi}^2] - [E(\hat{\psi})])$

Oracle Coverage: $E(I_{OCI_{lb} \leq \Psi_0 \leq OCI_{ub}})$

Where the Oracle upper (OCI_{ub}) and lower bound (OCI_{lb}) are calculated using the standard deviation of the estimator across simulation repetitions: $\Psi_i \pm 1.96\sigma_{est}$

Code used for the simulation is available on Github.[1]

6 Simulation results

6.1 Scenario 1: Realistic simulation, protective effect of GLP1 on dementia

True disk difference -0.0.00719; Outcome prevalence 0.83% in treated group, 1.55% in control group

6.1.1 Performance of difference point estimators (scenario 1)

The estimator chosen, based on the lowest variance among estimators with an oracle coverage within 1% of 95%, is bolded and highlighted with lines. Here, for the risk difference with data scenario 1, the method chosen from the simulation results is L-TMLE using undersmoothed ridge with g-truncation. Importantly, the undersmoothed option for various algorithms tend to perform better than the default lambda parameter selection. *The bias and variance columns are scaled by 100 for better readability.

Estimator	Algorithm	Lambda	Truncation	RD bias x 100*	RD variance x 100*	RD bias SE ratio	RD oracle 95% coverage
TMLE	**Ridge**	**Undersmoothed**	**Less than 0.01**	**0.185123**	**0.000491**	**0.835818**	**95.00000**
TMLE	Elastic Net	Undersmoothed	Less than 0.01	0.196316	0.000400	0.981080	95.60000
TMLE	LASSO	Undersmoothed	Less than 0.01	0.196423	0.000401	0.980587	95.60000
TMLE	GLM	N/A	Less than 0.01	0.204626	0.000440	0.975826	95.80000
TMLE	Ridge	Undersmoothed	Untruncated	0.225100	0.001060	0.691243	96.40000
TMLE	Elastic Net	Undersmoothed	Untruncated	0.247762	0.000891	0.829886	96.20000
TMLE	LASSO	Undersmoothed	Untruncated	0.248398	0.000900	0.827790	96.20000
TMLE	Elastic Net	CV-minimum SE	Untruncated	0.253776	0.000268	1.549389	73.40000
TMLE	Elastic Net	CV-minimum SE	Less than 0.01	0.254413	0.000258	1.585283	71.00000
TMLE	Ridge	CV-minimum SE	Untruncated	0.255141	0.000447	1.207114	76.20000
TMLE	GLM	N/A	Untruncated	0.255202	0.001072	0.779565	95.60000
TMLE	LASSO	CV-minimum SE	Untruncated	0.255850	0.000260	1.585686	73.00000
TMLE	LASSO	CV-minimum SE	Less than 0.01	0.256470	0.000251	1.618536	70.20000
TMLE	Ridge	CV-minimum SE	Less than 0.01	0.261435	0.000371	1.356658	68.60000
TMLE	Random Forest	N/A	Less than 0.01	0.277126	0.000142	2.325057	92.00000
IPTW	GLM	N/A	Untruncated	0.421628	0.001177	1.228763	54.80000
IPTW	Ridge	Undersmoothed	Less than 0.01	0.421628	0.000508	1.870010	54.80000
IPTW	GLM	N/A	Less than 0.01	0.421628	0.000456	1.974381	54.80000
IPTW	LASSO	Undersmoothed	Less than 0.01	0.421628	0.000415	2.070313	54.80000

IPTW	Elastic Net	Undersmoothed	Less than 0.01	0.421628	0.000414	2.072648	54.80000
IPTW	Ridge	CV-minimum SE	Less than 0.01	0.421628	0.000379	2.166498	54.80000
IPTW	Elastic Net	CV-minimum SE	Untruncated	0.421628	0.000272	2.556702	54.80000
IPTW	LASSO	CV-minimum SE	Untruncated	0.421628	0.000264	2.596233	54.80000
IPTW	Elastic Net	CV-minimum SE	Less than 0.01	0.421628	0.000260	2.613575	54.80000
IPTW	LASSO	CV-minimum SE	Less than 0.01	0.421628	0.000254	2.647554	54.80000
IPTW	Ridge	CV-minimum SE	Untruncated	0.432669	0.000456	2.026333	53.77778
IPTW	Ridge	Undersmoothed	Untruncated	0.433075	0.001127	1.290068	53.58491
IPTW	Random Forest	N/A	Less than 0.01	0.434636	0.000127	3.864224	53.60000
IPTW	LASSO	Undersmoothed	Untruncated	0.446782	0.000945	1.453555	50.60606
IPTW	Elastic Net	Undersmoothed	Untruncated	0.446782	0.000934	1.462082	50.60606

6.1.2 Performance of difference variance estimators (scenario 1)

Comparing different variance estimators using the estimator and estimation method chosen above, we find that the influence curve (ic) method is anti-conservative, and the TMLE method is conservative, with very wide confidence interval widths. The bootstrap estimator performs the best of the different variance options.

Variance.estimator	RD Coverage
Influence curve	86.4
Bootstrap	95
TMLE	99.4

6.2 Scenario 2: Realistic simulation, no exposure-outcome relationship

The second simulation was done to benchmark different estimator performance using data with a realistic prevalence of dementia and 5 years of follow-up, but with a known null association (True Risk Difference=0) between treatments and the dementia outcome.

6.2.1 Comparison of different estimators' performance (scenario 2)

In Scenario 2, g truncation does not make a sizeable difference for CV-min SE estimators with penalized regression. The undersmoothed truncated penalized regression algorithms here perform similarly. *The bias and variance columns are scaled by 100 for better readability.

Estimator	Algorithm	Lambda	Truncation	RD bias x 100*	RD variance x 100*	RD bias SE ratio	RD oracle 95% coverage
TMLE	LASSO	CV-minimum SE	Less than 0.01	0.159940	0.000315	0.901633	96.0
TMLE	LASSO	CV-minimum SE	Untruncated	0.159944	0.000315	0.901413	96.0
TMLE	Elastic Net	CV-minimum SE	Less than 0.01	0.160203	0.000328	0.884646	96.0
TMLE	Elastic Net	CV-minimum SE	Untruncated	0.160220	0.000328	0.884224	96.0
TMLE	Ridge	CV-minimum SE	Less than 0.01	0.161497	0.000592	0.663841	95.4
TMLE	Ridge	CV-minimum SE	Untruncated	0.162864	0.000605	0.662149	95.8
TMLE	**Ridge**	**Undersmoothed**	**Less than 0.01**	**0.168897**	**0.000601**	**0.688838**	**94.4**
TMLE	Elastic Net	Undersmoothed	Less than 0.01	0.173586	0.000362	0.912383	94.4
TMLE	LASSO	Undersmoothed	Less than 0.01	0.173656	0.000362	0.913133	94.4
TMLE	Ridge	Undersmoothed	Untruncated	0.173755	0.000658	0.677415	95.0
TMLE	GLM	N/A	Less than 0.01	0.174666	0.000399	0.874241	95.0
TMLE	Elastic Net	Undersmoothed	Untruncated	0.181858	0.000421	0.886583	94.8
TMLE	LASSO	Undersmoothed	Untruncated	0.182133	0.000422	0.886609	94.8
TMLE	GLM	N/A	Untruncated	0.186551	0.000494	0.839255	95.4
IPTW	Ridge	Undersmoothed	Untruncated	0.264168	0.000678	1.014801	88.2

IPTW	Ridge	CV-minimum SE	Untruncated	0.264168	0.000623	1.058681	88.2
IPTW	Ridge	Undersmoothed	Less than 0.01	0.264168	0.000615	1.064925	88.2
IPTW	Ridge	CV-minimum SE	Less than 0.01	0.264168	0.000607	1.071845	88.2
IPTW	GLM	N/A	Untruncated	0.264168	0.000511	1.168190	88.2
IPTW	LASSO	Undersmoothed	Untruncated	0.264168	0.000435	1.266778	88.2
IPTW	Elastic Net	Undersmoothed	Untruncated	0.264168	0.000433	1.268788	88.2
IPTW	GLM	N/A	Less than 0.01	0.264168	0.000406	1.310899	88.2
IPTW	Elastic Net	Undersmoothed	Less than 0.01	0.264168	0.000368	1.376521	88.2
IPTW	LASSO	Undersmoothed	Less than 0.01	0.264168	0.000368	1.377126	88.2
IPTW	Elastic Net	CV-minimum SE	Untruncated	0.264168	0.000335	1.443315	88.2
IPTW	Elastic Net	CV-minimum SE	Less than 0.01	0.264168	0.000334	1.444487	88.2
IPTW	LASSO	CV-minimum SE	Untruncated	0.264168	0.000321	1.474387	88.2
IPTW	LASSO	CV-minimum SE	Less than 0.01	0.264168	0.000321	1.475018	88.2

7 Baseline demographic characteristics by baseline treatment

SGLT2i

Variable	Level	SGLT2i (n=22434)	no SGLT2i (n=82494)	Total (n=104928)
Age groups	Below 55	3,329 (14.8)	11,628 (14.1)	14,957 (14.3)
	55-60	4013 (17.9)	13260 (16.1)	17,273 (16.5)
	60-65	4164 (18.6)	14361 (17.4)	18,525 (17.7)
	65-70	3769 (16.8)	14231 (17.3)	18,000 (17.2)
	70-75	3526 (15.7)	12109 (14.7)	15,635 (14.9)
	75-80	2296 (10.2)	8505 (10.3)	10,801 (10.3)
	80-85	1003 (4.5)	5126 (6.2)	6,129 (5.8)
	85-90	269 (1.2)	2351 (2.8)	2,620 (2.5)
	Above 90	65 (0.3)	923 (1.1)	988 (0.9)
Education	Basic	8,022 (35.8)	35,376 (42.9)	43,398 (41.4)
	Medium	11113 (49.5)	36401 (44.1)	47,514 (45.3)
	High	3299 (14.7)	10717 (13.0)	14,016 (13.4)
Income (tertile)	Q1	5,757 (25.7)	29,269 (35.5)	35,026 (33.4)
	Q2	7163 (31.9)	27788 (33.7)	34,951 (33.3)
	Q3	9514 (42.4)	25437 (30.8)	34,951 (33.3)
Diabetes duration (years)	Below 1	5,489 (24.5)	30,053 (36.4)	35,542 (33.9)
	1-5	8193 (36.5)	30473 (36.9)	38,666 (36.9)
	5-10	6501 (29.0)	17513 (21.2)	24,014 (22.9)
	Above 10	2251 (10.0)	4455 (5.4)	6,706 (6.4)
Diagnoses				
Heart failure	Yes	1599 (7.1)	4901 (5.9)	6,500 (6.2)
Renal disease	Yes	138 (0.6)	1087 (1.3)	1,225 (1.2)
Chronic pulmonary disease	Yes	1224 (5.5)	5477 (6.6)	6,701 (6.4)
Any malignancy	Yes	1316 (5.9)	5739 (7.0)	7,055 (6.7)
Ischemic heart disease	Yes	2635 (11.7)	10091 (12.2)	12,726 (12.1)
Myocardial infarction	Yes	1561 (7.0)	4263 (5.2)	5,824 (5.6)
Hypertension	Yes	4573 (20.4)	18552 (22.5)	23,125 (22.0)
Stroke	Yes	929 (4.1)	3651 (4.4)	4,580 (4.4)
Medications				
Beta blockers	Yes	7047 (31.4)	23793 (28.8)	30,840 (29.4)
Calcium channel blockers	Yes	6938 (30.9)	24274 (29.4)	31,212 (29.7)
Renin-angiotensin system-acting inhibitors	Yes	14811 (66.0)	51451 (62.4)	66,262 (63.1)
Loop diuretics	Yes	3114 (13.9)	13536 (16.4)	16,650 (15.9)
Mineralocorticoid receptor antagonists	Yes	1797 (8.0)	4595 (5.6)	6,392 (6.1)
COPD medications	Yes	3002 (13.4)	10887 (13.2)	13,889 (13.2)

DPP4

Variable	Level	DPP4 (n=40569)	no DPP4 (n=64359)	Total (n=104928)
Age groups	Below 55	5,438 (13.4)	9,519 (14.8)	14,957 (14.3)
	55-60	6326 (15.6)	10947 (17.0)	17,273 (16.5)
	60-65	7009 (17.3)	11516 (17.9)	18,525 (17.7)
	65-70	6998 (17.2)	11002 (17.1)	18,000 (17.2)
	70-75	6145 (15.1)	9490 (14.7)	15,635 (14.9)
	75-80	4327 (10.7)	6474 (10.1)	10,801 (10.3)
	80-85	2633 (6.5)	3496 (5.4)	6,129 (5.8)
	85-90	1224 (3.0)	1396 (2.2)	2,620 (2.5)
	Above 90	469 (1.2)	519 (0.8)	988 (0.9)
Education	Basic	16,919 (41.7)	26,479 (41.1)	43,398 (41.4)
	Medium	18331 (45.2)	29183 (45.3)	47,514 (45.3)
	High	5319 (13.1)	8697 (13.5)	14,016 (13.4)
Income (tertile)	Q1	13,884 (34.2)	21,142 (32.9)	35,026 (33.4)
	Q2	14095 (34.7)	20856 (32.4)	34,951 (33.3)
	Q3	12590 (31.0)	22361 (34.7)	34,951 (33.3)
Diabetes duration (years)	Below 1	11,377 (28.0)	24,165 (37.5)	35,542 (33.9)
	1-5	16287 (40.1)	22379 (34.8)	38,666 (36.9)
	5-10	10307 (25.4)	13707 (21.3)	24,014 (22.9)
	Above 10	2598 (6.4)	4108 (6.4)	6,706 (6.4)
Diagnoses				
Heart failure	Yes	2445 (6.0)	4055 (6.3)	6,500 (6.2)
Renal disease	Yes	626 (1.5)	599 (0.9)	1,225 (1.2)
Chronic pulmonary disease	Yes	2661 (6.6)	4040 (6.3)	6,701 (6.4)
Any malignancy	Yes	2953 (7.3)	4102 (6.4)	7,055 (6.7)
Ischemic heart disease	Yes	4799 (11.8)	7927 (12.3)	12,726 (12.1)
Myocardial infarction	Yes	1991 (4.9)	3833 (6.0)	5,824 (5.6)
Hypertension	Yes	9215 (22.7)	13910 (21.6)	23,125 (22.0)
Stroke	Yes	1788 (4.4)	2792 (4.3)	4,580 (4.4)
Medications				
Beta blockers	Yes	11801 (29.1)	19039 (29.6)	30,840 (29.4)
Calcium channel blockers	Yes	11979 (29.5)	19233 (29.9)	31,212 (29.7)
Renin-angiotensin system-acting inhibitors	Yes	25598 (63.1)	40664 (63.2)	66,262 (63.1)
Loop diuretics	Yes	6553 (16.2)	10097 (15.7)	16,650 (15.9)
Mineralocorticoid receptor antagonists	Yes	2162 (5.3)	4230 (6.6)	6,392 (6.1)
COPD medications	Yes	5319 (13.1)	8570 (13.3)	13,889 (13.2)

Chapter 3

Simulations to pre-specify estimators in causal analyses: an illustration of considerations using a diabetes medication and cardiovascular outcomes from registry data

3.1 Introduction

Simulation has long been used to inform study design. One of the most common uses of simulation in study design is power calculations[133], but increasingly simulations have been used to guide other aspects of the analysis as well. These uses include defining the primary endpoint in a randomized trial in light of data support concerns[134, 135], or determining the number and timing of interim analyses in adaptive trials[136, 137], among others. There is also a growing body of work focused on the use of simulation for adaptive specification or validation of estimation strategies.[138, 139, 86, 140]

The planning, implementation, and adequate reporting of simulation studies to inform real analyses plays a critical role in the movement toward transparent and reproducible science. This is particularly true in regulatory affairs, where careful pre-specification of estimation choices is needed in randomized-controlled trials (e.g.,[29, 30, 31]), and is becoming increasingly used in observational data analyses (e.g.,[33, 34]). An area of particular emerging importance is the use of simulations to support valid inference for causal questions with complex (e.g., hierarchical) data structures. This is particularly true for per-protocol effects in longitudinal data, where we seek to "intervene" on treatments over time. In non-randomized data with long-term follow-up, we are often most interested in the effects of treatments over time. However, requiring sustained treatment over time under shifting usage patterns can reduce data support—i.e., the proportion of data that follows the regime of interest. Estimating these quantities well, particularly under observed data challenges, requires complex and flexible estimators. Failing to flexibly adjust for confounding or to utilize estimators that reflect the complexity of the data will result in biased inference. However, using such estimation methods can produce their own challenges. It is easiest for applied researchers to take the path of least resistance–relying blindly on estimator defaults, or choosing a favorite algorithm or library. However, relying on estimation defaults can risk invalid inference. Carefully designed simulations offer a tool to avoid these pitfalls, allowing for estimator fine-tuning while still pre-specifying the analysis.

At the Forum on the Integration of Observational and Randomized Data (FIORD) meeting held in Washington, D.C. in November 2022, statisticians from industry, academia, and regulatory agencies came together to discuss best practices for conducting applied simulations for causal analyses. The meeting members produced guidance that called for conducting outcome-blind simulations to guide estimator pre-specification.[38] There are some recommendations for how to use simulation to inform estimator specification (eg.[77, 141]), and some simulation examples for longitudinal data (e.g.,[142, 143]), but no tutorials on estimator selection in this setting

with specific guidance and code. This chapter aims to address this research gap by presenting a detailed case study for the causal effect of prolonged drug exposure using registry data—specifically, data simulated from the Danish national health registry. The data example includes complex time-dependent confounding, high dimensional confounders, and prolonged follow-up. As a result, estimators are needed that maintain their performance in the face of practical positivity violations.[65] The estimand of interest is a per-protocol estimand (i.e., sustained exposure over time). The goal of this chapter is to provide clear guidance, through a worked example, of how simulations can help inform practical analytic decisions and estimator specification. Our focus is on evaluating the statistical performance of candidate estimators using data generating processes (DGPs) in which the causal estimand is identified (i.e., the causal gap, or difference between the true value of the causal estimand and the true value of the statistical estimand, is zero by design of the simulation). Extensions to incorporate causal gap analysis are also possible (e.g., to estimate the amount of feasible/likely confounding), but are not the focus of current paper. There is a considerable literature on causal gap analysis, and other sensitivity analyses, covered elsewhere (e.g.,[38, 144]).

This chapter is organized as follows. In section 3.2, we give an overview of the motivation behind the case study used throughout the paper. In section 3.3, we give an overview of simulation design considerations; we walk through setting goals, generating the data, and calculating the benchmark truth and estimation comparison, including method comparison and evaluation metrics considered. In section 3.4, we conclude with some points of discussion and future directions of this work.

3.2 Case study

Treatment for type 2 diabetes is not only key to glycemic control, but can also have important downstream effects on other comorbidities like cardiovascular disease. Cardiovascular disease (CVD) accounts for 17.9 million deaths annually, and remains the leading cause of death globally.[145, 146] This risk is particularly high among those with other metabolic risk factors like diabetes; people with diabetes are two to four times as likely to die of CVD than the general population.[147]

Randomized trial data has shown that newer antidiabetic second-line therapies, such as Glucagon-like Peptide-1 Receptor Agonists (GLP-1RA), can lower the risk of major adverse cardiovascular events (MACE), including stroke, heart attack, and cardiovascular death.[16, 17, 18] However, most of these trials have been placebo-controlled studies of single drugs rather than comparisons of different treatment

options; direct comparisons of drugs can help inform clinician treatment decisions. Additionally, these trials demonstrate efficacy and not effectiveness of these therapies (i.e., the effect under highly controlled circumstances, which limits generalizability of results to real-world settings). The populations included in these studies were also not representative of the broader population which is at a lower risk for MACE.[118] This is particularly important as these newer therapies become more popular and widely available. In Denmark in particular, there has been a notable increase in use of newer second-line antidiabetic therapies in recent years.[106]

In this illustrative case study we used synthetic observed data designed to mimic the underlying registry data, as the "ground truth" for data privacy reasons and so that we could post reproducible code, available on Github.[148] The synthetic data used was created to mimic a population of over 100,000 insulin-naive diabetes patients in Denmark over the age of 50, who were initiating a second-line therapy for the first time between 2012 and 2021.

3.3 Designing an applied simulation study

Defining simulation goals

Any simulation for estimator specification requires a researcher to: define the goals of the simulation; simulate relevant data (specify a data generating process or DGP that mimics the real data in key respects); calculate the true value of the causal estimand; apply alternative estimators to generate estimates based on the simulated data and compare the performance of these estimators; implement the final estimator selection on the real data and report the results of the simulation and final estimation (Table 3.1).

The above simulation approach is executed only after a causal estimand, causal model, and observed data have been specified through a rigorous causal framework, such as the Causal Roadmap (Figure 3.1), and the causal estimand has been identified as a statstical estimand (parameter of the observed data distribution)—in other words, after the statistical estimation problem has been fully defined. The Causal Roadmap guides researchers in specifying: (1) a clear causal question, causal model and causal contrast; (2) the observed data available for estimation; (3) the identifiability assumptions needed in order to estimate the causal contrast using the observed data; (4) the statistical estimand and estimator, (5) estimation of the statistical quantity of interest, and (6) interpretation of results in the context of the assumptions and limitations. A rigorous framework like the Causal Roadmap is an essential component of conducting thoughtful and thorough causal research; simula-

tions can help improve the rigor of the research by guiding key estimation choices at the estimation phase of the Causal Roadmap.

Once the goals of the study are defined through the Roadmap, we can consider our various estimation strategies and what we would like to compare. For a given statistical estimand, there are many potential estimators from which to choose. This includes parametric g-computation, inverse probability weighting (including inverse probability of treatment and censoring weighting, but also double-robust methods like augmented inverse probability weighting), targeted minimum loss-based estimation or TMLE (including for example, cross-validated TMLE or collaborative TMLE). Within each estimator, there are other options such as weight truncation and nuisance parameter estimation algorithm.

We also must choose from a selection of potential methods for inference. For plug-in causal estimators, some popular options include empirical influence curve-based estimation, or a non-parametric bootstrap. There are also other options that can be specific to each estimator. For example, there is a robust variance option for TMLE.[80] For a discussion of options and automated data-adaptive estimation in the point-treatment, randomized trial case, see van Lacker et al. 2024.[149]

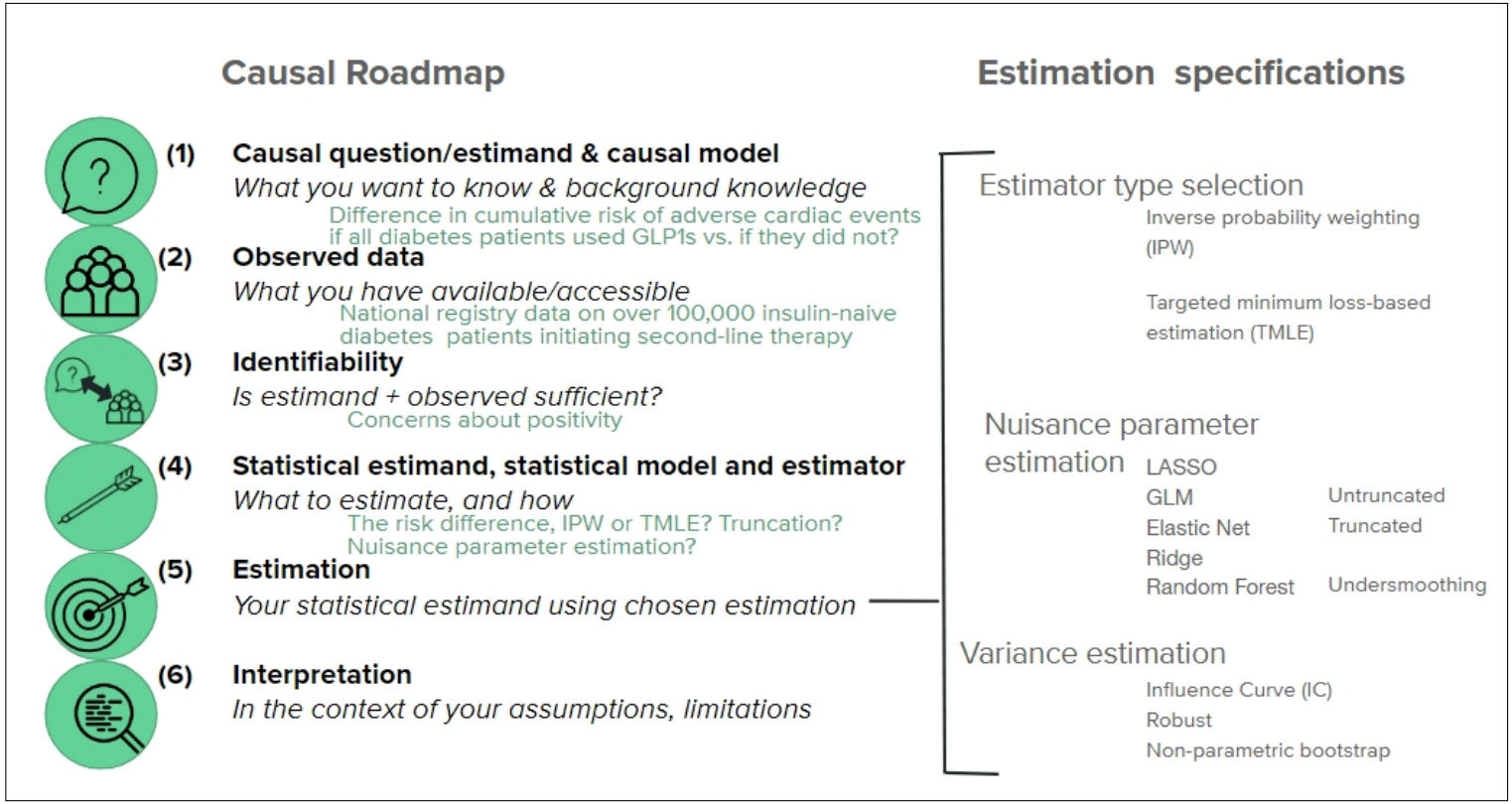


Figure 3.1: The Causal Roadmap and applied simulations for estimator pre-specification

Case study Roadmap and goals

In the case of our particular case study, we are interested in learning whether there is a difference in the two-year cumulative risk of adverse cardiac events if all diabetes patients over 50 years old used GLP1s vs. if they did not. The data available in the example is the Danish registry data: over 100,000 insulin-naive diabetes patients initiating second-line therapy. We are concerned about practical positivity violations due to rare GLP1 use relative to other medications, particularly over time. We are interested in estimating the risk difference in adverse cardiac events, comparing if all patients used GLP1 throughout follow up, compared to if they did not. Our statistical model is nonparametric. We are interested in comparing between different estimators, truncation levels for the treatment weights, and nuisance parameter estimation strategies. (Figure 3.1)

Table 3.1: Steps for planning and executing an applied simulation for estimation specification

Step	Considerations	Applied example
1. Define simulation goals	To guide estimator specification, the investigator should detail the specific estimators to be compared, as well as the metrics that will be used to choose between them.	To compare and choose between distinct: estimators (TMLE; IPW), Super Learner libraries, g truncation bounds, and confidence interval estimation.
2. Define a data generating process from which to simulate data	For a given longitudinal data structure, the process of specifying a data generating process can be broken down into approaches to simulating: 1. distribution of the baseline data (before any intervention) 2. conditional distribution of the treatment/intervention at each time point given the past (treatment mechanism) 3. conditional distribution of the non intervention nodes (covariates and outcomes) given the past at each time point. How you choose to estimate the above will imply some tradeoffs.	We defined the structural model of interest and a corresponding data generating process by estimating each of the structural equations using generalized linear models. (See Appendix A)

3. Find benchmark "truth" (i.e., the value of the causal estimand)	Can get value of the true causal estimand from Monte Carlo simulation (with very big N), intervening on variables of interest (e.g., exposure) and using the known DGP likelihood components.	Intervened on the models above, setting A to 1 (sustained GLP1 use) throughout follow up and sampling large N from the distribution, then setting A to all 0 (no GLP1 use) and sampling large N from the distribution. Took the expectation of the final outcome in each dataset and subtracting them to get the causal risk difference.
4. Estimate	Draw multiple "samples" from underlying DGP of observed sample size; implement proposed estimator on each (including all aspects of estimator); repeat for all estimation options of interest.	We simulated 500 datasets from the DGP described above and estimated the statistical estimand of interest.
5. Compare and select among candidate estimation strategies	Evaluate performance of candidate estimators by comparing using specified performance metrics. Employ two tiered approach: point estimation, then inference (such that point estimation choice is made irrespective of variance estimation).	Compared the oracle coverage and variance of the estimators, and then compared the coverage of the variance estimators. Chose the Super Learner library with machine learning (adaptive splines, gradient boosting, decision trees), with g truncation an robust variance estimation.

6. Implement selected estimator & report findings	Implement estimation and report simulation procedure transparently (eg. with framework like ADEMP[43]). Use: Markup language, reproducible and documented Github repository, pre-registration.	Git repository[148]

The goal of simulation in this case study is to provide insight into the performance of different estimation strategies. We seek to pre-select the point estimation strategy that has nominal coverage (and therefore bias low enough that it does not undermine inference), and minimal variance among various estimation strategies. Specifically, in this example, we are interested in comparing estimation strategies including the estimator (inverse probability weighting or IPW, and double-robust targeted maximum likelihood estimation or TMLE[150]), the nuisance parameter estimation (distinct Super Learner libraries), the weight truncation (truncation at different levels, no truncation), and the inferential procedure (empirical variance of the influence curve, non-parametric bootstrap, robust variance estimator).

Defining a data generating process

Our goal is to specify a DGP for the simulation that will result in estimator performance on the simulated data that reflects estimator performance on the real data. We want a DGP that will result in an unbiased estimate of the performance of the range of estimators we will consider. To do this, we want a simulation DGP that accurately reflects the aspects of the true DGP that will drive estimator performance (e.g., exposure-covariate relationships, in the case of propensity scores), but without having a DGP that is fit so closely to the DGP that it results in a biased estimate of performance due to overfitting the estimator to the data at hand without incorporating this in the estimation process itself (and corresponding inference).

In navigating between these, we can think about a spectrum of options in terms of the extent to which the observed data are used, assuming data have already been collected.

On one extreme of this spectrum is a data-generating process of fully investigator-specified (e.g., parametric) functions. Here, we use observed data only to align on basic metrics, such as: the marginal distribution of outcomes (time-varying for survival outcomes); the marginal distribution of treatment; the degree of practical positivity (conditional probability of following each treatment regime of interest given the measured past); and/or the number and type of baseline and time-varying covariates. An advantage of this approach is that it is blinded to the actual observed data, which protects the researcher against "p-hacking" or data dredging. However, this approach may result in a DGP that differs substantially from the data-generating process, making the results from the simulation more difficult to generalize to the real data.

On the other end of this spectrum of options is fitting the likelihood using the observed data and/or sampling empirically from the observed data (i.e., parametric bootstrap[65]). This approach mimics observed DGP and accommodates high di-

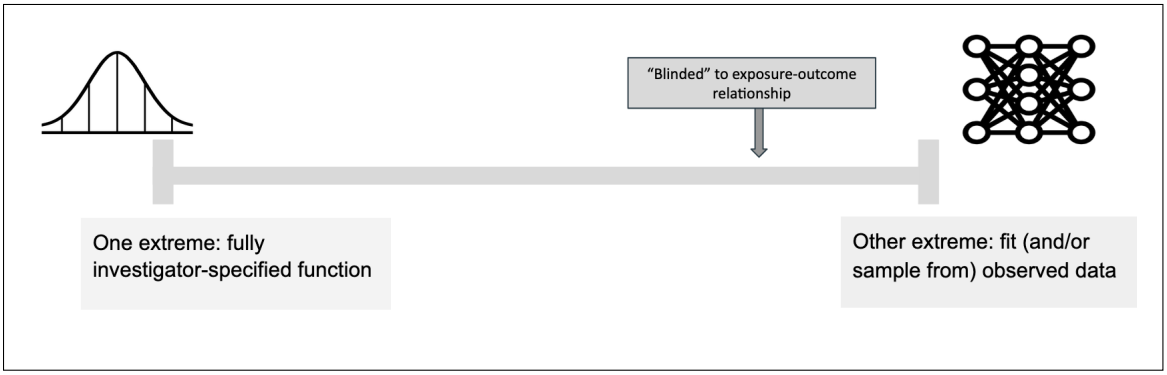


Figure 3.2: Spectrum of complexity in simulating the data generating process for an applied analysis

mensionality. However, if the outcome mechanism is fit, the researcher is no longer blinded to the exposure-outcome relationship.

Between these two extremes lies another option: to draw baseline covariates from the empirical distribution, fit the treatment and censoring mechanisms from observed data, use investigator-specified DGP for the outcome regression portions of likelihood (i.e., Q or non-interventional components). However, it is important to still maintain the marginal distribution of the outcome over time. One advantage of this approach is that the relationship between the components of the DGP that are most often related to practical positivity violations (i.e., the treatment/censoring mechanism and covariates) are maintained, while the components of the DGP that drive the effect size (i.e., the outcome) are investigator specified, improving the degree to which the simulation process is blinded.

Finally, another option may be to split the sample and reserve one portion of the data to fit all the components of the likelihood and the other to estimate. This can ensure both similarity to the true DGP and that the simulation is based on fully independent data (albeit drawn from the same underlying DGP). This may be possible when working with datasets with large sample sizes like that of a registry; however, even with these, it may not be feasible if the treatment regime of interest suffers from low data support.

An additional challenge that faces all of these approaches is the risk that the performance of some of the estimation choices may be overestimated by choosing the same functional form to simulate the data and to estimate the nuisance parameters. For example, simulating the data with a generalized linear model may mean that a generalized linear model will perform better in the comparisons than it otherwise

would in the real data.

An additional consideration worthy of mention is how to assess the "quality" of the data you are simulating. Metrics used for assessing the quality of the data being generated depend on the use case, and the extent to which one seeks to approximate the true underlying data distribution. There is a vast literature exploring the assessment of the fidelity of synthetic data (i.e., data that protects patient privacy while maintaining a similar data structure)[151], and there is no consensus in the research community on the best or most appropriate metrics. Univariate or bivariate statistics are helpful, but notably this does not illustrate the complex multivariate relationships between all variables at once. This is of particular importance in datasets with complex hierarchical structures. Popular methods for assessing quality include distance metrics (e.g., Hellinger, Wasserstein, euclidean) or correlation metrics (e.g., Pearson's). Of note, these metrics are only useful for proximal methods (i.e., methods that seek to replicate the exact data generating process of the underlying data). We provide a few illustrative examples in Appendix B for reference. Further discussion of these metrics and use cases can be found elsewhere (e.g.,[152, 153]), and we believe this area of research merits further study.

Case study data generation

In our case study we simulated the data unblinded, using the equations from the structural equation models provide in Appendix A. In brief, we fit a regression for each equation, modelling each node on its measured past. We then used these models to predict each post-baseline time point sequentially. We examined univariate statistics and bivariate correlations to compare the simulated data with the true data. For further detail on this process, see Appendix A.

Defining the "truth" in applied simulations

In order to compare estimation strategies, we must have a benchmark "truth" through which we can calculate metrics like bias and mean-squared error. To do this, we intervene on the intervention components of the data generating process defined in the previous section to reflect our desired contrasts of interest. Of note, in the case of a user-specified function that is set irrespective of the other variables, the true value should be zero.

Defining the "truth" in the case study

For simplicity, in our case study we simulate the truth by intervening on the structural equations in our model and drawing a large sample from the DGP. We set the intervention of interest, GLP1, to 1 over time. We used our trained models (per previous section), and starting with the first time point, set GLP1 use to 1 and simulated the data forward in time. We then did the same for the other contrast of interest, setting GLP1 to 0 over time. We did this for a large sample size (n=1000000), and then took the mean of the risk of MACE outcome at the end of the follow up (two years) for each contrast, and subtracted them to get the risk difference.

Estimation

Once our DGP and underlying truth are defined, we are ready to implement our estimators of interest as defined by our simulation goals. We recommend examining options in a factorial fashion where possible, so that all possible combinations are considered.

Case study estimation

In our case study, for point estimation, we compare: TMLE and IPW estimators; three Super Learner libraries (one containing only a simple logistic GLM, one containing a GLM and a GLM with interaction terms, and one containing those models plus multivariate adaptive regression splines[154, 155], gradient boosting[156, 157], and random forest[158]); truncation level of the estimated treatment probabilities (no truncation, more truncation i.e. truncation of probabilities <0.01, and less truncation i.e. truncation of probabilities <0.001). We implement all combinations of our estimation strategies on the generated data, using our benchmark truth.

Estimator comparison and selection

Here, we introduce a two-stage approach for estimation specification. First, we evaluate point estimation in a way that is agnostic to the variance estimation by using Monte Carlos simulation to estimate the standard error of the estimator (i.e., using sample variance of the point estimates from the simulation iterations) and corresponding oracle coverage (the 95% confidence interval coverage when using this estimated standard error of the estimator). We select the point estimator with the lowest standard error among those with approximately nominal oracle coverage. Then, using the selected point estimation strategy, we compare approaches to 95% confidence interval construction, aiming for an approach that achieves approximately

nominal coverage. A list of evaluation metrics and further discussion of options can be found in Nance et. al.[141].

Case study comparison and selection

We compare the estimation strategies of interest (as defined above) using oracle coverage. Among estimators that achieve nominal oracle coverage, we then (as defined above) select the estimator with the lowest variance. We display MSE and bias for completeness, however, if the estimator achieves nominal oracle coverage, then we can feel confident that the bias does not meaningfully affect inference. Below, we present results from 500 simulation iterations.

Estimator	SL.library	Truncation levels	Bias	MSE	Oracle coverage	SD of estimator
TMLE	Simple GLM	[0.01,1]	-0.007	0.002	0.958	0.046
		[0.001,1]	-0.008	0.002	0.966	0.048
		[0,1]	-0.007	0.002	0.964	0.049
	GLM with interactions	[0.01,1]	-0.008	0.002	0.954	0.045
		[0.001,1]	-0.008	0.002	0.966	0.047
		[0,1]	-0.008	0.002	0.964	0.047
	GLM + machine learning	[0.01,1]	-0.008	0.002	0.946	0.044
		[0.001,1]	-0.008	0.002	0.954	0.044
		[0,1]	-0.008	0.002	0.954	0.044
IPW	Simple GLM	[0.01,1]	0.026	0.003	0.916	0.047
		[0.001,1]	-0.004	0.003	0.962	0.053
		[0,1]	-0.005	0.003	0.958	0.056
	GLM with interactions	[0.01,1]	0.013	0.002	0.950	0.046
		[0.001,1]	-0.013	0.003	0.948	0.051
		[0,1]	-0.015	0.003	0.948	0.053
	GLM + machine learning	[0.01,1]	0.045	0.004	0.828	0.048
		[0.001,1]	0.041	0.004	0.854	0.048
		[0,1]	0.042	0.004	0.850	0.048

Table 3.2: Results comparing estimation strategies, averaging over 500 simulations iterations

We found that the TMLE estimator with the more complex Super Learner library that included machine learning algorithms was the most precise (i.e., had the lowest standard error (SE)) among those with close to 95% confidence interval coverage. There were no meaningful differences between different truncation levels; thus, we chose not to truncate.

We then compared strategies for inference. We implemented the above chosen specifications on the 500 simulated datasets using both an empirical influence curve (IC)[7] and robust standard error[80] variance estimation. We evaluated the variance estimator using 95% confidence interval coverage. We found robust variance estimation to have overly conservative coverage (98.2%), while the empirical influence curve variance estimation was anti-conservative (82.8%). This is consistent with previous findings,[80] and suggests for this particular case, neither of these approaches would be sufficient. Additional options that would be needed (and will be explored in future iterations) includes a non-parametric bootstrap, as well as approaches that integrate sample splitting(e.g.,[159, 149]).

Implementing selected estimator & report findings

Once we have defined our simulation goal, a DGP with ground truth, and compared the estimation strategies, we are able to implement the estimator on the real data. We estimate and report findings. Details of the simulation methods, along with reproducible code) should be included in addition to the real analytic results (e.g., as an appendix file).

Case study selected estimator & findings

In our applied example, we implemented the complex Super Learner stack (i.e. more learners and incorporating machine learning) with no g-bound truncation and robust variance estimation, and found a risk difference of -0.19 (95% confidence interval -0.27, -0.10). Of note, this estimate is similar to the effect magnitude found in the SELECT randomized trial of semaglutide, a GLP1, on MACE reduction.[18]

3.4 Discussion

Simulation decisions depend on the use case and needs of the researcher—there is no "one-size fits all" approach. Each dataset and estimand of interest have both advantages and estimation challenges. For this reason, it is important to tailor the analytic strategy to fit the estimation challenges in a particular example. Here, we outline

how this can be done. Through the case study of second-line diabetes medication and MACE, we highlight the use of simulation for pre-specification of analytic strategies. We define simulation goals, simulate the data, define our empirical "truth", choose estimation strategies to compare, estimate, select a method using pre-determined criteria, and implement the estimation strategy on the real data.

There are several key open questions that require further investigation and researcher consensus. Further research is needed on the best methods for simulating complex data structures (including longitudinal data but also other correlation structures, like multi-level data) for this kind of applied simulation. The creation and curation of software tools that will better automate the production of applied simulations for these kinds of complex estimands (and estimators) of interest is also needed, though creating such usable and flexible software is not trivial. However, without better tools to execute these simulations, they will remain infeasible for many applied researchers. We must address these barriers by (1) providing guidance on how to conduct these simulations, (2) providing more practical and usable tools for these simulations, and (3) improving the culture around applied simulation. Current research incentives prioritize significant results, at times over the research process. However, rushing to the finish line will continue to produce results that are haphazard and not reproducible. This will also be an important step in the standardization of applied simulation methods and reporting.

There are multitudinous other uses for simulation in applied data analysis that, however useful, have fallen outside the scope of this paper. An active area of research is sensitivity analysis to determine feasible bounds for unobserved confounding. Sensitivity analysis is an important step in applied epidemiology and in the Causal Roadmap.[38, 144] Furthermore, there is a need for greater guidance from regulatory bodies on what simulations to run and how to report them in analytic plans. Greater buy-in and guidance from these bodies would increase the transparency and reproducibility of applied research.

This paper builds on prior work to improve the rigor and reproducibility of real-data applications via simulation. We outline the simulation process and highlight several key areas of future research. We hope that this guide can serve as a reference for practitioners looking to implement these applied simulations in their own research.

3.5 Supplemental materials

Appendix A: Data simulation process

Structural Causal Model

We define the following exogenous nodes for $t = 1...k + 1$:

$$X = W, L(t), A(t)$$

where: W includes all time-constant covariates

$$A(t) = (\tilde{A}(t), C(t))$$

- $\tilde{A}(t)$ is GLP1-RA use at time t
- $C(t)$ is administrative censoring at time t

$$L(t) = (\tilde{L}(t), Y(t))$$

- $\tilde{L}(t)$ is the time-varying covariate status at time t
- $Y(t)$ is the outcome, adverse cardiac event, at time t

And exogenous error:

$$U = (U_A, U_L)$$

Structural equations are as follows:

$$L(t) = f_{L(t)}(\bar{A}(t-1), \bar{L}(t-1), U_{L(t)}), t = 1, ..., K+1$$

$$A(t) = f_{A(t)}(\bar{A}(t-1), \bar{L}(t-1), U_{A(t)}), t = 1, ..., K$$

Procedure

Here we provide a description for simulating data from the observed data distribution, given the above structural causal model. Specifically, we:

- factorize the likelihood into distinct component parts that need to be estimated;
- sample empirical baseline covariates;
- iteratively estimate each factor at each time point, from earliest to latest, using measured past;

- we use the models generated above to create predictions for the simulated data at each time point, using the resulting predictions as input for the subsequent ones.

Appendix B: Example metrics for data quality

Univariate statistics

You can examine the distribution of each covariate in the real and simulated data through a simple table, as seen below.

	"Real" data	Simulated data
Censor (t=1) (%)	1000 (20.0)	1637 (32.7)
GLP1 (t=1) (%)	734 (14.7)	877 (17.5)
MACE (t=1) (%)	1286 (25.7)	1408 (28.2)
Insulin (t=1) (%)	1299 (26.0)	1541 (30.8)
Censor (t = 2) (%)	2389 (47.8)	2482 (49.6)
GLP1 (t=2) (%)	747 (14.9)	638 (12.8)
MACE (t=2) (%)	1630 (32.6)	1905 (38.1)
Insulin (t=2) (%)	1115 (22.3)	1048 (21.0)
Censor (t = 3) (%)	3295 (65.9)	3350 (67.0)
GLP1 (t=3) (%)	628 (12.6)	569 (11.4)
MACE (t=3) (%)	1984 (39.7)	2370 (47.4)
Insulin (t=3) (%)	1064 (21.3)	2010 (40.2)
Censor (t = 4) (%)	3765 (75.3)	3908 (78.2)
GLP1 (t=4) (%)	532 (10.6)	464 (9.3)
MACE (t=4) (%)	2314 (46.3)	2378 (47.6)

In our example, there is substantially more censoring in the simulated data than the real data at the first time point. The other univariate distributions appear to be similar.

Bivariate statistics

Correlation plots are easy to generate and allow for a visual comparison of pairwise correlations between all the variables. An example is provided here:

In this example, the patterns of correlation structure in the two dataset are similar, but the bivariate relationships in the simulated data are slightly stronger.

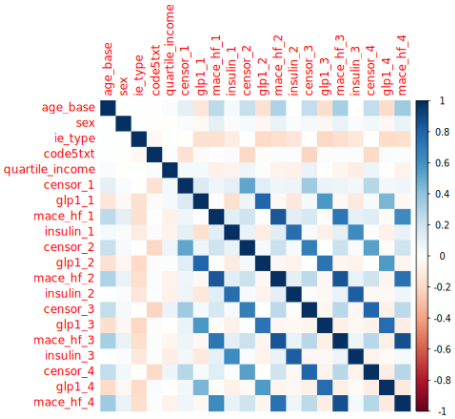


Figure 3.3: Bivariate correlation plot of "real" data

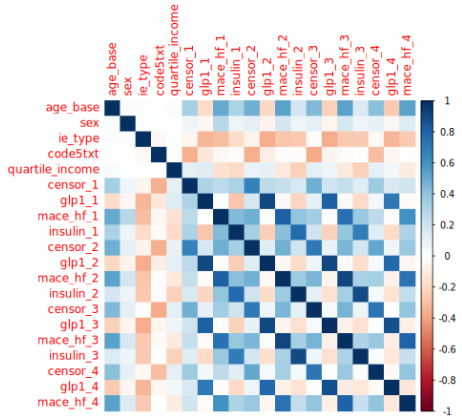


Figure 3.4: Bivariate correlation plot of simulated data

Conclusion

Through this work I provide guidance on how to use simulations to understand the performance of estimators under real-data conditions. I also include two key case studies of second-line diabetes treatments and their effects on downstream outcomes, namely dementia and adverse cardiac events. While the case studies are interesting in their own right, they illustrate the use of applied simulations, which I believe can aid pharmacoepidemiology researchers with ensuring adequate performance of their estimation strategy for their own real world examples.

In Chapter 1, I rallied practitioners to apply simulations in the specification of analytic strategies in statistical analysis plans. I walked through the Causal Roadmap and two relevant examples—one, a randomized trial with clustered data, and the other, an observational longitudinal medical record study. In the randomized trial, I illustrated how treatment-blind simulations can help compare strategies to estimate a two-stage TMLE with alternative approaches to account for missingness, adaptively adjust for covariates, and obtain inference. In the observational medical record study, I used outcome-blind simulations to illustrate choosing between nuisance parameter estimation and obtaining inference. In both examples, I show how these choices can be pre-specified and included in a statistical analysis plan to ensure research transparency.

In Chapter 2, I answered a relevant and timely substantive question of interest: whether there is a causal effect between sustained GLP1 use over time and dementia risk among diabetes patients in Denmark. I applied the Causal Roadmap, which gave a structured framework through which to ask the causal question, explicitly state needed assumptions, statistically estimate the quantity of interest, and interpret results in the context of our needed assumptions. In answering this question, I had available estimators with strong theoretical properties implemented in robust software. However, I was still faced with key threats to valid estimation. In order to ensure valid inference, I employed simulation as a tool through which to assess the performance of these estimators, and ultimately pre-select the estimator with superior performance. Substantive results showed a potentially protective relation-

ship between sustained GLP1 use and dementia diagnosis at five years. These results must be interpreted with caution, however; despite rigorous methodology, the study did not account for key confounders like smoking status. The study was also conducted in a relatively homogenous population. Future work may consider replicating this question in a more diverse cohort and with a more expansive set of covariates.

In Chapter 3, I provide more specific, pragmatic guidance for researchers on how to conduct and report an applied simulation study. I use a relevant case study of antidiabetic medications and adverse cardiac events. With the case study, I walk through the data simulation, truth definition, and implementation of distinct estimation choices. I detail the comparison and selection of an estimation strategy, and the reporting and implementation of this estimation strategy on the "real data".

While this work advances the field of applied pharmacoepidemiology and diabetes research, there are several noteworthy future directions. Firstly, there is a need for further research on the most appropriate way to approximate the data generating process. Secondly, there is a need for packages that automate this kind of "tuning" of the analytic choices for a given real-data example. Approaches such as "auto-TMLE" are being developed to dynamically evaluate estimators in simulations and implement the optimal estimator in a single step.[96, 97] However, this work is still forthcoming. It is essential that better tools that automate this process be created in order to ensure the adoption of simulation for real-data analyses. Without this, such work will remain outside of the scope of interest or ability of many applied researchers. Further work in this area must make simulations more accessible and easy to implement.

There are two key policy implications of this work that are worthy of mention. First and foremost, I hope that this work will influence guidelines for the use of simulations in applied analytic work, as well as encourage more research in this field. The second implication is relevant to applied pharmacoepidemiologic research on antidiabetic medications, specifically to the study of GLP1 and dementia. While the results on the relationship between GLP1 and dementia in Chapter 2 must be interpreted with caution, they suggest a potential protective relationship between GLP1 use and dementia diagnosis. I hope this work will not only add to the current RWE on this topic, but also encourage more research that addresses the limitations present in our study while utilizing a rigorous causal framework to conduct the research.

Bibliography

[1] Mònica Sabaté and Eva Montané. "Pharmacoepidemiology: an overview". In: *Journal of clinical medicine* 12.22 (2023). ISBN: 2077-0383 Publisher: Multidisciplinary Digital Publishing Institute, p. 7033.

[2] Office of the Commissioner. *Real-World Evidence*. en. Publisher: FDA. Apr. 2023. URL: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence (visited on 07/05/2024).

[3] *Use of real-world evidence in regulatory decision making – EMA publishes review of its studies — European Medicines Agency*. URL: https://www.ema.europa.eu/en/news/use-real-world-evidence-regulatory-decision-making-ema-publishes-review-its-studies (visited on 07/05/2024).

[4] *Methodological Guide - European Union*. en. URL: https://encepp.europa.eu/encepp-toolkit/methodological-guide_en (visited on 07/04/2024).

[5] Mark J. van der Laan and Daniel Rubin. "Targeted Maximum Likelihood Learning". en. In: *The International Journal of Biostatistics* 2.1 (Dec. 2006). Publisher: De Gruyter. ISSN: 1557-4679. DOI: 10.2202/1557-4679.1043. URL: https://www.degruyter.com/document/doi/10.2202/1557-4679.1043/html (visited on 11/09/2021).

[6] Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. en. Google-Books-ID: RGnSX5aCAgQC. Springer Science & Business Media, June 2011. ISBN: 978-1-4419-9782-1.

[7] Mark J. van der Laan and Sherri Rose. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. en. Google-Books-ID: vK-FTDwAAQBAJ. Springer, Mar. 2018. ISBN: 978-3-319-65304-4.

[8] Mei Xue et al. "Diabetes mellitus and risks of cognitive impairment and dementia: A systematic review and meta-analysis of 144 prospective studies". In: *Ageing Research Reviews* 55 (2019), p. 100944. ISSN: 1568-1637. DOI:

https://doi.org/10.1016/j.arr.2019.100944. URL: https://www.sciencedirect.com/science/article/pii/S1568163719300157.

[9] G. Cheng et al. "Diabetes as a risk factor for dementia and mild cognitive impairment: a meta-analysis of longitudinal studies". en. In: *Internal Medicine Journal* 42.5 (2012), pp. 484–491. ISSN: 1445-5994. DOI: 10.1111/j.1445-5994.2012.02758.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1445-5994.2012.02758.x (visited on 10/19/2022).

[10] Puja Ghosh et al. "Targeting redox imbalance in neurodegeneration: characterizing the role of GLP-1 receptor agonists". eng. In: *Theranostics* 13.14 (2023), pp. 4872–4884. ISSN: 1838-7640. DOI: 10.7150/thno.86831.

[11] Xiaonan Kang et al. "Exendin-4 ameliorates tau hyperphosphorylation and cognitive impairment in type 2 diabetes through acting on Wnt/-catenin/NeuroD1 pathway". eng. In: *Molecular Medicine (Cambridge, Mass.)* 29.1 (Sept. 2023), p. 118. ISSN: 1528-3658. DOI: 10.1186/s10020-023-00718-2.

[12] A. I. Duarte et al. "Crosstalk between diabetes and brain: glucagon-like peptide-1 mimetics as a promising therapy against neurodegeneration". eng. In: *Biochimica Et Biophysica Acta* 1832.4 (Apr. 2013), pp. 527–541. ISSN: 0006-3002. DOI: 10.1016/j.bbadis.2013.01.008.

[13] Olusola F. Onoviran et al. "Effects of glucagon-like peptide 1 receptor agonists on comorbidities in older patients with diabetes mellitus". en. In: *Therapeutic Advances in Chronic Disease* 10 (Jan. 2019). Publisher: SAGE Publications, p. 2040622319862691. ISSN: 2040-6223. DOI: 10.1177/2040622319862691. URL: https://doi.org/10.1177/2040622319862691 (visited on 11/09/2021).

[14] Francesco Vadini et al. "Liraglutide improves memory in obese patients with prediabetes or early type 2 diabetes: a randomized, controlled study". en. In: *International Journal of Obesity* 44.6 (June 2020). Number: 6 Publisher: Nature Publishing Group, pp. 1254–1263. ISSN: 1476-5497. DOI: 10.1038/s41366-020-0535-5. URL: https://www.nature.com/articles/s41366-020-0535-5 (visited on 10/20/2022).

[15] Tali Cukierman-Yaffe et al. "Effect of dulaglutide on cognitive impairment in type 2 diabetes: an exploratory analysis of the REWIND trial". English. In: *The Lancet Neurology* 19.7 (July 2020). Publisher: Elsevier, pp. 582–590. ISSN: 1474-4422, 1474-4465. DOI: 10.1016/S1474-4422(20)30173-3. URL: https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(20)30173-3/fulltext (visited on 11/14/2022).

[16] Bernard Zinman et al. "Liraglutide and Glycaemic Outcomes in the LEADER Trial". In: *Diabetes Therapy* 9.6 (Dec. 2018), pp. 2383–2392. ISSN: 1869-6953. DOI: 10.1007/s13300-018-0524-z. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6250637/ (visited on 03/24/2021).

[17] Hertzel C. Gerstein et al. "Dulaglutide and cardiovascular outcomes in type 2 diabetes (REWIND): a double-blind, randomised placebo-controlled trial". English. In: *The Lancet* 394.10193 (July 2019). Publisher: Elsevier, pp. 121–130. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(19)31149-3. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(19)31149-3/abstract (visited on 03/24/2021).

[18] Lincoff A. Michael et al. "Semaglutide and Cardiovascular Outcomes in Obesity without Diabetes". In: *New England Journal of Medicine* 389.24 (Dec. 2023). Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa2307563, pp. 2221–2232. DOI: 10.1056/NEJMoa2307563. URL: https://www.nejm.org/doi/full/10.1056/NEJMoa2307563 (visited on 07/09/2024).

[19] American Diabetes Association Professional Practice Committee. "10. Cardiovascular Disease and Risk Management: Standards of Care in Diabetes—2024". In: *Diabetes Care* 47.Supplement_1 (Dec. 2023), S179–S218. ISSN: 0149-5992. DOI: 10.2337/dc24-S010. URL: https://doi.org/10.2337/dc24-S010 (visited on 07/27/2024).

[20] *Recommendations to reduce cardiovascular risk in patients with diabetes published today*. URL: https://www.escardio.org/The-ESC/Press-Office/Press-releases/Recommendations-to-reduce-cardiovascular-risk-in-patients-with-diabetes-published-today,%20https://www.escardio.org/The-ESC/Press-Office/Press-releases/Recommendations-to-reduce-cardiovascular-risk-in-patients-with-diabetes-published-today (visited on 07/27/2024).

[21] Clive Ballard et al. "Liraglutide and semaglutide: Pooled post hoc analysis to evaluate risk of dementia in patients with type 2 diabetes". en. In: *Alzheimer's & Dementia* 16.S9 (2020), e042909. ISSN: 1552-5279. DOI: 10.1002/alz.042909. (Visited on 10/19/2022).

[22] Caroline Holm Nørgaard et al. "Treatment with glucagon-like peptide-1 receptor agonists and incidence of dementia: Data from pooled double-blind randomized controlled trials and nationwide disease and prescription registers". en. In: *Alzheimer's & Dementia: Translational Research & Clinical In-*

terventions 8.1 (2022), e12268. ISSN: 2352-8737. DOI: 10.1002/trc2.12268. (Visited on 09/19/2022).

[23] 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". en. In: *Mathematical Modelling* 7.9 (Jan. 1986), pp. 1393–1512. ISSN: 0270-0255. DOI: 10.1016/0270-0255(86)90088-6. URL: https://www.sciencedirect.com/science/article/pii/0270025586900886 (visited on 05/02/2023).

[24] Maya L. Petersen and Mark J. van der Laan. "Causal Models and Learning from Data". In: *Epidemiology (Cambridge, Mass.)* 25.3 (May 2014), pp. 418–426. ISSN: 1044-3983. DOI: 10.1097/EDE.0000000000000078. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4077670/ (visited on 07/25/2023).

[25] Miguel A. Hernán and James M. Robins. "Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available". In: *American Journal of Epidemiology* 183.8 (Apr. 2016), pp. 758–764. ISSN: 0002-9262. DOI: 10.1093/aje/kwv254. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4832051/ (visited on 09/22/2022).

[26] Matthew P. Fox et al. "The critical importance of asking good questions: The role of epidemiology doctoral training programs". en. In: *American Journal of Epidemiology* 189.4 (2020), pp. 261–264.

[27] Lauren E. Dang and Laura B. Balzer. "Start with the target trial protocol; then follow the Roadmap for causal inference". eng. In: *Epidemiology (Cambridge, Mass.)* (May 2023). ISSN: 1531-5487. DOI: 10.1097/EDE.0000000000001637.

[28] ICH Harmonised Tripartite Guideline. *Statistical Principles for Clinical Trials E9*. Feb. 1998.

[29] European Medicines Agency. *ICH E9 (R1) addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials*. Feb. 2020.

[30] U.S. Food and Drug Administration (FDA). *Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry*. Tech. rep. May 2023. URL: https://www.fda.gov/media/148910/download.

[31] National Institutes of Health (NIH). *Research Methods Resources*. URL: https://researchmethodsresources.nih.gov/ (visited on 07/11/2023).

[32] Lauren E. Dang et al. *A Causal Roadmap for Generating High-Quality Real-World Evidence*. arXiv:2305.06850 [stat]. May 2023. DOI: 10.48550/arXiv.2305.06850. URL: http://arxiv.org/abs/2305.06850 (visited on 10/20/2023).

[33] Marcus R. Munafò et al. "A manifesto for reproducible science". In: *Nature human behaviour* 1 (Jan. 2017), p. 0021. ISSN: 2397-3374. DOI: 10.1038/s41562-016-0021. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7610724/ (visited on 09/28/2023).

[34] Bart Hiemstra et al. "DEBATE-statistical analysis plans for observational studies". In: *BMC Medical Research Methodology* 19.1 (Dec. 2019), p. 233. ISSN: 1471-2288. DOI: 10.1186/s12874-019-0879-5. URL: https://doi.org/10.1186/s12874-019-0879-5 (visited on 09/04/2023).

[35] Iván Díaz et al. "Sensitivity analysis for causality in observational studies for regulatory science". en. In: *Journal of Clinical and Translational Science* 7.1 (Jan. 2023), e267. ISSN: 2059-8661. DOI: 10.1017/cts.2023.688. URL: https://www.cambridge.org/core/journals/journal-of-clinical-and-translational-science/article/sensitivity-analysis-for-causality-in-observational-studies-for-regulatory-science/75D253DC98BBA2C48624E35B4D83C34C# (visited on 04/05/2024).

[36] Maya B Mathur and Matthew P Fox. "Toward Open and Reproducible Epidemiology". In: *American Journal of Epidemiology* 192.4 (Apr. 2023), pp. 658–664. ISSN: 0002-9262. DOI: 10.1093/aje/kwad007. URL: https://doi.org/10.1093/aje/kwad007 (visited on 09/03/2023).

[37] Susan Gruber et al. "Developing a Targeted Learning-Based Statistical Analysis Plan". In: *Statistics in Biopharmaceutical Research* 15.3 (July 2023). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19466315.2022.2116104, pp. 468–475. ISSN: null. DOI: 10.1080/19466315.2022.2116104. URL: https://doi.org/10.1080/19466315.2022.2116104 (visited on 09/01/2023).

[38] Lauren E. Dang et al. *A Causal Roadmap for Hybrid Randomized and Real-World Data Designs: Case Study of Semaglutide and Cardiovascular Outcomes*. arXiv:2305.07647 [stat]. May 2023. DOI: 10.48550/arXiv.2305.07647. URL: http://arxiv.org/abs/2305.07647 (visited on 10/20/2023).

[39] Theodore Ruel et al. "A multilevel health system intervention for virological suppression in adolescents and young adults living with HIV in rural Kenya and Uganda (SEARCH-Youth): a cluster randomised trial". English. In: *The*

Lancet HIV 10.8 (Aug. 2023). Publisher: Elsevier, e518–e527. ISSN: 2352-3018. DOI: 10.1016/S2352-3018(23)00118-2. URL: https://www.thelancet.com/journals/lanhiv/article/PIIS2352-3018(23)00118-2/fulltext (visited on 11/30/2023).

[40] Heejung Bang and James M. Robins. "Doubly robust estimation in missing data and causal inference models". eng. In: *Biometrics* 61.4 (Dec. 2005), pp. 962–973. ISSN: 0006-341X. DOI: 10.1111/j.1541-0420.2005.00377.x.

[41] Laura B. Balzer et al. "Two-Stage TMLE to Reduce Bias and Improve Efficiency in Cluster Randomized Trials". In: *Biostatistics* kxab043 (2021).

[42] Maya Petersen et al. "Targeted Maximum Likelihood Estimation for Dynamic and Static Longitudinal Marginal Structural Working Models". eng. In: *Journal of Causal Inference* 2.2 (June 2014), pp. 147–185. ISSN: 2193-3677. DOI: 10.1515/jci-2013-0007.

[43] Tim P. Morris, Ian R. White, and Michael J. Crowther. "Using simulation studies to evaluate statistical methods". en. In: *Statistics in Medicine* 38.11 (2019). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8086, pp. 2074–2102. ISSN: 1097-0258. DOI: 10.1002/sim.8086. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8086 (visited on 10/21/2022).

[44] Joseph D. Y. Kang and Joseph L. Schafer. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data". In: *Statistical Science* 22.4 (2007). Publisher: Institute of Mathematical Statistics, pp. 523–539. ISSN: 0883-4237. URL: https://www.jstor.org/stable/27645858 (visited on 06/27/2023).

[45] J. S. Sekhon et al. "Propensity-Score-Based Estimators and C-TMLE". In: *Targeted Learning: Causal Inference for Observational and Experimental Data*. Ed. by M.J. van der Laan and S. Rose. New York, NY: Springer, 2011, pp. 343–364.

[46] A.I. Naimi, A.E. Mishler, and E.H. Kennedy. "Challenges in Obtaining Valid Causal Effect Estimates withMachine Learning Algorithms". In: *Am J Epidemiol* In press (2021).

[47] Laura B Balzer and Ted Westling. "Demystifying Statistical Inference When Using Machine Learning in Causal Research". In: *American Journal of Epidemiology* (July 2021), kwab200. ISSN: 0002-9262. DOI: 10.1093/aje/kwab200. URL: https://doi.org/10.1093/aje/kwab200 (visited on 07/11/2023).

[48] Vincent Dorie et al. "Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition". In: *Statistical Science* 34.1 (Feb. 2019). Publisher: Institute of Mathematical Statistics, pp. 43–68. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/18-STS667. URL: https://projecteuclid.org/journals/statistical-science/volume-34/issue-1/Automated-versus-Do-It-Yourself-Methods-for-Causal-Inference/10.1214/18-STS667.full (visited on 08/25/2023).

[49] Matthew P Fox et al. "Illustrating How to Simulate Data From Directed Acyclic Graphs to Understand Epidemiologic Concepts". In: *American Journal of Epidemiology* 191.7 (June 2022), pp. 1300–1306. ISSN: 0002-9262. DOI: 10.1093/aje/kwac041. URL: https://doi.org/10.1093/aje/kwac041 (visited on 07/11/2023).

[50] Cleo Anastassopoulou et al. "Data-based analysis, modelling and forecasting of the COVID-19 outbreak". In: *PLoS ONE* 15.3 (Mar. 2020), e0230405. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0230405. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7108749/ (visited on 08/25/2023).

[51] Keri N Althoff et al. "Life-Expectancy Disparities Among Adults With HIV in the United States and Canada: The Impact of a Reduction in Drug- and Alcohol-Related Deaths Using the Lives Saved Simulation Model". In: *American Journal of Epidemiology* 188.12 (Dec. 2019), pp. 2097–2109. ISSN: 0002-9262. DOI: 10.1093/aje/kwz232. URL: https://doi.org/10.1093/aje/kwz232 (visited on 10/26/2023).

[52] Roch A. Nianogo and Onyebuchi A. Arah. "Investigating the Role of Childhood Adiposity in the Development of Adult Type 2 Diabetes in a 64-year Follow-up Cohort: An Application of the Parametric G-formula Within an Agent-based Simulation Study". eng. In: *Epidemiology (Cambridge, Mass.)* 30 Suppl 2 (Nov. 2019), S101–S109. ISSN: 1531-5487. DOI: 10.1097/EDE.0000000000001062.

[53] Katsiaryna Bykov et al. "Comparison of self-controlled designs for evaluating outcomes of drug-drug interactions: simulation study". In: *Epidemiology (Cambridge, Mass.)* 30.6 (Nov. 2019), pp. 861–866. ISSN: 1044-3983. DOI: 10.1097/EDE.0000000000001087. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6768702/ (visited on 11/02/2023).

[54] Paul N. Zivich et al. "Transportability without positivity: a synthesis of statistical and simulation modeling". en-US. In: *Epidemiology* (), 10.1097/EDE.0000000000001677. ISSN: 1044-3983. DOI: 10.1097/EDE.0000000000001677. URL: https://journals.lww.com/epidem/abstract/

9900/transportability_without_positivity__a_synthesis.185.aspx (visited on 11/02/2023).

[55] Teresa J. Filshtein et al. "Proof of Concept Example for Use of Simulation to Allow Data Pooling Despite Privacy Restrictions". en-US. In: *Epidemiology* 32.5 (Sept. 2021), p. 638. ISSN: 1044-3983. DOI: 10.1097/EDE.0000000000001373. URL: https://journals.lww.com/epidem/abstract/2021/09000/proof_of_concept_example_for_use_of_simulation_to.4.aspx (visited on 11/02/2023).

[56] Shein-Chung Chow et al. *Sample Size Calculations in Clinical Research*. 3rd ed. New York: Chapman and Hall/CRC, Aug. 2017. ISBN: 978-1-315-18308-4. DOI: 10.1201/9781315183084.

[57] Office of the Commissioner. "Modeling & Simulation at FDA". en. In: *FDA* (Nov. 2022). Publisher: FDA. URL: https://www.fda.gov/science-research/about-science-research-fda/modeling-simulation-fda (visited on 07/11/2023).

[58] Laura B. Balzer et al. *Statistical Analysis Plan for SEARCH Phase I: Health Outcomes among Adults*. arXiv:1808.03231 [stat]. July 2018. DOI: 10.48550/arXiv.1808.03231. URL: http://arxiv.org/abs/1808.03231 (visited on 04/29/2024).

[59] Matthew P. Fox, Richard F. MacLehose, and Timothy L. Lash. *Applying quantitative bias analysis to epidemiologic data*. Springer, 2021. ISBN: 3-030-82672-4.

[60] Ruvani T. Jayaweera et al. "Accounting for Misclassification and Selection Bias in Estimating Effectiveness of Self-managed Medication Abortion". en-US. In: *Epidemiology* 34.1 (Jan. 2023), p. 140. ISSN: 1044-3983. DOI: 10.1097/EDE.0000000000001546. URL: https://journals.lww.com/epidem/Fulltext/2023/01000/Accounting_for_Misclassification_and_Selection.18.aspx (visited on 07/25/2023).

[61] Lina M. Montoya et al. "Efficient and robust approaches for analysis of sequential multiple assignment randomized trials: Illustration using the ADAPT-R trial". en. In: *Biometrics* n/a.n/a (). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13808. ISSN: 1541-0420. DOI: 10.1111/biom.13808. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13808 (visited on 07/19/2023).

[62] Brian D. Williamson et al. "An application of the Causal Roadmap in two safety monitoring case studies: Causal inference and outcome prediction using electronic health record data." eng. In: *Journal of clinical and translational science* 7.1 (2023). Place: England, e208. ISSN: 2059-8661. DOI: 10.1017/cts.2023.632.

[63] Nerissa Nance et al. *Applying the causal roadmap to longitudinal national Danish registry data: a case study of second-line diabetes medication and dementia*. arXiv:2310.03235 [stat]. Oct. 2023. DOI: 10.48550/arXiv.2310.03235. URL: http://arxiv.org/abs/2310.03235 (visited on 11/02/2023).

[64] Anne-Laure Boulesteix et al. "Introduction to statistical simulations in health research". en. In: *BMJ Open* 10.12 (Dec. 2020). Publisher: British Medical Journal Publishing Group Section: Epidemiology, e039921. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2020-039921. URL: https://bmjopen.bmj.com/content/10/12/e039921 (visited on 08/25/2023).

[65] Maya L. Petersen et al. "Diagnosing and responding to violations in the positivity assumption". eng. In: *Statistical Methods in Medical Research* 21.1 (Feb. 2012), pp. 31–54. ISSN: 1477-0334. DOI: 10.1177/0962280210386207.

[66] Kara E. Rudolph et al. "When Effects Cannot be Estimated: Redefining Estimands to Understand the Effects of Naloxone Access Laws". eng. In: *Epidemiology (Cambridge, Mass.)* 33.5 (Sept. 2022), pp. 689–698. ISSN: 1531-5487. DOI: 10.1097/EDE.0000000000001502.

[67] D. G. Horvitz and D. J. Thompson. "A generalization of sampling without replacement from a finite universe". In: *Journal of the American Statistical Association* 47 (1952), pp. 663–685. DOI: 10.2307/2280784.

[68] PAUL R. ROSENBAUM and DONALD B. RUBIN. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (Apr. 1983), pp. 41–55. ISSN: 0006-3444. DOI: 10.1093/biomet/70.1.41. URL: https://doi.org/10.1093/biomet/70.1.41 (visited on 09/04/2023).

[69] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed". In: *Journal of the American Statistical Association* 89.427 (1994). Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 846–866. ISSN: 0162-1459. DOI: 10.2307/2290910. URL: https://www.jstor.org/stable/2290910 (visited on 09/28/2023).

[70] Alejandra Benitez et al. "Defining and estimating effects in cluster randomized trials: A methods comparison". eng. In: *Statistics in Medicine* 42.19 (Aug. 2023), pp. 3443–3466. ISSN: 1097-0258. DOI: 10.1002/sim.9813.

[71] Joshua R Nugent et al. "Blurring cluster randomized trials and observational studies: Two-Stage TMLE for subsampling, missingness, and few independent units". In: *Biostatistics* (Aug. 2023), kxad015. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxad015. URL: https://doi.org/10.1093/biostatistics/kxad015 (visited on 04/29/2024).

[72] Brennan C. Kahan et al. "Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study". en. In: *Trials* 17.1 (Sept. 2016), p. 438. ISSN: 1745-6215. DOI: 10.1186/s13063-016-1571-2. URL: https://doi.org/10.1186/s13063-016-1571-2 (visited on 09/07/2023).

[73] A. E. Hubbard et al. "To GEE or Not to GEE Comparing Population Average and Mixed Models for Estimating the Associations Between Neighborhood Risk Factors and Health". In: *Epidemiology* 21.4 (2010), pp. 467–474.

[74] Garrett Fitzmaurice et al. "Estimation of the causal effects of time-varying exposures". In: *Longitudinal Data Analysis*. Chapman and Hall/CRC, 2008, pp. 567–614. ISBN: 0-429-14267-6.

[75] Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. "Super Learner". en. In: *Statistical Applications in Genetics and Molecular Biology* 6.1 (Sept. 2007). Publisher: De Gruyter. ISSN: 1544-6115. DOI: 10.2202/1544-6115.1309. URL: https://www.degruyter.com/document/doi/10.2202/1544-6115.1309/html (visited on 09/04/2023).

[76] Rachael V. Phillips et al. "Practical considerations for specifying a super learner". In: *International Journal of Epidemiology* 52.4 (Aug. 2023). arXiv:2204.06139 [stat], pp. 1276–1285. ISSN: 0300-5771, 1464-3685. DOI: 10.1093/ije/dyad023. URL: http://arxiv.org/abs/2204.06139 (visited on 09/07/2023).

[77] Susan Gruber et al. "Data-Adaptive Selection of the Propensity Score Truncation Level for Inverse-Probability-Weighted and Targeted Maximum Likelihood Estimators of Marginal Point Treatment Effects". eng. In: *American Journal of Epidemiology* 191.9 (Aug. 2022), pp. 1640–1651. ISSN: 1476-6256. DOI: 10.1093/aje/kwac087.

[78] Laura B. Balzer et al. "Adaptive pre-specification in randomized trials with and without pair-matching". eng. In: *Statistics in Medicine* 35.25 (Nov. 2016), pp. 4528–4545. ISSN: 1097-0258. DOI: 10.1002/sim.7023.

[79] D Benkeser et al. "Doubly robust nonparametric inference on the average treatment effect". In: *Biometrika* 104.4 (Dec. 2017), pp. 863–880. ISSN: 0006-3444. DOI: 10.1093/biomet/asx053. URL: https://doi.org/10.1093/biomet/asx053 (visited on 09/07/2023).

[80] Linh Tran et al. "Robust variance estimation and inference for causal effect estimation". en. In: *Journal of Causal Inference* 11.1 (Jan. 2023). Publisher: De Gruyter. ISSN: 2193-3685. DOI: 10.1515/jci-2021-0067. URL: https://www.degruyter.com/document/doi/10.1515/jci-2021-0067/html (visited on 08/31/2023).

[81] Laura B. Balzer et al. "Adaptive Selection of the Optimal Strategy to Improve Precision and Power in Randomized Trials". In: *Biometrics* In Press (May 2023).

[82] Samuel D. Lendle et al. "ltmle: An R Package Implementing Targeted Minimum Loss-Based Estimation for Longitudinal Data". en. In: *Journal of Statistical Software* 81 (Oct. 2017), pp. 1–21. ISSN: 1548-7660. DOI: 10.18637/jss.v081.i01. URL: https://doi.org/10.18637/jss.v081.i01 (visited on 01/23/2024).

[83] Reuven Y. Rubinstein. *Simulation and the Monte Carlo method*. eng. Third edition. Wiley series in probability and statistics. Hoboken, New Jersey: Wiley, 2017. ISBN: 978-1-118-63216-1.

[84] Nicholas Schreck et al. *Statistical Plasmode Simulations – Potentials, Challenges and Recommendations*. arXiv:2305.06028 [stat]. May 2023. DOI: 10.48550/arXiv.2305.06028. URL: http://arxiv.org/abs/2305.06028 (visited on 09/07/2023).

[85] David Benkeser et al. "Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes". en. In: *Biometrics* 77.4 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13377, pp. 1467–1481. ISSN: 1541-0420. DOI: 10.1111/biom.13377. URL: http://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13377 (visited on 06/30/2022).

[86] Richard Wyss et al. "Synthetic Negative Controls: Using Simulation to Screen Large-scale Propensity Score Analyses". eng. In: *Epidemiology (Cambridge, Mass.)* 33.4 (July 2022), pp. 541–550. ISSN: 1531-5487. DOI: 10.1097/EDE.0000000000001482.

[87] David Benkeser and Mark Van Der Laan. "The Highly Adaptive Lasso Estimator". In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Oct. 2016, pp. 689–696. DOI: 10.1109/DSAA.2016.93. URL: https://ieeexplore.ieee.org/abstract/document/7796956?casa_token=n3pF5FtqZgAAAAAA:0IqjP0Eg971ZI7074TvTEGl3WdBmBVzyuhL6ZwmHRFHwYPUK-8JyFMw_35vnQqY2Ta4pKrnXmUI (visited on 04/29/2024).

[88] A. W. van der Vaart. *Asymptotic Statistics*. en. Google-Books-ID: SYlmEAAAQBAJ. Cambridge University Press, June 2000. ISBN: 978-0-521-78450-4.

[89] Aaron Fisher. *Treatment Effect Bias from Sample Snooping: Blinding Outcomes is Neither Necessary nor Sufficient*. en. July 2020. URL: https://arxiv.org/abs/2007.02514v4 (visited on 12/14/2023).

[90] L. B. Balzer et al. "Statistical Analysis Plan for Primary and Selected Secondary Health Endpoints of the SEARCH-Youth Study". In: *arXiv* https://arxiv.org/abs/2211.02771 (2022).

[91] James M. Robins. "The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies". In: *Health service research methodology: a focus on AIDS* (1989). Publisher: US Public Health Service, pp. 113–159.

[92] Charles F. Manski. "Nonparametric bounds on treatment effects". In: *The American Economic Review* 80.2 (1990). ISBN: 0002-8282 Publisher: JSTOR, pp. 319–323.

[93] Elizabeth A. Stuart et al. "The use of propensity scores to assess the generalizability of results from randomized trials". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 174.2 (2011). ISBN: 0964-1998 Publisher: Oxford University Press, pp. 369–386.

[94] Elias Bareinboim and Judea Pearl. "A general algorithm for deciding transportability of experimental results". In: *Journal of causal Inference* 1.1 (2013). ISBN: 2193-3685 Publisher: De Gruyter, pp. 107–134.

[95] Sonja A. Swanson et al. "Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes". In: *Journal of the American Statistical Association* 113.522 (2018). ISBN: 0162-1459 Publisher: Taylor & Francis, pp. 933–947.

[96] Susan M. Shortreed and Erica EM Moodie. "Automated analyses: Because we can, does it mean we should?" In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 35.3 (2020). Publisher: NIH Public Access, p. 499.

[97] David Benkeser, Weixin Cai, and Mark J. van der Laan. "A Nonparametric Super-Efficient Estimator of the Average Treatment Effect". In: *Statistical Science* 35.3 (Aug. 2020). Publisher: Institute of Mathematical Statistics, pp. 484–495. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/19-STS735. URL: https://projecteuclid.org/journals/statistical-science/volume-35/issue-3/A-Nonparametric-Super-Efficient-Estimator-of-the-Average-Treatment-Effect/10.1214/19-STS735.full (visited on 02/27/2024).

[98] Miguel A. Hernán. "The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data". In: *American Journal of Public Health* 108.5 (May 2018). Publisher: American Public Health Association, pp. 616–619. ISSN: 0090-0036. DOI: 10.2105/AJPH.2018.304337. URL: https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2018.304337 (visited on 04/08/2023).

[99] Camila Olarte Parra et al. "Consistency of causal claims in observational studies: a review of papers published in a general medical journal". en. In: *BMJ Open* 11.5 (May 2021). Publisher: British Medical Journal Publishing Group Section: Research methods, e043339. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2020-043339. URL: https://bmjopen.bmj.com/content/11/5/e043339 (visited on 04/08/2023).

[100] amastyleinsider. *The Use of Cause-and-Effect Language in the JAMA Network Journals*. en-US. Sept. 2017. URL: https://amastyleinsider.com/2017/09/19/use-cause-effect-language-jama-network-journals/ (visited on 04/08/2023).

[101] Maya L. Petersen. "Applying a Causal Road Map in Settings with Time-dependent Confounding". In: *Epidemiology (Cambridge, Mass.)* 25.6 (Nov. 2014), pp. 898–901. ISSN: 1044-3983. DOI: 10.1097/EDE.0000000000000178.

URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4460577/ (visited on 10/31/2022).

[102] Morten Schmidt et al. "The Danish National Patient Registry: a review of content, data quality, and research potential". In: *Clinical Epidemiology* 7 (Nov. 2015), pp. 449–490. ISSN: 1179-1349. DOI: 10.2147/CLEP.S91125. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4655913/ (visited on 11/14/2022).

[103] Sandra K. Thygesen et al. "The predictive value of ICD-10 diagnostic coding used to assess Charlson comorbidity index conditions in the population-based Danish National Registry of Patients". eng. In: *BMC medical research methodology* 11 (May 2011), p. 83. ISSN: 1471-2288. DOI: 10.1186/1471-2288-11-83.

[104] *Dementia*. en. URL: https://www.who.int/news-room/fact-sheets/detail/dementia (visited on 12/24/2022).

[105] Geert Jan Biessels et al. "Risk of dementia in diabetes mellitus: a systematic review". eng. In: *The Lancet. Neurology* 5.1 (Jan. 2006), pp. 64–74. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(05)70284-2.

[106] Jakob S. Knudsen et al. "Changes in SGLT2i and GLP-1RA real-world initiator profiles following cardiovascular outcome trials: A Danish nationwide population-based study". In: *PLoS ONE* 15.3 (Mar. 2020), e0229621. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0229621. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7055862/ (visited on 06/02/2023).

[107] Lan Wen et al. "Parametric g-formula implementations for causal survival analyses". In: *Biometrics* 77.2 (June 2021), pp. 740–753. ISSN: 0006-341X. DOI: 10.1111/biom.13321. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9044498/ (visited on 04/21/2023).

[108] Camila Olarte Parra et al. "Trial emulation and survival analysis for disease incidence registers: A case study on the causal effect of pre-emptive kidney transplantation". eng. In: *Statistics in Medicine* 41.21 (Sept. 2022), pp. 4176–4199. ISSN: 1097-0258. DOI: 10.1002/sim.9503.

[109] Marios Rossides et al. "Infection risk in sarcoidosis patients treated with methotrexate compared to azathioprine: A retrospective 'target trial' emulated with Swedish real-world data". eng. In: *Respirology (Carlton, Vic.)* 26.5 (May 2021), pp. 452–460. ISSN: 1440-1843. DOI: 10.1111/resp.14001.

[110] Antoine Gavoille et al. "Investigating the Long-term Effect of Pregnancy on the Course of Multiple Sclerosis Using Causal Inference". In: *Neurology* 100.12 (Mar. 2023), e1296–e1308. ISSN: 0028-3878. DOI: 10.1212/WNL.0000000000206774. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10033171/ (visited on 10/11/2023).

[111] International Diabetes Federation. *Denmark diabetes report 2000 — 2045*. EN. URL: https://www.diabetesatlas.org/data/ (visited on 12/26/2022).

[112] Moien Abdul Basith Khan et al. "Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends". In: *Journal of Epidemiology and Global Health* 10.1 (Mar. 2020), pp. 107–111. ISSN: 2210-6006. DOI: 10.2991/jegh.k.191028.001. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7310804/ (visited on 11/09/2021).

[113] CDC. *What is Diabetes?* en-us. Mar. 2020. URL: https://www.cdc.gov/diabetes/basics/diabetes.html (visited on 11/09/2021).

[114] Hiroyuki Umegaki. "Type 2 diabetes as a risk factor for cognitive impairment: current insights". In: *Clinical Interventions in Aging* 9 (June 2014), pp. 1011–1019. ISSN: 1176-9092. DOI: 10.2147/CIA.S48926. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4085321/ (visited on 04/04/2023).

[115] Andrea Tumminia et al. "Type 2 Diabetes Mellitus and Alzheimer's Disease: Role of Insulin Signalling and Therapeutic Implications". eng. In: *International Journal of Molecular Sciences* 19.11 (Oct. 2018), E3306. ISSN: 1422-0067. DOI: 10.3390/ijms19113306.

[116] Paula L. McClean et al. "The diabetes drug liraglutide prevents degenerative processes in a mouse model of Alzheimer's disease". eng. In: *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 31.17 (Apr. 2011), pp. 6587–6594. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.0529-11.2011.

[117] Paula L. McClean et al. "Glucagon-like peptide-1 analogues enhance synaptic plasticity in the brain: a link between diabetes and Alzheimer's disease". eng. In: *European Journal of Pharmacology* 630.1-3 (Mar. 2010), pp. 158–162. ISSN: 1879-0712. DOI: 10.1016/j.ejphar.2009.12.023.

[118] Kristina S. Boye et al. "Generalizability of glucagon-like peptide-1 receptor agonist cardiovascular outcome trials to the overall type 2 diabetes population in the United States". eng. In: *Diabetes, Obesity & Metabolism* 21.6 (June 2019), pp. 1299–1304. ISSN: 1463-1326. DOI: 10.1111/dom.13649.

[119] Geert Jan Biessels et al. "Effect of Linagliptin on Cognitive Performance in Patients With Type 2 Diabetes and Cardiorenal Comorbidities: The CARMELINA Randomized Trial". In: *Diabetes Care* 42.10 (Aug. 2019), pp. 1930–1938. ISSN: 0149-5992. DOI: 10.2337/dc19-0783. URL: https://doi.org/10.2337/dc19-0783 (visited on 09/09/2022).

[120] Ida Kim Wium-Andersen et al. "Antidiabetic medication and risk of dementia in patients with type 2 diabetes: a nested case-control study". eng. In: *European Journal of Endocrinology* 181.5 (Nov. 2019), pp. 499–507. ISSN: 1479-683X. DOI: 10.1530/EJE-19-0259.

[121] Miguel A. Hernán. "The Hazards of Hazard Ratios". In: *Epidemiology (Cambridge, Mass.)* 21.1 (Jan. 2010), pp. 13–15. ISSN: 1044-3983. DOI: 10.1097/EDE.0b013e3181c1ea43. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3653612/ (visited on 04/08/2023).

[122] James M. Robins. "Causal Inference from Complex Longitudinal Data". en. In: *Latent Variable Modeling and Applications to Causality*. Ed. by Maia Berkane. Lecture Notes in Statistics. New York, NY: Springer, 1997, pp. 69–117. ISBN: 978-1-4612-1842-5. DOI: 10.1007/978-1-4612-1842-5_4.

[123] Klaus Kähler Holst and Esben Budtz-Jørgensen. "Linear latent variable models: the lava-package". en. In: *Computational Statistics* 28.4 (Aug. 2013), pp. 1385–1452. ISSN: 1613-9658. DOI: 10.1007/s00180-012-0344-y. URL: https://doi.org/10.1007/s00180-012-0344-y (visited on 10/21/2022).

[124] Wenjing Zheng and Mark van der Laan. "Asymptotic Theory for Cross-validated Targeted Maximum Likelihood Estimation". In: *U.C. Berkeley Division of Biostatistics Working Paper Series* (Nov. 2010). URL: https://biostats.bepress.com/ucbbiostat/paper273.

[125] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". en. In: *Journal of Statistical Software* 33 (Feb. 2010), pp. 1–22. ISSN: 1548-7660. DOI: 10.18637/jss.v033.i01. URL: https://doi.org/10.18637/jss.v033.i01 (visited on 08/06/2023).

[126] J. Kenneth Tay, Balasubramanian Narasimhan, and Trevor Hastie. "Elastic Net Regularization Paths for All Generalized Linear Models". en. In: *Journal of Statistical Software* 106 (Mar. 2023), pp. 1–31. ISSN: 1548-7660. DOI: 10.18637/jss.v106.i01. URL: https://doi.org/10.18637/jss.v106.i01 (visited on 08/06/2023).

[127] Marvin N. Wright and Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". en. In: *Journal of Statistical Software* 77 (Mar. 2017), pp. 1–17. ISSN: 1548-7660. DOI: 10.18637/jss.v077.i01. URL: https://doi.org/10.18637/jss.v077.i01 (visited on 08/06/2023).

[128] Susan Gruber and Mark van der Laan. "tmle: An R Package for Targeted Maximum Likelihood Estimation". en. In: *Journal of Statistical Software* 51 (Nov. 2012), pp. 1–35. ISSN: 1548-7660. DOI: 10.18637/jss.v051.i13. URL: https://doi.org/10.18637/jss.v051.i13 (visited on 11/14/2022).

[129] Andrew Mertens. *registry_simulations*. original-date: 2023-06-22T10:36:35Z. June 2023. URL: https://github.com/amertens/registry_simulations (visited on 09/08/2023).

[130] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: https://www.R-project.org/.

[131] Frederik Persson et al. "Different patterns of second-line treatment in type 2 diabetes after metformin monotherapy in Denmark, Finland, Norway and Sweden (D360 Nordic): A multinational observational study". en. In: *Endocrinology, Diabetes & Metabolism* 1.4 (2018). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/edm2.36, e00036. ISSN: 2398-9238. DOI: 10.1002/edm2.36. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/edm2.36 (visited on 02/15/2023).

[132] Michael Gejl et al. "In Alzheimer's Disease, 6-Month Treatment with GLP-1 Analog Prevents Decline of Brain Glucose Metabolism: Randomized, Placebo-Controlled, Double-Blind Clinical Trial". In: *Frontiers in Aging Neuroscience* 8 (2016). ISSN: 1663-4365. URL: https://www.frontiersin.org/articles/10.3389/fnagi.2016.00108 (visited on 09/09/2022).

[133] Benjamin F. Arnold et al. "Simulation methods to estimate design power: an overview for applied research". In: *BMC Medical Research Methodology* 11.1 (June 2011), p. 94. ISSN: 1471-2288. DOI: 10.1186/1471-2288-11-94. URL: https://doi.org/10.1186/1471-2288-11-94 (visited on 07/24/2023).

[134] Stefan D. Anker et al. "Traditional and new composite endpoints in heart failure clinical trials: facilitating comprehensive efficacy assessments and improving trial efficiency". eng. In: *European Journal of Heart Failure* 18.5 (May 2016), pp. 482–489. ISSN: 1879-0844. DOI: 10.1002/ejhf.516.

[135] Geraldine Rauch et al. "Opportunities and challenges of clinical trials in cardiology using composite primary endpoints". In: *World Journal of Cardiology* 7.1 (Jan. 2015), pp. 1–5. ISSN: 1949-8462. DOI: 10.4330/wjc.v7.i1.1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4306200/ (visited on 03/03/2024).

[136] Cristiana Mayer et al. "Simulation Practices for Adaptive Trial Designs in Drug and Device Development". In: *Statistics in Biopharmaceutical Research* 11.4 (Oct. 2019). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19466315.2018.1560359, pp. 325–335. ISSN: null. DOI: 10.1080/19466315.2018.1560359. URL: https://doi.org/10.1080/19466315.2018.1560359 (visited on 05/28/2024).

[137] Kristian Thorlund et al. "Highly Efficient Clinical Trials Simulator (HECT): Software application for planning and simulating platform adaptive trials". In: *Gates Open Research* 3 (Mar. 2019), p. 780. ISSN: 2572-4754. DOI: 10.12688/gatesopenres.12912.2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6556760/ (visited on 05/28/2024).

[138] Martin Huber, Michael Lechner, and Conny Wunsch. "The performance of estimators based on the propensity score". In: *Journal of Econometrics* 175.1 (July 2013), pp. 1–21. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2012.11.006. URL: https://www.sciencedirect.com/science/article/pii/S0304407613000390 (visited on 05/01/2024).

[139] Alejandro Schuler et al. *Synth-Validation: Selecting the Best Causal Inference Method for a Given Dataset*. arXiv:1711.00083 [stat]. Oct. 2017. DOI: 10.48550/arXiv.1711.00083. URL: http://arxiv.org/abs/1711.00083 (visited on 07/09/2024).

[140] Harsh Parikh et al. "Validating Causal Inference Methods". en. In: *Proceedings of the 39th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, June 2022, pp. 17346–17358. URL: https://proceedings.mlr.press/v162/parikh22a.html (visited on 02/18/2024).

[141] Nerissa Nance and Laura Balzer. *The Causal Roadmap and simulation studies to inform the Statistical Analysis Plan for real-data applications*. arXiv:2309.03952 [stat]. Nov. 2023. DOI: 10.48550/arXiv.2309.03952. URL: http://arxiv.org/abs/2309.03952 (visited on 11/09/2023).

[142] Jessica M. Franklin et al. "Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases". In: *Computational Statistics & Data Analysis* 72 (Apr. 2014), pp. 219–226. ISSN:

0167-9473. DOI: 10.1016/j.csda.2013.10.018. URL: https://www.sciencedirect.com/science/article/pii/S0167947313003721 (visited on 02/26/2024).

[143] Youssra Souli et al. "Longitudinal plasmode algorithms to evaluate statistical methods in realistic scenarios: an illustration applied to occupational epidemiology". en. In: *BMC Medical Research Methodology* 23.1 (Oct. 2023), p. 242. ISSN: 1471-2288. DOI: 10.1186/s12874-023-02062-9. URL: https://doi.org/10.1186/s12874-023-02062-9 (visited on 11/15/2023).

[144] Susan Gruber et al. "Evaluating and improving real-world evidence with Targeted Learning". In: *BMC Medical Research Methodology* 23.1 (Aug. 2023), p. 178. ISSN: 1471-2288. DOI: 10.1186/s12874-023-01998-2. URL: https://doi.org/10.1186/s12874-023-01998-2 (visited on 04/02/2024).

[145] World Health Organization (WHO). *Cardiovascular diseases*. en. URL: https://www.who.int/westernpacific/health-topics/cardiovascular-diseases (visited on 03/24/2021).

[146] World Health Organization (WHO). *The top 10 causes of death*. en. URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (visited on 03/24/2021).

[147] Raghavan Sridharan et al. "Diabetes Mellitus–Related All-Cause and Cardiovascular Mortality in a National Cohort of Adults". In: *Journal of the American Heart Association* 8.4 (Feb. 2019). Publisher: American Heart Association, e011295. DOI: 10.1161/JAHA.118.011295. URL: https://www.ahajournals.org/doi/10.1161/JAHA.118.011295 (visited on 03/22/2021).

[148] *nerissanance/fiord_sim_code*. en. URL: https://github.com/nerissanance/fiord_sim_code (visited on 07/09/2024).

[149] Kelly Van Lancker, Iván Díaz, and Stijn Vansteelandt. "Automated, efficient and model-free inference for randomized clinical trials via data-driven covariate adjustment". In: *arXiv preprint arXiv:2404.11150* (2024).

[150] Matthew J. Smith et al. "Application of targeted maximum likelihood estimation in public health and epidemiological studies: a systematic review". In: *Annals of Epidemiology* 86 (Oct. 2023), 34–48.e28. ISSN: 1047-2797. DOI: 10.1016/j.annepidem.2023.06.004. URL: https://www.sciencedirect.com/science/article/pii/S1047279723001151 (visited on 03/31/2024).

[151] Theodora Kokosi and Katie Harron. "Synthetic data in medical research". en. In: *BMJ Medicine* 1.1 (Sept. 2022). Publisher: BMJ Specialist Journals Section: Methods primer. ISSN: 2754-0413. DOI: 10.1136/bmjmed-2022-000167. URL: https://bmjmedicine.bmj.com/content/1/1/e000167 (visited on 04/30/2024).

[152] Khaled El Emam et al. "Utility metrics for evaluating synthetic health data generation methods: validation study". In: *JMIR medical informatics* 10.4 (2022). ISBN: 2291-9694 Publisher: JMIR Publications Toronto, Canada, e35734.

[153] Vibeke Binz Vallevik et al. "Can I trust my fake data–A comprehensive quality assessment framework for synthetic tabular data in healthcare". In: *International Journal of Medical Informatics* (2024). ISBN: 1386-5056 Publisher: Elsevier, p. 105413.

[154] Jerome H. Friedman. "Multivariate Adaptive Regression Splines". In: *The Annals of Statistics* 19.1 (Mar. 1991). Publisher: Institute of Mathematical Statistics, pp. 1–67. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176347963. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-19/issue-1/Multivariate-Adaptive-Regression-Splines/10.1214/aos/1176347963.full (visited on 07/07/2024).

[155] *Multivariate Adaptive Regression Splines*. URL: http://www.milbo.users.sonic.net/earth/ (visited on 07/07/2024).

[156] *XGBoost — Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. URL: https://dl.acm.org/doi/10.1145/2939672.2939785 (visited on 07/07/2024).

[157] *dmlc/xgboost*. original-date: 2014-02-06T17:28:03Z. July 2024. URL: https://github.com/dmlc/xgboost (visited on 07/07/2024).

[158] Leo Breiman. "Random Forests". en. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324 (visited on 07/07/2024).

[159] Wenjing Zheng and Mark J. van der Laan. "Cross-Validated Targeted Minimum-Loss-Based Estimation". en. In: *Targeted Learning: Causal Inference for Observational and Experimental Data*. Ed. by Mark J. van der Laan and Sherri Rose. Springer Series in Statistics. New York, NY: Springer, 2011, pp. 459–474. ISBN: 978-1-4419-9782-1. DOI: 10.1007/978-1-4419-9782-1_27. URL: https://doi.org/10.1007/978-1-4419-9782-1_27 (visited on 11/02/2023).