New approaches to factual and counterfactual prediction modeling

HARVARD UNIVERSITY

Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Committee on Higher Degrees in Population Health Sciences,
have examined a dissertation entitled

"New Approaches to Factual and Counterfactual Prediction"

presented by

CHRISTOPHER BOYER

candidate for the degree of Doctor of Philosophy
and hereby certify that it is worthy of acceptance.

Goodarz Danaei
Goodarz Danaei (Jan 11, 2023 10:20 EST)

Dr. Goodarz Danaei, Sc.D., Committee Chair, Harvard T.H. Chan School of Public Health

Andrew Beam
Andrew Beam (Jan 12, 2023 11:41 EST)

Dr. Andrew Beam, Ph.D., Harvard Medical School, Harvard T.H. Chan School of Public Health

James Robins (Jan 11, 2023 08:10 PST)

Dr. James M. Robins, M.D., Harvard T.H. Chan School of Public Health

Date: 06 January 2023

New approaches to factual and counterfactual prediction modeling

A dissertation presented
by

Christopher Boyer

to

The Department of Epidemiology

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Population Health Sciences

Harvard University
Cambridge, Massachusetts
January 2023

Dissertation Advisor:
Goodarz Danaei

Author:
Christopher Boyer

New approaches to factual and counterfactual prediction modeling

Abstract

Over the past half century, new methods for quantitative risk prediction and validation were formalized and the number of models, both statistical and algorithmic, increased exponentially. However, this literature has largely focused on descriptive predictions of the world as it is, what I term factual prediction, instead of the world as it would be if we intervened, or counterfactual prediction. In this dissertation, I argue that in many instances counterfactual predictions are desired, but targeting them requires new methods based on causal inference.

In Chapter 1, I take a method traditionally associated with causal inference, the g-formula, and repurpose it as a model for factual and counterfactual prediction. In doing so, I highlight the potential of the g-formula as unifying framework for prediction as well as the assumptions required. Through simulation and an applied data example in the Framingham Offspring Study, I show how the g-formula can estimate factual and counterfactual quantities and leverage multiple repeated measurements over time to produce predictions that update dynamically.

In Chapter 2, I consider an example of a common clinical prediction task, i.e. developing a model for risk-based treatment decisions, where the ideal target is counterfactual. Building on prior work, I clarify the single-arm target trial of interest and propose two estimation methods that allow for separation between the causal and prediction tasks. I apply these methods to predict the statin-naive risk of cardiovascular disease using an emulated trial based on the Multi-Ethnic Study of Atherosclerosis. I find that traditional methods lead to underallocation of treatment at common thresholds by 5 percentage points.

Finally, in Chapter 3, I tackle the theoretical question of how to train and validate models for counterfactual prediction when the relevant potential outcomes are not observed for all units. I discuss how to tailor a model for use in the same population under a counterfactual shift in treatment policy, how to assess its performance, and how to perform model and tuning parameter selection. I also provide identifiability results for measures of counterfactual performance for a potentially misspecified prediction model. I illustrate the methods using simulation and apply them to validate the performance of the statin-naive risk prediction model from Chapter 2.

Contents

List of Tables

List of Figures

Acknowledgments

The enormity of what I owe to those who helped me reach this point cannot possibly be conveyed in a few brief paragraphs. And yet here I will try. Never one for fiscal prudence, it seems inevitable that I shall fail to account for at least some of the many debts I incurred along the way. So first, to those I forgot to mention, mea culpa.

None of this would have happened without the incredible support of my advisor Goodarz Danaei. Goodarz took a chance on me when I was just a name on the waitlist. He has been a patient mentor and a great teacher. He provided invaluable comments on all my drafts, even the early inscrutable ones. Beyond work, he was supremely kind and always there to remind me of what was important when life events intervened.

One of the driving forces behind my decision to come to Harvard was to learn from the inimitable Jamie Robins. With the benefit of hindsight, I can now safely say he did not disappoint. Taking his class was a revelation and I was honored to continue on for another two years as a teaching fellow. To my dissertation committee, he brought precision and rigor and always asked the tough questions.

Andy Beam is one of the nicest people that I've met in academia and a faithful committee member. I am grateful that he gave me freedom to pursue my own intellectual passions even when they diverged from those I originally planned.

My best friend and wife, Jama, supported me, both financially and emotionally throughout my time in the PhD program. Every time that I thought that I couldn't go on any longer or got lost in the frenetic void of academia she pulled me back from the brink and helped remind me why I started on this path in the first place. She is my tireless champion. This accomplishment is as much hers as it is my own.

Thanks to my family who know that I have dreamed of getting my PhD since the fourth grade and have supported me throughout. To my mom who always believed in me, even when I didn't. To my aunt Carol, who was my academic role model growing up, and who graciously donated her regalia to me. And to my brothers Alex and Zane who always kept me humble.

I am eternally grateful to the many friends and colleagues in my PhD program who provided the intellectual and social spark. Eva Rumpler is a true friend, kept me sane during the pandemic, and wrangled my many ideas on our paper into something coherent. Wenze Tang, who took every single methods class with me and made it all worthwhile. Ilkania Paulino Chowdhury, who is smart and funny and amazingly kind. To my amazing PQE study group, Wenze, Ruibin, and Jessie, who made me a better methods thinker. The upper G-years, my forebearers, who modeled towering intellectual achievements and grace under duress, chief among them Louisa Smith, Gabe Schwartz, Keletso Makofane, and Aaron Sarvet. The students in PHS2000 and EPI207 who endured my lectures and taught me how to be an effective teacher.

I am thankful to my mentors and collaborators beyond Harvard, who shaped my professional and academic development. The incredible scholars on the *Becoming One* research team: Betsy Levy Paluck, Jeannie Annan, Jasper Cooper, Lori Heise, Tvisha Nevatia, and Jackline Namubiru. You provided a real-world proving ground for the many ideas and methods I accumulated during my graduate training and you are still the best team of investigators I have ever worked with. At the Mailman School of Public Health, James Phillips and Les Roberts helped me decide that I wanted to be an epidemiologist.

I learned a love for research and rigor from my great colleagues at Innovations for Poverty Action. Lindsey Shaughnessy and Kristi Post were great friends and the best bosses I have ever had. Pace Phillips was instrumental in helping me become an investigator and had a southern zen that always brightend my day. Biljana Bogicevic and Rosemarie Sandino were my first mentees and I couldn't be prouder of all they accomplished. Erica Chuang was a true friend and intellectual companion, but also let me sleep on her couch far too many times when I was commuting.

Finally, to my son Bruno, who was born during the last year of my PhD and provided the last jolt to get me over the line. I love you more than I ever thought possible.

To Jama and Bruno,
who endure my endless musings,
both *factual* and *counterfactual*.

Introduction

Predicting who is likely to get disease and who isn't is a foundational goal in clinical medicine and epidemiology. In the past, clinicians largely relied on a combination of biological theory, historical experience, and subject matter expertise to form opinions about a patient's prognosis. But starting with the landmark early investigations into the multivariable risk of cardiovascular disease using data from the Framingham Heart Study in the 1960s and 1970s, over the past half century, new methods for the quantitative prediction of disease were formalized [1, 2]. Models, either statistical or algorithmic, proliferated across a variety of specialties and, to varying degrees, were incorporated into clinical practice guidelines [3]. By 2016, more than 350 prediction models had been published in cardiology alone [4].

Over time a culture[1] of prediction emerged, both within statistics, but also in fields as diverse as computer science, epidemiology, and marketing [5, 6]. Often, but not always, adherents organized under the cross-disciplinary banner of "machine learning" [2]. This culture emphasized the differences between prediction and other tasks of statistics such as description or causal explanation. Unlike classical statistics, which often preceded from the premise that the data were generated according to a particular mechanism, proponents of the prediction culture were agnostic about the inherent validity of any model or algorithm. Rather they tended to focus on rigorous evaluation of model performance through techniques such as bootstrapping or cross-validation. Algorithms were chosen on the basis of their predictive accuracy rather than interpretability or relation to causality. Some even

[1] Those less well disposed may even call it a cult

championed the black box nature of modern prediction algorithms as they diverted attention away from persistent distractions such as the sign and value of coefficients. Marquee accomplishments of this new culture included the remarkable performance of many machine learning algorithms on previously challenging prediction tasks such as computer vision and natural language processing. In particular, so-called *deep learning* algorithms, neural networks with billions of parameters and trained on truly massive datasets, now dominate across a number of prediction domains and benchmarks. In clinical practice, deep learning has been less successful outside computer vision tasks, but more traditional methods have been successfully deployed and integrated into standard care.

With the increased visibility of the culture of prediction came more sophisticated critiques [5, 7, 8]. Prominent among these was the inability of prediction models to readily generalize beyond the settings in which they were trained. For example, in his response Leo Breiman's famous paper describing the "two cultures" of statistics, Sir David Cox wrote of the pure prediction approach

> The success of a theory is best judged from its ability to predict in new contexts. [...] If the prediction is localized to situations directly similar to those applying to the data there is then an interesting and challenging dilemma. Is it preferable to proceed with a directly empirical black-box approach, as favored by Professor Breiman, or is it better to try to take account of some underlying explanatory process? The answer must depend on the context but I certainly accept, although it goes somewhat against the grain to do so, that there are situations where a directly empirical approach is better. [...] However, much prediction is not like this. [5]

The degradation of model performance when applied to a new setting, sometimes called dataset shift [9], can be a result of any of a number of factors. For instance, in the clinical settings, it could be due to differences in distribution of baseline prognostic factors [10] or differences in policies or interventions [11]. While dataset shift can be at least partially addressed through re-calibration or re-fitting the model in the new setting, both are often infeasible either because collecting data in the new setting is difficult or expensive or because the dynamics of the applied setting are constantly shifting.

In his response, Professor Cox also mentions another critique of prediction: namely, that prediction is often performed without respect for the “underlying explanatory” (i.e. causal) proccess. This is important as often the true target of prediction is not simply predicting in conditions exactly as they occurred during training but to instead posit some change and predict the response. Indeed, Professor Cox continues

> Often the prediction is under quite different conditions from the data. what is the likely progress of the incidence of the epidemic of v-CJD in the United Kingdom, what would be the effect on annual incidence of cancer in the United States of reducing by 10% the medical use of X-rays, etc.? That is, it may be desired to predict the consequences of something only indirectly addressed by the data available for analysis. As we move toward such more ambitious tasks, prediction, always hazardous, without some understanding of underlying process and linking with other sources of information, becomes more and more tentative. [5]

Such predictions are *counterfactual* in the sense that they are concerned not with what occurred but what would occur under a hypothetical intervention. For instance, in clinical practice, a patient may not only wish to know their prognosis but what their prognosis *might be* if they were to initiate treatment or change their diet. However, accurately forming predictions involving counterfactuals requires a more systematic approach to the estimation of causal effects.

Interestingly, at the same time that a prediction culture was emerging within statistics, another band of iconoclasts from computer science, epidemiology, and economics were revolutionizing the statistician’s notion of cause and effect [12–14]. Throughout most of the 20th century, statisticians had steadfastly refused to entertain causal inference outside the narrow confines of a randomized trial. But what started as a rogue set of techniques for estimating causal effects from observational data in various fields, by the early 21st century had gradually developed into unified theory of causal inference. Using the formalism of causal graphs and potential outcomes, the new field of causal inference clarified the causal queries at the heart of many scholarly investigations and made precise the assumptions necessary to estimate causal effects from observational data.

This dissertation lies at the intersection of prediction and causal inference. As such, it is part of a growing literature which views prediction through a causal lens [11, 15–17]. A major contention of this dissertation is that many of the prominent issues in prediction may be more fruitfully addressed by introducing the causal formalisms pioneered by those in causal inference. Throughout I focus on instances where problems in prediction can be recast as problems involving estimation of counterfactual quantities. In this section, in what follows, I formalize the distinction between factual and counterfactual prediction. I then provide an overview of the succeeding chapters and how they relate to emerging issues in factual and counterfactual prediction.

Assume the data O available to the analyst has the following structure:

$$O = \{(X_i, Y_i)\}_{i=1}^{n}$$

where X_i is a p-dimensional vector of predictors taking values in $\mathcal{X} \in \mathbb{R}^p$, and $Y_i \in \{0, 1\}$ is an indicator of disease. A prediction model is an arbitrary map[2] of inputs X to probabilistic outputs about Y, i.e.

$$\Pr[Y = 1 \mid X] = \mu(X).$$

Adopting the terminology from Dickerman and Hernan [15], we call this target a *factual* prediction. Factual predictions typically answer "what is?" questions such as "what is the 10-year risk of developing cardiovascular disease for a patient with characteristics X". Crucially traditional methods for factual prediction typically assume that new observations $\{X_{new_i}\}_{i=1}^{m}$ are independent samples from the same population process. Therefore they are valid mostly in deployment settings that are approximately the same as those in which the model was trained. Agnostic validation of factual predictions produced by model $\mu(X)$ is possible by comparing observed Y against model predictions, generally in an independent test set, for instance by estimating the mean squared error

$$E[(Y - \mu(X))^2].$$

[2] While there are a number of ways in which the word *model* is used in statistics and aligned fields, we use this definition to be as expansive as possible as to the range of possible algorithms and estimators *model* can encompass.

Counterfactual predictions, by contrast, target quantites such as

$$\Pr[Y^g = 1 \mid X] = \mu^g(X)$$

where Y^g denotes the potential outcome in a world in which everyone received intervention g. They typically answer "what if?" or "what would?" questions such as "what would the 10-year risk of developing cardiovascular disease be for a patient with characteristics X if they were to quit smoking". They posit a world in which we intervened to change the conditions prevalent in the source population at the time the model was trained. As we will see, a principal problem is that Y^g is generally not observed for all individuals. Unbiased estimation will therefore require untestable assumptions as essentially we must borrow data from a subset of those for whom Y^g is observed.

In Chapter 1, I take a method traditionally associated with causal inference, the g-formula, and repurpose it as a model for factual and counterfactual prediction. In doing so, I highlight the potential of the g-formula as a unifying framework for prediction as well as the assumptions required. Through simulation and an applied data example in the Framingham Offspring Study, I show how the g-formula can estimate factual and counterfactual quantities and leverage multiple repeated measurements over time to produce predictions that update dynamically.

In Chapter 2, I consider an example of a common clinical prediction task, i.e. developing a model for risk-based treatment decisions, where the ideal target is counterfactual. More specifically, when using risk thresholds to determine who should be treated, the ideal risk is the so-called "treatment-naive" risk, i.e. the risk if treatment is never initiated. When, as is often the case, participants in the training data do initiate treatment, this risk is represented by the counterfactual quantity

$$\Pr[Y^{\bar{a}_k=0} = 1 \mid X]$$

or equivalently

$$\Pr[T^{\bar{a}_k=0} \leq t \mid X]$$

for failure time outcome T. Using the target trial framework [18, 19], I identify the hypothetical single-arm trial corresponding to these quantities and show how an analog can be constructed from observational data. I then propose two estimation methods based on inverse probability weighting and g-estimation of structural nested accelerated failure time models. Importantly, both methods allow for effective separation between the causal and prediction tasks, permitting one to use many of the black-box pure prediction algorithms mentioned above. I apply these methods to predict the statin-naive risk of cardiovascular disease using an emulated trial based on the Multi-Ethnic Study of Atherosclerosis. I find that traditional methods lead to underallocation of treatment at common thresholds by about 5 percentage points.

Finally, in Chapter 3, I tackle the theoretical question of how to train and validate models for counterfactual prediction when the relevant potential outcomes are not observed for all units. For instance, to evaluate the mean squarred error of a model targeting counterfactual outcome Y^g we need to estimate

$$E[(Y^g - \mu(X))^2].$$

However, under consistency, Y^g is only observed among those who follow regime g. In this chapter, I show that counterfactual performance metrics such as that above can sometimes be estimated directly from the test data. I provide identiability results for most commonly used metrics. Importantly, throughout I allow for the model $\mu(X)$ to be potentially misspecified, effectively separating the choice of prediction algorithm from the estimation of its performance. This opens up more agnostic evaluation of counterfactual prediction models that is in the spirit of the prediction "culture". I also show how to tailor a model for use in the same population under a counterfactual shift in treatment policy and how to perform model and tuning parameter selection. Lastly, I illustrate the methods using simulation and apply them to estimate the counterfactual performance of the statin-naive risk prediction model from Chapter 2.

Chapter 1

Factual and counterfactual risk prediction using the parametric g-formula[1]

Over the last several decades, clinical prediction models have proliferated and are now prominent in research and clinical care. However, most models have several limitations: (1) they use a single baseline examination cycle despite the fact that the underlying data sources often include longitudinal assessments of risk factors over time, (2) they make predictions under the "natural course" for the population from which they were derived without making explicit how risk factors and treatments changed over time, and (3) they are not appropriate for predictions under counterfactual interventions such as if an individual were to initiate treatment or adopt healthy lifestyle changes. In this study, we consider an alternative approach to risk prediction based on a modified version of the parametric g-formula which resolves these limitations and compare it to conventional modeling approaches. We argue that the g-formula provides a useful unifying framework for targeting a variety of prediction estimands of interest to clinicians and researchers. We provide guidance for estimation

[1] Co-authored with James Robins, Andrew Beam, and Goodarz Danaei

and modeling and discuss the assumptions required. We then compare a g-formula-based approach to prediction modeling to standard approaches via Monte Carlo simulation as well as in empirical datasets.

1.1 Introduction

Models predicting a patient's risk of developing a disease or other clinical outcome are common in the medical literature [4]. Most follow the same basic archetype: an analyst acquires data from a cohort of individuals at risk for an outcome, including a set of plausible predictors or "risk factors" measured at baseline as well as detailed follow up data on the occurence of the outcome. Then drawing on a vast and varied literature on regression modeling strategies [1, 20, 21], they fit a model for the probability of the outcome conditional on the values of the predictors at baseline. Once developed, the model is then applied prospectively to similar individuals to estimate their risk of developing the outcome with the intention that this information can be used to guide clinical decision-making — either as the basis for counseling individuals about treatment options or to stratify or prioritize individuals by risk. Alternatively, model predictions can be used at the population level to guide public health policy.

While in many cases a carefully-developed risk prediction model can provide valuable clinical information, there are also several limitations which may contribute to the reluctance to fully adopt them in clinical practice. First, there is often a mismatch between the descriptive information provided by the model and the types of questions that clinicians and their patients have regarding their treatment plan. For instance, while knowing one's 10-year risk of developing cardiovascular disease may be useful for actuarial purposes, it doesn't answer the natural follow up question of what one's risk might be if certain treatment or intervention strategies were pursued. Furthermore, the risk provided may be misleading if the population in which the model was trained includes individuals who initiate treatment over the follow-up period. Existing approaches are not designed or appropriate for answering these sorts of questions, many of which are counterfactual [15].

Second, by relying on a single baseline examination, models often neglect subsequent changes in treatment patterns, risk factors, or other determinants. They therefore represent risk under the "natural course" of the population on which the model was originally fit. For instance, models fit to data from the Framingham Heart Study typically assume the constancy of treatment patterns prevalent in the area during follow up decades later. Therefore, most models may be rendered invalid due to changes in coverage of treatments. This is further compounded by the fact that clinicians are often unaware of this natural course assumption, or of the treatment or risk factor patterns that predominate in the derivation cohort. They may als inappropriately apply the estimated risk reductions under treatment from randomized trials to this risk, rather than to the true risk under no treatment. Finally, this also present challenges for transporting models to new settings, even when some specific changes in the "natural course", such as shifts in treatment policy, are known in advance [16].

Finally, despite the fact that the underlying data sources often include repeated observations on individual over time, most existing modeling approaches do not account for changes within individuals over time. They also do not dynamically update when new information becomes available incorporating an individual's complete examination history. This ignores relevant information on risk factor trajectory which may lead to over or underestimation of risk.

In this paper, we argue that many of these limitations can be addressed by using a Robins' g-computation algorithm, or more specifically the parametric g-formula [22]. While traditionally used to estimate population-level causal effects, we show how it can be modified to target a range of prediction estimands, both factual and counterfactual. In this respect, our work contributes to a larger project which attempts to bridge the divide between causal inference and prediction methods [5, 6, 15, 23].

We begin by distinguishing between estimands of interest and highlight the requisite assumptions to estimate them using the g-formula. We then describe the modeling process for estimating the components of the g-formula. Using simulation, we explore the finite sample properties of the g-formula and compare it to a more conventional approach across a range of prediction scenarios. Finally we investigate the real-world performance of the g-formula in the prediction setting by means of an applied example using data from the Framingham Offspring Study.

1.2 Theory

1.2.1 Setup and Notation

Consider the common model development setting in which an analyst observes i.i.d. longitudinal samples $\{O_i\}_{i=1}^{n}$ from n participants following distribution $\mathbb{P}$. For each observation, let $O_i = \{(X_k, C_{k+1}, D_{k+1}, Y_{k+1})\}_{k=0}^{K}$ where X_k is a vector of, possibly time-varying, covariates measured at time k, C_k is an indicator of loss to follow-up by interval k, D_k as an indicator that a competing event has occurred by interval k, and Y_{k+1} is an indicator of whether the outcome of interest has occurred by time $k+1$ for $k = 0, \ldots, K$. By definition $C_0 \equiv 0$, $D_0 \equiv 0$, and $Y_0 \equiv 0$ as we restrict to those who are event-free at the start of follow up. We assume that covariates in X_k can be further categorized into modifiable variables for which we can imagine interventions (A_k), predictors of the outcome which do not act as confounders (P_k), as well as other important determinants of the outcome and the intervention variables which may act as confounders (L_k), i.e. $X_k = (L_k, P_k, A_k)$. Throughout, we use overbars to denote past history of a variable and underbars to denote future history, such that $\overline{X}_k = (X_0, \ldots, X_k)$ and $\underline{X}_k = (X_k, \ldots, X_K)$. Capital letters represent random variables and their lower case equivalents are realizations.

Our main interest is in the case that Y_k is a survival outcome, i.e. $\underline{Y}_{k+1} = (1, \ldots, 1)$ if $Y_k = 1$, as this is the most common prediction outcome in epidemiology. However, for simplicity of presentation, we begin by assuming no loss to follow up ($\overline{C}_K = 0$) and no competing events ($\overline{D}_K = 0$), in which case we can treat Y_{K+1} as an end of follow up outcome. An example directed acyclic graph for a two time point process under these simplyifying assumptions is shown in Figure 3.1.

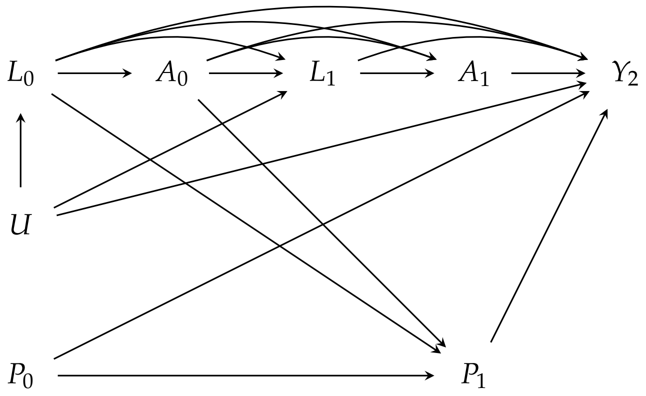


Figure 1.1: *Example two time point directed acyclic graph for prediction.*

1.2.2 Factual prediction estimands

Classically, the goal of a risk prediction model is to estimate the probability that an individual develops a clinical outcome at future time $K + 1$ conditional on their baseline risk factor levels. Generally, we assume that this probability is well approximated by the proportion of people with the same risk factor levels who develop the outcome in the population from which the individual is derived. Using the notation above, this corresponds to the estimand

$$\Pr(Y_{K+1} = 1 \mid X_0 = x_0).$$

In the time-dependent setting, where repeated measurements of an individual's risk factor levels are collected over time, a further goal might be to estimate an "updated" sequence of probabilities that the individual develops a clinical outcome conditional on their risk factor history up until the present time k (where $k < K$). Under similar logic, this can be represented by the population estimand

$$\Pr(Y_{K+1} = 1 \mid \overline{X}_k = \overline{x}_k).$$

Notice, in both instances the estimands are written in terms of the observed data and therefore relate strictly to facts about the population; facts which we can hope to learn to an arbitrary level of precision by drawing samples and applying the standard tools for statistical estimation and inference. As long as the relationship between covariates and the outcome are stable across time and the prediction is applied to member of the original population, we require no additional assumptions.

1.2.3 Counterfactual prediction estimands

By contrast, there are many clinical questions in which the relevant prediction is not the extant risk among similar individuals, but rather a "hypothetical" assessment of what their risk would be under different intervention strategies. For instance, in addition to the 10-year risk of developing coronary heart disease among those with risk factor levels similar to our own, we may also be interested in knowing the risk if we quit smoking, started taking statins, or committed to exercising more regularly. Again, assuming this individual prediction is well approximated by the proportion of people with same or similar risk factor levels who would develop the outcome in a population in which everyone followed the strategy of interest, this suggests targeting counterfactual estimands of the form

$$\Pr(Y_{K+1}^{\underline{a}_k} = 1 \mid \overline{X}_k = \overline{x}_k)$$

where $Y_{K+1}^{\underline{a}_k}$ denotes the potential outcome that would be observed at time $K+1$ if, possibly contrary to the fact, the intervention sequence $\underline{A}_k = \underline{a}_k$ were followed at times between k and K among those with same history up to to time k.

Note that, in addition to the fixed intervention sequences such as those above, we can also imagine a more general class of counterfactual regimes in which interventions may be probabilistically assigned or in which intervention assignment can be a function of previous covariate history. For example, we might be interested in predicting cardiovascular risk under a regime in which we initiated anti-hypertensive medication only once our blood pressure crossed a particular threshold. Following the definitions in Hernán & Robins [12], we define a regime which depends on previous covariate values to be *dynamic* and one which does not to be *static* as well as defining a regime in which an intervention is deterministically assigned to be *deterministic* and a regime in which it is probabilitically assigned to be *random*. For a particular regime g where $g \in \{g_k(\bar{l}_k, \bar{a}_{k-1}), k = k, \ldots, K\}$ we can write our target estimand as

$$\Pr(Y^g_{K+1} = 1 \mid \overline{X}_k = \overline{x}_k).$$

Unlike the factual estimands of the previous sections, estimation and inference of these estimands requires additional untestable assumptions, namely

1. *Exchangeability:* $Y^g_{K+1} \perp\!\!\!\perp A_k \mid \overline{L}_k, \overline{P}_k, \overline{A}_{k-1}$

2. *Consistency:* $Y_{K+1} = Y^g_{K+1}$, $\overline{L}_k = \overline{L}^g_k$, and $\overline{P}_k = \overline{P}^g_k$ if $\overline{A}_k = \overline{a}^g_k$

3. *Positivity:* $\Pr(A_k = a_k \mid \overline{L}_k = \bar{l}_k, \overline{P}_k = \overline{p}_k, \overline{A}_{k-1} = \overline{a}_{k-1}) > 0$

for all k from k to K. These assumptions are discussed in more detail elsewhere [12, 22]. Briefly, exchangeability implies that treatment assignment at time k is as good as randomized within strata of covariate histories $\overline{L}_k$, $\overline{P}_k$, and $\overline{A}_{k-1}$. Consistency implies that observed outcomes reflect potential outcomes under the observed treatment sequence and could be violated if there were interference between units such that outcomes may be affected by another unit's treatment assignment, or alternatively if there are multiple hidden versions of treatment. Finally, positivity means that there must be a positive nonzero probability of observing all possible treatment conditions a_k at time k conditional on covariate history for all values of k; it would be violated if, for instance, within some stratum of $\overline{L}_k$, $\overline{P}_k$, and $\overline{A}_{k-1}$ it is not possible to receive treatment.

1.2.4 The g-formula, identification, and the natural course

In a now classic result, Robins [22] showed that —under the assumptions of exchangeability, consistency and positivity— counterfactual estimands for time-varying interventions are consistently estimated by the g-formula, i.e.

$$\Pr(Y^{\overline{a}_k}_{K+1} = 1) = \sum_{\bar{l}_k} \Pr(Y_{k+1} = 1 \mid \overline{L}_k = \bar{l}_k, \overline{A}_k = \overline{a}_k) \times \prod_{j=0}^{K} f(l_j \mid \bar{l}_{j-1}, \overline{a}_{j-1}) \tag{1.1}$$

where, in a slight abuse of notation, here we use $f(\cdot)$ to represent the (conditional) probability density function for covariate L_k given it's past. This results can be generalized to random and dynamic regimes, i.e.

$$\begin{aligned}\Pr(Y^g_{K+1} = 1) = \sum_{\bar{l}_k}\sum_{\bar{a}_k} & \Pr(Y_{k+1} = 1 \mid \overline{L}_k = \bar{l}_k, \overline{A}_k = \bar{a}^g_k) \times \\ & \prod_{j=0}^{K}\left\{ f(l_j \mid \bar{l}_{j-1}, \bar{a}^g_{j-1}) \times f^g(a_j \mid \bar{l}_j, \bar{a}_{j-1}) \right\}\end{aligned} \tag{1.2}$$

where $f^g(a_j \mid \bar{l}_j, \bar{a}_{j-1})$ is the intervention density specified by the random regime (i.e. the probability rule for how treatment is assigned).

This generalized g-formula expression is written in terms of the marginal counterfactual distribution. This makes sense as in most settings in which the g-formula has been applied the targets are population-level counterfactual effects. However, in the context of time-dependent risk prediction the target estimand is conditional on past covariate history. Therefore, we modify it slightly to produce what we term here the *conditional g-formula*, i.e.

$$\begin{aligned}&\Pr(Y^g_{K+1} = 1 \mid \overline{L}_k = \bar{l}_k, \overline{P}_k = \bar{p}_k, \overline{A}_k = \bar{a}_k) = \\ &\quad \sum_{\underline{l}_k}\sum_{\underline{p}_k}\sum_{\underline{a}_k}\Pr(Y_{k+1} = 1 \mid \overline{L}_k = \bar{l}_k, \overline{P}_k = \bar{p}_k, \overline{A}_k = \bar{a}_k, \underline{L}_k = \underline{l}_k, \underline{P}_k = \underline{p}_k, \underline{A}_k = \underline{a}^g_k) \times \\ &\quad\quad \prod_{j=k}^{K}\left\{ f(l_j \mid \bar{l}_{j-1}, \bar{p}_{j-1}, \bar{a}^g_{j-1}) \times f(p_j \mid \bar{l}_j, \bar{p}_{j-1}, \bar{a}^g_{j-1}) \times f^g(a_j \mid \bar{l}_j, \bar{p}_j, \bar{a}_{j-1}) \right\}\end{aligned} \tag{1.3}$$

where sums (or integrals in the case of continuous covariates) are now taken over just the period between k and K, where k is the examination cycle history of interest and K is selected based on the follow-up for the relevant cumulative risk. We also further condition on predictor history $\overline{P}_k$ and sum out future values. In section C.1 we show that this conditional g-formula recovers the conditional counterfactual prediction estimand under the modified exchangeability, consistency, and positivity assumptions in section 1.2.3, motivating it's use as a risk prediction model for counterfactual estimands.

Perhaps less appreciated, is that the g-formula can also recover factual risk prediction estimands. Indeed, it can be shown that the risk under no intervention is equivalent to the g-formula under a random dynamic regime in which the intervention density is equivalent to the observed probability of treatment, i.e. $f^{g}(a_j \mid \bar{l}_j, \overline{p}_j, \bar{a}_{j-1}) = f^{obs}(a_j \mid \bar{l}_j, \overline{p}_j, \bar{a}_{j-1}))$, where the latter is often termed the *natural course*. We can write the g-formula under the natural course as

$$\begin{aligned}
&\Pr(Y^{g}_{K+1} = 1 \mid \overline{L}_k = \bar{l}_k, \overline{P}_k = \overline{p}_k, \overline{A}_k = \bar{a}_k) = \\
&\quad \sum_{\underline{l}_k}\sum_{\underline{p}_k}\sum_{\underline{a}_k} \Pr(Y_{k+1} = 1 \mid \overline{L}_k = \bar{l}_k, \overline{P}_k = \overline{p}_k, \overline{A}_k = \bar{a}_k, \underline{L}_k = \underline{l}_k, \underline{P}_k = \underline{p}_k, \underline{A}_k = \underline{a}^{g}_k) \times \\
&\quad\quad \prod_{j=k}^{K} \left\{ f(l_j \mid \bar{l}_{j-1}, \overline{p}_{j-1}, \bar{a}^{g}_{j-1}) \times f(p_j \mid \bar{l}_j, \overline{p}_{j-1}, \bar{a}^{g}_{j-1}) \times f^{obs}(a_j \mid \bar{l}_j, \overline{p}_j, \bar{a}_{j-1}) \right\}
\end{aligned} \tag{1.4}$$

or, recognizing that we defined $X_k = (L_k, P_k, A_k)$ previously, more simply

$$\Pr(Y_{K+1} = 1 \mid \overline{X}_k = \bar{x}_k) = \sum_{\underline{x}_k} \Pr(Y_{k+1} = 1 \mid \overline{X}_k = \bar{x}_k, \underline{X}_k = \underline{x}_k) \times \prod_{j=k}^{K} f(x_j \mid \bar{x}_{j-1}). \tag{1.5}$$

The connection between random dynamic regimes and traditional factual prediction estimands through the concept of the natural course is useful for thinking about the role of modifiable variables (e.g. treatments) in factual prediction settings. For instance, a common problem in clinical risk prediction modeling is how to interpret estimates when some participants in a training data set receive treatment during follow up. One way to think about these risk estimates is that they represent the risk under the natural course of treatment, i.e. under a hypothetical dynamic regime in which treatment is assigned probabilistically according to observed treatment and covariate history, and therefore should be applicable in settings in which the "natural course" mechanism, $f^{obs}(a_j \mid \bar{l}_j, \overline{p}_j, \bar{a}_{j-1})$, is the same or similar.

In the prediction literature, dynamic prediction rules are said to be *consistent*[2] when a prediction at future time K conditional on covariates through time k can be obtained by summing or integrating over the conditional probability distribution of the longitudinal outcome in the interval between k and K [24]. This is a desirable property as it implies a degree of coherence among longitudinal models. Although this has generally been applied to joint modeling of a single biomarker, a consequence of the result above is that factual

[2] Meant here in a different sense than it is used in section 1.2.4.

predictions using the g-formula also satisfy this condition.

1.2.5 Censoring and competing events

In the real world, participants are lost to follow up or their covariate data may be missing. Additionally, competing events may preclude a subset of participants from experiencing the outcome of interest. Both affect the quality of predictions [25, 26]. Additionally, in both cases a nominally factual prediction estimand may be re-conceived as a counterfactual prediction estimand [27]. However, to do so invariably involves invoking additional untestable assumptions about the nature of the data generation process.

Following the framework of Young *et al.* [27] we conceive of two types of estimands involving competing events, the risk with and without elimination of competing events. The former is a counterfactual estimand which imagines a world in which no one ever had a competing event, for instance if there were an intervention which could eliminate the possibility of the event occurring. While this can make sense as a target of inference in certain circumstances, when the competing event is death, as is commonly the case in epidemiological studies, this estimand may be less plausible. In the statistics literature, the former (i.e. the risk under elimination) is also known as the marginal cumulative incidence or the net risk, while the latter is often referred to as the cause-specific cumulative incidence or the subdistribution function.

$$\textit{Elimination risk:}\ \Pr(Y_{K+1}^{\bar{d}=0} \mid \overline{X}_k = \overline{x}_k)$$
$$\textit{Outcome-specific risk:}\ \Pr(Y_{K+1} \mid \overline{X}_k = \overline{x}_k)$$

Loss to follow up is generally understood as a form of censoring and often the analytical goal is to remove the influence of censoring from predictions. This can be conceived as a counterfactual estimand that posits a world in which we intervened to follow everyone fully until the conclusion of the study, i.e.

$$\Pr(Y_{K+1}^{\bar{c}=0} \mid \overline{X}_k = \overline{x}_k).$$

While traditional risk prediction techniques for competing risks and missing data and/or loss to follow up exist, they are often estimand-specific and therefore figuring out which to use can be confusing. On the other hand, the g-formula can readily be adapted to target any of the estimands discussed above, depending on which components of the joint distribution are explicitly modeled versus which are treated as censoring events. When a competing event is considered a censoring event, the g-formula targets the risk under the elimination of competing events, just as it targets the risk under the elimination of loss to follow up. When a competing event is considered an element of L_k, and therefore explicitly modeled, the g-formula targets the risk without elimination of competing events. Modified identification assumptions and g-formula expressions under censoring and competing events are given in section A.1.2 of the appendix.

1.2.6 Variable selection

Often, in clinical prediction tasks, the predictors in X_k are determined by the information available to the decision-maker rather than what might be optimal from a statistical or theoretical point-of-view. For instance, certain laboratory values may be cost prohibitive or may take too long to collect relative to the decision timeline. For factual prediction tasks, this is not an issue and the covariates in X_k can be determined by the operating constraints of the decision-maker. However, for counterfactual prediction tasks, during training X_k must include all L_k sufficient to ensure exchangeability of potential outcomes with respect to treatments A_k in order to yield unbiased predictions. In practice, this may necessitate selecting training data where either (1) exchangeability is assured by design, such as in a randomized controlled trial, or (2) covariate data for X_k is sufficiently rich to make identification plausible. Once accomplished, the g-formula can be modified to produce predictions based on a subset $V_k \subset X_k$ of covariates available to the decision-maker by summing/integrating out the covariates that are not available $V_k^* = X_k - V_k$ as shown in section A.2 of the Appendix.

1.2.7 Estimation

When the number of possible risk factor histories is small, we can estimate the components of the g-formula nonparametrically and sum over all possible histories. However, when the number of histories is large or when they include continuous covariates, we often rely on parametric models to estimate the components of the g-formula and approximate the high-dimensional sum/integral via Monte Carlo simulation, this estimation procedure is often referred to as the *parametric g-formula* and is described in detail in [22, 27–30]. Depending on the estimand the parametric g-formula requires us to specify models for the conditional hazard of the outcome $p(\overline{x}_k, \underline{x}_k, k; \theta)$, for the time-varying covariate trajectories for treatments $h(\overline{x}_{j-1}, j; \alpha)$ as well as possible confounders or other predictors $\ell(\overline{x}_{j-1}, j; \lambda)$, and for the conditional hazard of competing events $q(\overline{x}_{j-1}, j; \eta)$, which we index by parameters θ, α, λ, and η respectively. For instance, when targeting the outcome-specific risk under the natural course the parametric g-formula estimates

$$\sum_{\underline{x}} \sum_{k=k}^{K} p(\overline{x}_k, \underline{x}_k, k; \hat{\theta}) \prod_{j=k}^{K} [1 - p(\overline{x}_j, j; \hat{\theta})][1 - q(\overline{x}_{j-1}, j; \hat{\eta})] \ell(\overline{x}_{j-1}, j; \hat{\lambda}) h(\overline{x}_{j-1}, j; \hat{\alpha})$$

While the full sequence of steps have been outlined previously, here we modify the standard algorithm slightly to recover the appropriate conditional estimand for prediction. Example steps are:

1. Fit models for each component of the g-formula. Specifically,

   (a) Fit pooled (over persons and time) models $\ell(\overline{x}_{j-1}, j; \lambda)$ for the conditional distribution of each confounder in L_k as well as each predictor P_k at time k as a function of k, past treatment, confounder, and predictor history based on those who are alive, uncensored, and competing event free at k.

   (b) Fit pooled (over persons and time) models $h(\overline{x}_{j-1}, j; \alpha)$ for the conditional distribution of each possible treatment in A_k at time k as a function of k, past treatment, confounder, and predictor history based on those who are alive, uncensored, and competing event free at k.

(c) Fit a pooled logistic regression model $p(\overline{x}_k, \underline{x}_k, k; \theta)$ for the probability of the outcome by time $k+1$ as a function of k, past treatment, confounder, and predictor history based on those who are alive, uncensored, and competing event free at k.

(d) If there is a competing event and the target estimand is the risk without elimination of competing events, then also fit a pooled logistic regression model $q(\overline{x}_{j-1}, j; \eta)$ for the probability of the competing event by time $k+1$ as a function of k, past treatment, confounder, and predictor history based on those who are alive, uncensored, and competing event free at k.

2. For *each* conditional history of interest, approximate the sum (or integral) over *future* treatment and covariate histories by performing the following Monte Carlo simulation B number of times based on the intervention of interest. For $k > 0$:

   (a) Simulate covariate values from the fitted models $\ell(\overline{x}_{j-1}, j; \hat{\lambda})$ in Step 1(a) using previously simulated covariates and assigned treatment values through time $k-1$.

   (b) i. If the target is a deterministic regime, assign treatment according to the intervention based on simulated covariates and assigned treatment values through time $k-1$.

   ii. If the target is a random regime, simulate treatment according to the specified intervention density $f^g(a_j \mid \cdot)$.

   iii. Otherwise, if target is natural course, simulate treatment using fitted model for treatment $h(\overline{x}_{j-1}, j; \hat{\alpha})$ in Step 1(b) based on simulated covariates and treatment values through time $k-1$.

   (c) Compute the hazard of the outcome at time k from the fitted model $p(\overline{x}_k, \underline{x}_k, k; \hat{\theta})$ in Step 1(c) using previously simulated covariates and either assigned treatment values (deterministic regimes) or simulated treatment values (random regimes and natural course) through time k.

(d) If there is a competing event and the target estimand is the risk without elimination of competing events, compute the hazard of competing events at time k from the fitted model $q(\overline{x}_{j-1}, j; \hat{\eta})$ in Step 1(d) using previously simulated covariates and either assigned treatment values (deterministic regimes) or simulated treatment values (random regimes and natural course) through time k.

3. Calculate the average cumulative probability of the outcome by time K over all possible futures.

Confidence intervals for g-formula predictions can be obtained by using the bootstrap and re-fitting models $p(\overline{x}_k, \underline{x}_k, k; \theta)$, $h(\overline{x}_{j-1}, j; \alpha)$, $\ell(\overline{x}_{j-1}, j; \lambda)$, and $q(\overline{x}_{j-1}, j; \eta)$ in each resampled dataset, but calculating predictions using the same initial values at time k. Implementation of the modified g-formula algorithm above are available as an R package at https://github.com/boyercb/gmethods.

1.3 Simulations

To investigate the finite sample properties of risk prediction modeling using the g-formula as compared to a more conventional approach, we designed a simulation study. We generate data from a process meant to approximate the cardiovascular disease application in section 1.4. R code to reproduce the simulations can be found at https://github.com/boyercb/g-formula-prediction.

1.3.1 Setup and data generation process

We focus on a ten time point prediction process in which we have examination values for six variables $X_k = (L_0, L_{1_k}, L_{2_k}, P_0, P_{1_k}, A_k)$, consisting of five continous covariates $(L_0, L_{1_k}, L_{2_k}, P_0, P_{1_k})$, three of which ($L_{1_k}$, L_{2_k}, P_{1_k}) vary over time and two of which (L_0, P_0) are fixed at baseline, as well as a binary treatment variable (A_k). We allow for treatment-confounder feedback between L_{1_k}, L_{2_k}, and A_k as the former affect the decision to administer treatment and are themselves affected by treatment. We wish to predict cumulative risk of the survival outcome event (Y_k) for which there is optionally a competing risk (D_k) as well as the possibility of being lost to follow up (C_k). We allow past covariate values to affect future values.

More specifically, we draw baseline values L_0 and P_0 as well as initial values of L_{1_k}, L_{2_k} and P_{1_k} from Normal$(0, 1)$. Then for times $k \in \{1, \dots, 9\}$, we generate data according to the model

$$\begin{aligned}
L_{1_k} &\sim \text{Normal}\left(0.5 \cdot k + L_{1_{k-1}} + 0.25 \cdot L_{2_{k-1}} - 1.5 \cdot A_{k-1} - 0.5 \cdot A_{k-2}, 0.2\right) \\
L_{2_k} &\sim \text{Normal}\left(0.5 \cdot k + 0.25 \cdot L_{1_k} + L_{2_{k-1}}, 0.2\right) \\
P_{1_k} &\sim \text{Normal}\left(P_{1_{k-1}}, 0.2\right) \\
A_k &\sim \text{Bernoulli}(\text{expit}\{\log(0.001) + \log(4) \cdot L_{1_k} + \log(4) \cdot L_{2_k} + 4 \cdot A_{k-1} + \\
&\qquad \log(0.75) \cdot L_0 - \log(4) \cdot L_{1_k} \cdot A_{k-1} - \log(4) \cdot L_{2_k} \cdot A_{k-1}\}) \\
C_{k+1} &\sim \text{Bernoulli}\left(\text{expit}\{\log(0.01) + \log(0.5) \cdot A_{k-1}\}\right) \\
D_{k+1} &\sim \text{Bernoulli}(\text{expit}\{\log(0.005) + \log(1.25) \cdot L_{1_k} + \log(1.25) \cdot L_{2_k} + \log(1.25) \cdot P_{1_k} + \\
&\qquad \log(0.75) \cdot L_0 + \log(0.75) \cdot P_0 + \log(0.5) \cdot A_k\}) \\
Y_{k+1} &\sim \text{Bernoulli}(\text{expit}\{\log(0.0005) + \log(1.5) \cdot L_{1_k} + \log(1.5) \cdot L_{2_k} + \log(1.5) \cdot P_{1_k} + \\
&\qquad \log(1.25) \cdot L_0 + \log(0.5) \cdot P_0 + \log(0.5) \cdot A_k\})
\end{aligned}$$

Intercepts were chosen such that the cumulative risk of the outcome in untreated with mean covariate values was roughly 20% and treatment initiation over follow up was roughly 40%. For any of C_{k+1}, D_{k+1}, and Y_{k+1}, when the current value is one, future values are deterministically set to one and the others are set to zero. In simulations without competing risks D_k is set to zero at all time points.

In each simulation experiment, we generate training samples of size $N = 3000$ from the process above and fit all models. We then generate an independent test dataset of size $N = 3000$ and apply our fitted models to estimate the cumulative risk at time 10, $\Pr(Y_{10} \mid \overline{X}_{k^*} = \overline{x}_{k^*}) = p(\overline{x}_{k^*})$, conditioning on the first exam, first three exams, and first six exams, i.e. $k^* \in \{0, 3, 6\}$. We repeat this same sequence across $J = 500$ simulations and then evaluate the finite sample performance of each estimator based on time-dependent extensions of the mean-squared error and the area under the receiver operating characteristics curve, MSE(Δk, k^*) and AUC(Δk, k^*), which are ajusted for censoring and competing risks [31, 32] and further described in the Appendix.

1.3.2 Simulation scenarios and comparators

To evaluate the performance of the g-formula, we consider three prediction scenarios.

1. *Factual prediction.* Training and test data are generated from process above, except competing events are removed as D_{k+1} are set to zero at all time points. Target is cumulative risk $\Pr(Y_{10} \mid \overline{X}_{k^*} = \overline{x}_{k^*})$ for $k^* \in \{0, 3, 6\}$.

2. *Competing risk prediction.* Training and test data are generated from process above with competing events, i.e. D_{k+1} are drawn as described. Target is cumulative risk without elimination of competing risks $\Pr(Y_{10} \mid \overline{X}_{k^*} = \overline{x}_{k^*})$ for $k^* \in \{0, 3, 6\}$.

3. *Counterfactual prediction.* Training data are generated from process above but test data are generated from a population in which treatment is unnavailable, i.e. A_k are deterministically set to zero. Target is counterfactual cumulative risk $\Pr(Y_{10}^{a_k=0} \mid \overline{X}_{k^*} = \overline{x}_{k^*})$ for $k^* \in \{0, 3, 6\}$.

The third scenario is meant to mimic the scenario in which the estimand is the treatment-naive risk, but models are fit in a population in which treatment is initiated over the follow up period. In the second scenario, the target is the risk without elimination of competing events and performance metrics, i.e. MSE(Δk, k^*) and AUC(Δk, k^*), are adjusted for competing events.

The g-formula estimator is constructed using parametric models for covariates (L_{1_k}, L_{2_k}, P_{1_k}), treatment (A_k), the outcome (Y_k), and competing event (D_k) that are initially specified correctly relative to the data generation process. As comparators we use pooled landmark logistic regressions to model the discrete-time hazard of Y_k based on covariates $\overline{X}_{k^*}$ up to landmark times $k^* \in \{0, 3, 6\}$. That is we fit separate pooled logistic regressions conditional on covariate values up to k^* for the period from k^* to K. We separately consider landmark models using just the most recent values and those that include lagged terms for previous values.

To understand performance under model misspecification, we modify the data generation process above under two types of misspecification: misspecification of the covariate process and misspecification of the outcome process. In the case of the former, L_{1_k} is drawn from

$$L_{1_k} \sim \text{Normal}\big(\sqrt{0.5 \cdot k} + L_{1_{k-1}} + 0.05 \cdot \sum_{j=0}^{k-1} L_{2_j} - 1.5A_{k-1}\big)$$

but the models in the g-formula and landmark estimators are unchanged. In the latter, Y_{k+1} is drawn from

$$Y_{k+1} \sim \text{Bernoulli}\big(\text{expit}\{\log(0.00075) + \log(1.1) \cdot \sum_{j=0}^{k-1} L_{1_j} + \log(1.1) \cdot \sum_{j=0}^{k-1} L_{2_j} + \log(1.1) * \sum_{j=0}^{k-1} P_{1_j} + \log(1.25) \cdot P_0 + \log(0.5) \cdot L_0 + \log(0.5) \cdot A_k\}\big)$$

1.3.3 Simulation results

Simulation results under correct specification of parametric models are presented in Table 1.1. The best performing estimator in each row is highlighted in bold. In all scenarios the MSE and AUC of all estimators improved when we condition on additional exam values as a richer covariate history and a shortened prediction window yield better predictions. In the factual and competing risk prediction scenarios, the g-formula outperformed the landmark estimators with and without lags by a small margin, reflecting performance gains from the correct specification of the intermediate process as well as correct estimation of the subdistribution function when there are competing risks. In the counterfactual prediction scenario, the g-formula performed substantially better than the landmark methods, even when treatment was included in these models and naive adjustments were made (i.e. forcing treatment terms to zero). This was likely due to the fact that the g-formula correctly estimates the effect of removing treatment from the treated, including via paths mediated through time-varying counfounders L_{1_k} and L_{2_k}.

Results when models for the covariate process and the outcome process are misspecified are shown in Tables A.1 and A.2 in the Appendix. As expected performance of all estimators degrades under misspecification. However, they suggest that, at least under some data generation processes, the advantages of the g-formula for factual and competing risk estimands may be reduced when models are misspecified, as they inevitably are in practice. On the other hand, the advantages of the g-formula persist across all counterfactual prediction scenarios, as estimates remain less biased than those produced by the landmark estimators.

1.4 Application

In this section, we provide an example of factual and counterfactual prediction using the g-formula in real data.

The Framingham Study was initiated in 1948 as a longitudinal, population-based study of cardiovascular disease among 5,209 men and women in Framingham, Massachusetts. In 1971, 5,135 offspring of original participants of the study and their spouses were recruited to participate in the Framingham Offspring Study. Across nine examination cycles, members of the Framingham Offspring Study returned, on average, every 3 to 4 years for a physical examination, questionnaires, laboratory tests, and assessment of cardiovascular and other risk factors. Both the original study and the offspring study have been used extensively for the development of risk prediction models for cardiovascular disease [33]; however, most popular implementations use a single examination cycle rather than the full set.

We began follow up at the fifth examination (1991-1994) and included longitudinal data from examinations six (1994-1998) and seven (1998-2001). We restricted to 2,828 cohort members who were under 70 years of age, had complete baseline data, were not currently on lipid-lowering treatment and had no prior coronary heart disease event all at baseline (1991-1994). We used atherosclerotic cardiovascular disease (ASCVD), defined as a nonfatal myocardial infarction, coronary heart disease death, or ischemic stroke, as our primary outcome. There were 192 ASCVD events and 119 non-ASCVD deaths during follow up. Descriptive characteristics of our sample across examination cycles are shown in Table A.3.

To implement the g-formula, we used separate regressions to model ASCVD, non-

Table 1.1: *Monte carlo simulation results comparing g-formula and landmark approaches.*

	MSE(Δk, k^*)			AUC(Δk, k^*)		
k^*	g-formula	landmark	landmark (lags)	g-formula	landmark	landmark (lags)
Scenario 1: Factual prediction						
0	**0.112**	0.116	0.116	**0.881**	0.880	0.880
	(0.006)	(0.007)	(0.007)	(0.011)	(0.011)	(0.011)
3	**0.099**	0.104	0.104	**0.903**	0.901	0.900
	(0.006)	(0.006)	(0.006)	(0.010)	(0.010)	(0.010)
6	**0.087**	0.091	0.091	**0.918**	0.916	0.916
	(0.006)	(0.006)	(0.006)	(0.010)	(0.010)	(0.010)
Scenario 2: Competing risk prediction						
0	**0.102**	0.104	0.104	**0.893**	0.892	0.892
	(0.006)	(0.007)	(0.007)	(0.014)	(0.014)	(0.014)
3	**0.097**	0.099	0.099	**0.915**	0.913	0.911
	(0.006)	(0.006)	(0.006)	(0.013)	(0.013)	(0.013)
6	**0.089**	0.090	0.091	**0.925**	0.923	0.921
	(0.006)	(0.006)	(0.006)	(0.012)	(0.012)	(0.012)
Scenario 3: Counterfactual prediction						
0	**0.164**	0.388	0.388	**0.960**	0.943	0.943
	(0.009)	(0.018)	(0.018)	(0.006)	(0.007)	(0.007)
3	**0.116**	0.290	0.285	**0.970**	0.965	0.966
	(0.007)	(0.014)	(0.014)	(0.004)	(0.005)	(0.005)
6	**0.086**	0.158	0.158	**0.972**	0.971	0.971
	(0.006)	(0.012)	(0.015)	(0.004)	(0.004)	(0.004)

Note:
All results based on 500 Monte Carlo simulations using data generation process described in section 3. Standard deviations of Monte Carlo estimates are provided in parentheses. The best performing estimator is shown in **bold**. All simulations use correctly specified models. For results under misspecification see the appendix.

ASCVD death and each of the following time-varying risk factors: cigarette smoking, BMI, diabetes, anti-hyertension medication use, lipid-lowering medication use, serum LDL cholesterol, serum HDL cholesterol, and systolic blood pressure. We used pooled discrete-time logistic regression to model the probability of ASCVD and the probability of non-ASCVD death in each year. Each time-varying risk factor was classed as binary, binary-to-failure, or continuous, and then modelled using a generalized linear model as specified in Table A.4. To increase efficiency all models were pooled over all examination cycles. All models included, as predictors, age, the two previous values of all time-varying risk factors, and the fixed covariates baseline age and sex. Binary predictors were entered into the models as indicators; continuous predictors were entered as polynomials (linear, quadratic and cubic) and restricted cubic splines in sensitivity analyses. Table A.4 summarizes the information on the covariates included in the primary analysis. Tables tables A.5 and A.6 provide estimated model coefficient values and fit statistics for the outcome and covariate models respectively. We estimate the 10-year cumulative risk using the Monte Carlo procedure outline in section 1.2.7 with 500 simulations per individual, or 1,297,500 total.

Figure 1.2 shows an example of the predictions generated using the g-formula. In the figure, we present predictions for a single covariate profile: a 60 year-old male smoker with a BMI of 30 kg/m^2, no history of diabetes, no history of treatment and elevated risk factors levels. The first row shows the full 10-year predicted risk trajectory based on baseline values alone. The left panel shows the predicted risk of ASCVD and the right panel shows the expected trajectory of their risk factors. Unlike conventional methods, which only model the outcome, the latter gives us some insights into the expected "natural course" of similar individuals in Framingham dataset. For instance, the model suggests that men like him have a fair chance of starting lipid-lowering or anti-hypertensive medication over the follow up period. The second row shows an example of an updated risk prediction for the same covariate profile at a subsequent examination 3 years later in which the man has now developed diabetes, but the levels of other risk factors are the same. This prediction is conditional on the man's full examination history to date. Compared to the previous

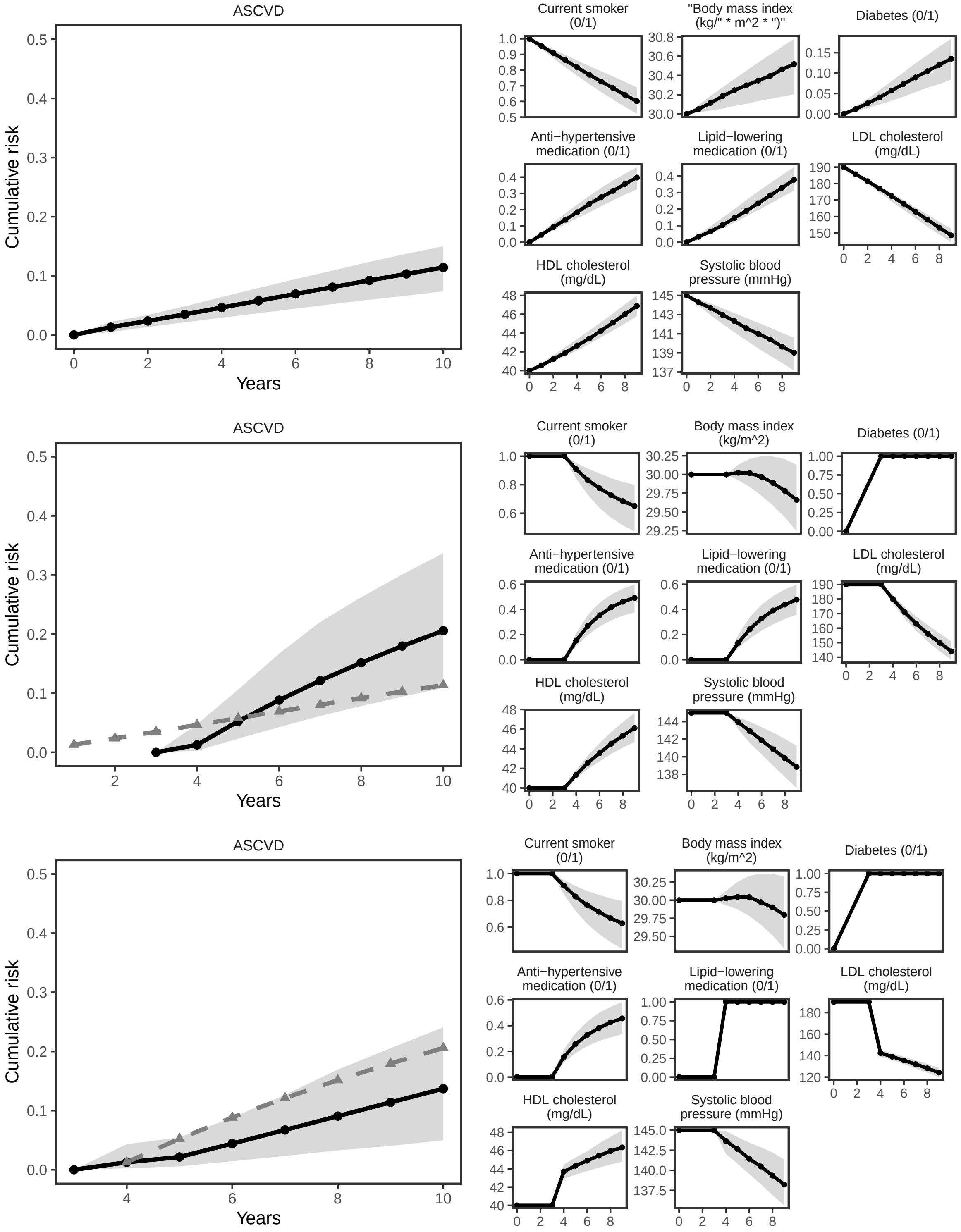


Figure 1.2: *Example risk predictions using the g-formula for a single covariate profile. Panel A is the 10-year risk under the natural course. Panel B is updated after 3 years and a diabetes diagnosis. Panel C is the counterfactual risk under statin-initiation at visit at year 3. Dotted lines show predictions from panel above for comparison.*

Table 1.2: *Optimism-corrected estimates of model performance for factual prediction in the Framingham Offspring Study.*

k^*	Model	MSE(Δk, k^*)	95% CI	AUC(Δk, k^*)	95% CI
0	g-formula	0.0607	(0.0551, 0.0671)	0.746	(0.719, 0.773)
	landmark	0.0613	(0.0537, 0.0698)	0.740	(0.707, 0.774)
	landmark (lags)	0.0613	(0.0537, 0.0698)	0.740	(0.707, 0.774)
3	g-formula	0.0488	(0.0435, 0.0552)	0.738	(0.708, 0.771)
	landmark	0.0493	(0.0420, 0.0568)	0.732	(0.694, 0.771)
	landmark (lags)	0.0497	(0.0427, 0.0572)	0.732	(0.694, 0.773)
6	g-formula	0.0272	(0.0227, 0.0319)	0.717	(0.661, 0.766)
	landmark	0.0276	(0.0215, 0.0339)	0.702	(0.641, 0.758)
	landmark (lags)	0.0283	(0.0225, 0.0343)	0.684	(0.621, 0.738)

prediction (dotted line), the man's risk of developing ASCVD has increased. Finally, the last row shows an example counterfactual prediction at the same examination which estimates the risk if the man were to initiate lipid-lowering treatment. Compared to the factual prediction at 3 years (dotted line), his risk is reduced. However, unlike previous predictions this one rests on the strong, untestable, assumptions outlined in section 1.2.4.

To assess the performance of factual predictions using the g-formula, we conducted the following internal validation. We drew 500 bootstrap replicates from the Framingham Offspring sample and repeated the entire modeling process to estimate the optimism-corrected dynamic MSE and AUC at baseline ($k^* = 0$), after four years ($k^* = 3$), and after seven years ($k^* = 6$). For comparison we repeated the same procedure using landmark Cox regressions using the same covariate set with and without lagged terms. The latter mimics models traditionally used in risk prediction for cardiovascular disease such as the Pooled Cohort Equations [34]. The results are displayed in Table 1.2. At all time points the g-formula outperforms traditional methods, although the differences are modest. In the appendix, we show that the g-formula predictions are also better calibrated against the ASCVD-specific risk than traditional methods.

To highlight the use of the g-formula for counterfactual prediction we estimated the treatment-naive risk, i.e. the risk if, contrary to fact, none of the participants initiated

Table 1.3: *Population-level risk estimates under lipid-lowering therapy interventions using the g-formula in the Framingham Offspring Study and then transported to Framingham Study.*

Intervention	Risk	95% CI	RR	95% CI	% intervened on
Framingham Offspring Cohort					
Never treat	7.6 %	(7.0%, 8.2%)	ref		13%
Natural course[1]	7.1%	(7.0%, 8.2%)	0.95	(7.0%, 8.2%)	0%
Always treat	6.0%	(7.0%, 8.2%)	0.79	(7.0%, 8.2%)	87%
Framingham Original Cohort					
Never treat[2]	11.2%	(7.0%, 8.2%)	ref		0%
Offspring course	9.9%	(7.0%, 8.2%)	0.88	(7.0%, 8.2%)	18%
Always treat	7.3%	(7.0%, 8.2%)	0.65	(7.0%, 8.2%)	100%

[1] For reference, the observed risk in the Framingham Offspring sample was 7.1% using an inverse probability of censoring weighted estimator.

[2] For reference, the observed risk in the (untreated) Framingham sample was 13.7% using an inverse probability of censoring weighted estimator.

lipid-lowering therapy over follow up. When counselling a patient on whether to start treatment, the treatment-naive risk is the more relevant, but harder to estimate, risk for weighing costs and benefits of treatment. To estimate the treatment-naive risk using the g-formula, we simulate the risk under an intervention which sets the indicator lipid-lowering therapy to zero in all intervals. Table 1.3 shows the population-level risk estimates under this intervention as compared to the natural course and an intevention which sets the indicator to one in all intervals (always treat). As expected the treatment-naive risk is larger than both the natural course and the always treat intervention. Compared with the factual predictions from tradiational methods, as shown in Figure 1.3, the treatment-naive predictions from the g-formula are higher.

Assessing the performance of counterfactual predictions is substantially more challenging as potential outcomes are not observed for those who do initiate treatment over follow up. On the one hand, we can take some comfort from the fact that the population-level estimate of the relative risk of lipid-lowering therapy (RR = 0.79) in Table 1.3 is consistent with those from cholesterol treatment trials. However, this could simply be co-incidence. It also doesn't speak to the validity of individual predictions, which are conditional and

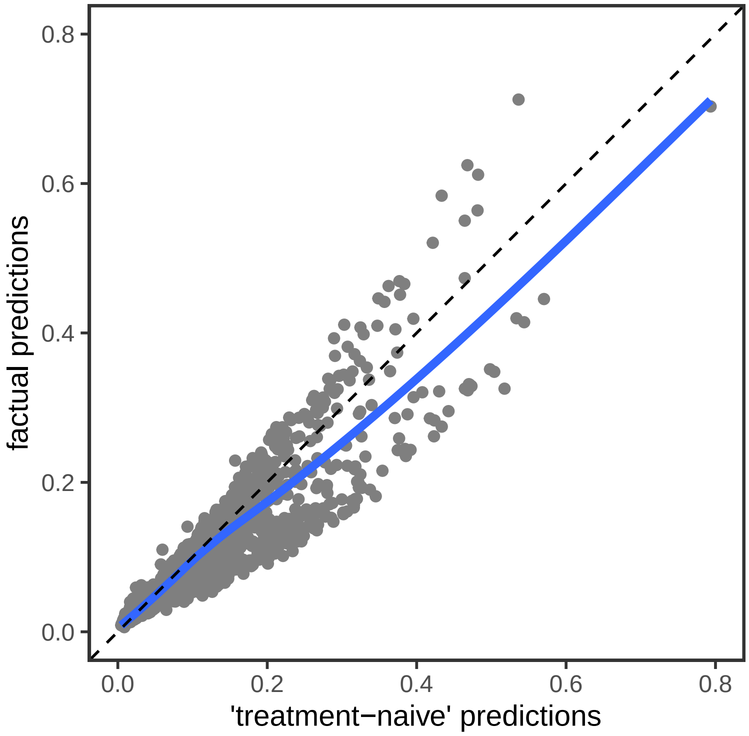


Figure 1.3: *Comparison of factual predictions using traditional methods and 'treatment-naive' predictions using the g-formula.*

require correct modeling of effect modifiers.

In lieu of better options and in the model-agnostic spirit in which most prediction models are deployed, we conduct the following validation exercise. In the original Framingham Study, modern lipid-lowering therapies, notably statins, were unavailable. Therefore the participants in the original study, who were direct relatives of those in the Offspring study, are a treatment-naive cohort that, in some sense, may resemble the counterfactual experience of the Offspring under no treatment. In practice, this is almost surely false, as there are many other differences in dietary patterns, education, and non-statin treatment options between the Original and Offspring cohorts. However, given that prediction models are often applied agnostically to new settings and judged on their "performance" rather than their "correctness" in an abstract sense, we investigate whether using a counterfactual treatment-naive model performs better than a simple factual model in the Original cohort. We used the 10th examination cycle (1968-1971) as a baseline as this was the only time a full serum cholesterol panel was conducted, and retained exams 11 through 16 for evaluation of model fit. We applied both the g-formula and the landmark models used previously to the original cohort. Compared to the observed 10-year risk in the Framingham Study of 13.7% the average of treatment-naive predictions was 11.2%, while the average using the landmark predictions was only 9.5%, indicating the treatment-naive predictions were closer to observed. In the appendix, we show that MSE and AUC of transported treatment-naive predictions outperform those of traditional methods. This suggests that, if the choice were between the two models only, the counterfactual model would be the better option. However, it's probable that a more structured approach to transporting the model which considered a richer set of possible differences in participants over follow up in the two cohorts would perform better still.

1.5 Discussion

We applied the parametric g-formula to provide more reliable estimates of time-dependent factual and counterfactual risk using longitudinal data. We demonstrated the flexibility of the g-formula in targeting different prediction estimands of interest to clinicians and researchers and discussed the assumptions necessary for unbiased estimation. Through simulation we showed some of the trade-offs of modeling the natural course of treatment and risk factors. When models are correctly specified (or approximately so) we find that the g-formula performs better than conventional approaches by making efficient use of longitudinal data. However, we also show that poor specification of covariate models can lead to bias. We further demonstrate the flexibility of the g-formula to target competing-risk and counterfactual estimands. We conclude by applying these insights to the practical problem of cardiovascular disease prediction in the Framingham Offspring Study.

Our work builds on a growing literature clarifying estimands for factual prediction and counterfactual risk prediction using the potential outcomes framework [11, 15] as well as other efforts to synthesize traditional prediction modeling approaches with modern casual inference methods. We focus here on the (non iterated conditional expectation) g-formula as we see it as most similar to traditional strategies based on outcome regression modeling, but there are other estimators that could also be used to target counterfactual prediction estimands, including the iterated conditional expectation version of the g-formula [35] as well as those based on maringal structural models or inverse-probability weighting [17, 36] which model the treatment process. There are also doubly- or multiply-robust estimators [37–40] which consistently estimate counterfactuals under correct specification of either a model for the outcome or a model for treatment. Finally, our work is also related to a large literature on the estimation of the conditional average treatment effect (CATE), which proposes estimators for the treatment *effect* rather than the expected outcome risk.

Dickerman *et al.* propose an alternative way to use the parametric g-formula to target counterfactual prediction estimands. Specifically, they use the joint distribution for the outcome under a proposed intervention implied by the parametric g-formula to generate alternative training data. This simulated data can then be used by any prediction algorithm to create a new model for the outcome under the proposed intervention. Their approach nicely separates the casual and prediction tasks. However, to date, there isn't a clear way to construct valid confidence intervals for the models fit on the simulated data.

In the prediction literature, when targeting factual prediction estimands our approach most resembles joint-modeling [41] as both the g-formula and joint modeling posit models for the joint distribution of a survival outcome and a set of time-varying predictors; however, the latter generally uses a random or mixed-effect modeling framework, assumes covariates are measured with error, has mostly been used to predict changes based on the time-evolution of a single biomarker, with a few exceptions. In theory though many of the advantages of adopting a joint modeling approach apply equally to the g-formula, such as the ability to dynamically update predictions over time [42]. Several previous studies have also showed the benefits of incorporating data from longitudinal assessments in clinical prediction models using traditional regression-based approaches [43, 44]. This is consistent with results in our simulation and application.

We note several important limitations of the present study. First, as emphasized throughout counterfactual predictions require untestable assumptions that may not be met in observational settings. Estimation of counterfactual predictions requires careful attention to all assumptions. Unlike in RCTs, the conditional exchangeability assumption is not satistifed by design and requires subject matter expertise to determine whether sufficient confounding control is, even approximately, possible in any given circumstance. Real-world analoges of the interventions estimated in observational settings may not exist or may not be applicable to the clinical decision in question [45]. Analysts must also pay careful attention to possible structural sources of non-positivity and modeling assumptions that may lead to poor extrapolations [46]. Finally, throughout we have implicitly assumed that covariates were measured without error, which is almost certainly violated in practice. In principle, measurement error could be accounted for within our framework, although example implementations of the g-formula which account for measurement error are lacking.

Second, as our simulations have shown, applying the parameteric g-formula to model the time-dependent evolution of treatment and risk factors requires the correct specification of several longitudinal models for the evolution of covariates over time. This is in contrast to a traditional regression approach which requires the correct specification of a single outcome model to yield valid estimates. The trade-off is that when these models are correct or nearly correct the g-formula-based predictions can substantially improve on traditional approaches. While previous literature has suggested ways to check for gross model misspecification [30] these are often harder to verify when there's loss to follow up and competing risks, although we do so here using inverse probability of censoring weighting. For counterfactual prediction, an additional challenge is that for even a moderate number of time points it may be impossible to correctly specify parametric models under the null, i.e. the so called *g-null paradox* [47, 48]. An alternative is to use more flexible modeling approaches for the components of the g-formula. One could also consider other nonparametric methods based on machine learning [49], however the asymptotic properties of these estimators is not well established. In some cases, applying machine learning methods to estimators that are not based on the efficient influence function, i.e. are not "doubly-robust", such as the g-formula can lead to estimators that actually perform worse than those based on parametric models [50].

While our application focused on a observational cohort, in principle our g-formula based approach could be equally applied to data from a randomized trial, in which case random assignment ensures exchangeability, at least at baseline, by design. This may increase the plausibility of the estimation of some counterfactual prediction estimands. However, given the challenges in recruitment and strictness of eligibility criteria, additional modeling assumptions may still be required to transport this model to a non-trial population without bias [51, 52].

In summary, the parametric g-formula can be a flexible tool for both factual and counterfactual risk prediction. As large longitudinal databases such as electronic health records become increasingly common, it is interesting to consider how more dynamic approaches to risk prediction such as the one considered in this paper may be integrated into existing systems.

Chapter 2

Target trials for prediction: emulating a trial to estimate the treatment-naive risk[1]

Clinical prediction models are commonly used to determine treatment eligibility. However, depending on the data used to train the model, predicted risks may include the possibility that treatment is initiated over follow up, which can lead to underallocation. A better option would be to determine treatment eligibility using a model for the treatment-naive risk, that is the counterfactual risk had no one received treatment. Yet, outside of a randomized or single arm trial, estimating this risk generally requires the tools of causal inference. In this paper, we propose methods for estimating the treatment-naive risk based on emulating a target trial corresponding to the clinical decision in question. We use inverse probability of censoring weighting and g-estimation of structural nested accelerated failure time models to estimate the effect of removing treatment from the treated. We apply these methods to create a statin-naive risk prediction model in the Multi-Ethnic Study in Atherosclerosis.

Clinical prediction models are commonly used to determine treatment eligibility. However,

[1] Co-authored with James Robins, Andrew Beam, and Goodarz Danaei

depending on the data used to train the model, predicted risks may include the possibility that treatment is initiated over follow up, which can lead to underallocation. A better option would be to determine treatment eligibility using a model for the treatment-naive risk, that is the counterfactual risk had no one received treatment. Yet, outside of a randomized or single arm trial, estimating this risk generally requires the tools of causal inference. In this paper, we propose methods for estimating the treatment-naive risk based on emulating a target trial corresponding to the clinical decision in question. We use inverse probability of censoring weighting and g-estimation of structural nested accelerated failure time models to estimate the effect of removing treatment from the treated. We apply these methods to create a statin-naive risk prediction model in the Multi-Ethnic Study in Atherosclerosis.

2.1 Introduction

A common use of prediction models in clinical care is to guide decisions about initiating treatment [1]. This generally involves a model-based estimate of a patient's risk and then a decision rule to determine whether they should start treatment [53]. For instance, the AHA guidelines on the treatment of cholesterol [3] state that patients aged 40 to 75 with serum LDL cholesterol levels above 70 mg/dL and no history of diabetes should initiate a moderate intensity statin if their predicted 10-year risk of cardiovascular disease exceeds 7.5% based on the pooled cohort equations.

Ideally, the risk prediction model would be estimated in a single arm trial in which treatment is withheld from all participants. Alternatively, it could be estimated in the control arm of a two arm randomized trial comparing treatment against a placebo. Likewise, the decision rule should be determined by identifying the risk threshold that maximizes expected clinical utility while minimizing any costs or adverse effects. However, for many reasons, this paradigm is rarely realized in practice. For one, changes in the clinical landscape and eligibility drift often lead to a mismatch between the eligibility criteria for trials, which are more restrictive, and indications for use in clinical care, which are more relaxed [54]. Additionally, trial data may not be conducive to the development of a prediction model as relevant predictors may not be collected, outcomes may be composite or surrogate, and follow up may be shorter than that deemed necessary for prediction. Finally, changes in risk over time due to shifts in population health or other treatment innovations may render original risks obsolete [9, 55]. Instead, observational data are often used to train the model and the decision rule is determined based on a combination of evidence summaries and expert consensus.

Yet, as previous work has noted [11, 15, 17, 56], an issue arises when the model is trained in a cohort where participants initiate treatment over the follow up period. In this case, risk estimates may no longer be appropriate as they include the possibility that the patient will be treated. We term these risks the *natural course* risk as they represent the risk under the natural course of treatment in the derivation cohort. In contrast, the risk desired is often called the *treatment-naive* risk as it is the risk that would be obtained in a cohort where treatment is unavailable or withheld [11].

If treatment reduces the risk, natural course estimates may understate the true risk, particularly among those most likely to initiate treatment. Depending on the decision rule, using the natural course estimates might yield a substantial fraction of people not receiving treatment, who otherwise might have been treated using the true treatment-naive risk. This fraction may grow further as treatment use increases, until eventually the decision rule is no longer useful. This has prompted efforts to estimate the treatment-naive risk from observational data [17, 36, 57]. However, estimating the treatment-naive risk in a non-naive cohort generally requires the tools of causal inference.

In this paper, we propose two methods for estimating the treatment-naive risk based on emulating a target trial [18, 19] that matches the conditions of the clinical decision under consideration. First, we use g-estimation of structural nested accelerated failure time models (SNAFTM) [58–61] to construct pseudo-outcomes representing a participant's hypothetical failure time under no treatment. These pseudo-outcomes can then be used to predict the treatment-naive risk conditional on the subset of predictors desired using any prediction algorithm. Second, we censor participants when they initiate treatment and use inverse probability of censoring weights to construct a treatment-naive pseudopopulation. These weights can also be used by many prediction algorithms to construct models for the treatment-naive risk. Both methods effectively separate the causal inference and prediction tasks.

2.2 Methods

2.2.1 Setup and Notation

Consider the common model development setting in which we observe i.i.d. longitudinal samples $\{O_i\}_{i=1}^n$ from n participants following distribution $\mathbb{P}$. For each observation, let

$$O_i = (\overline{X}_k, \overline{A}_k, \overline{C}_{k+1}, \overline{Y}_{k+1}, T)$$

where X_k is a vector of time-varying covariates measured at time k, A_k is an indicator of treatment in the interval $(k, k+1]$, C_{k+1} is an indicator of loss to follow-up by time $k+1$, Y_{k+1} is an indicator of whether the outcome of interest has occurred by time $k+1$, and T is the failure time for outcome Y_{k+1} that is either exactly observed or interval censored. Overbars denote past history of a variable and underbars to denote future history, such that $\overline{X}_k = (X_0, \ldots, X_k)$ and $\underline{X}_k = (X_k, \ldots, X_K)$. Capital letters represent random variables and their lower case equivalents are realizations.

By definition $C_0 \equiv 0$ and $Y_0 \equiv 0$ as we restrict to those who are uncensored and event-free at the start of follow up. We assume that covariates in X_k can be further categorized into predictors of the outcome which do not act as confounders (P_k), as well as joint determinants of the outcome and treatment variables which act as confounders (L_k), i.e. $X_k = (L_k, P_k)$. By convention, when $Y_k = 1$ then $\underline{Y}_{k+1} = 1$, $\underline{X}_{k+1} = 0$, and $\underline{C}_{k+1} = 0$, and likewise when $C_k = 1$ then $\underline{Y}_{k+1} = 0$, $\underline{X}_{k+1} = 0$, and $\underline{C}_{k+1} = 1$. An example directed acyclic graph for a two time point process under the above process is shown in Figure 3.1.

To define causal estimands of interest, let Y^a be the potential outcome under an inter-

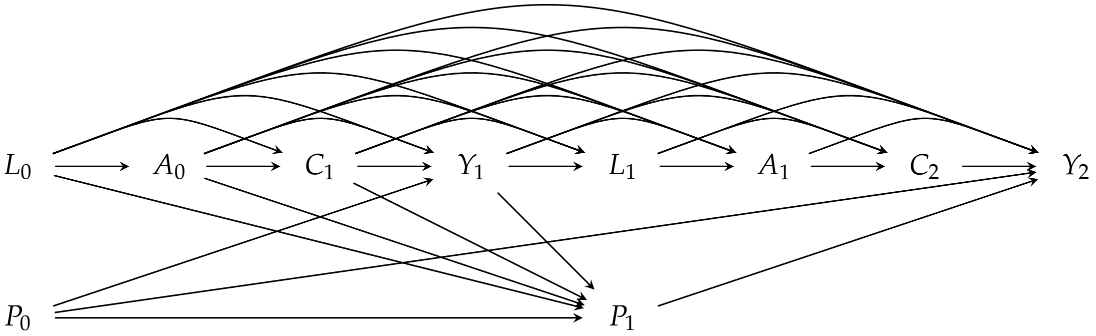


Figure 2.1: *Example two time point directed acyclic graph for prediction. In the main text, we omit the possibility of an unmeasured confounder U with arrows into L_0, Y_1, L_1, and Y_2 for simplicity, but note that the treatment-naive risk could still be identified if such U were present.*

vention which sets treatment A to a. For a sequence of time-varying treatments $\overline{A}_k$, we further define a *treatment regime* as a collection of functions $\{g_k(\overline{a}_{k-1}, \overline{x}_k) : k = 0, \ldots, K\}$ for determining treatment assignment at each time k, possibly based on past treatment and covariate history. In this paper, we are primarily concerned with the "never treat" regime $g = (0, 0, \ldots, 0)$ also denoted $\overline{a} = 0$.

2.2.2 Defining the treatment-naive risk

To better inform clinical decision-making, we'd like to estimate the treatment-naive risk, that is, the risk of the outcome at time t if, possibly contrary to fact, all participants remained untreated during the follow up period, or

$$\Pr(T^{\overline{a}=0} \leq t \mid X^*)$$

conditional on a subset of clinical predictors, X^* where $X^* \subset X_0$, that are available during a preliminary examination visit and are commonly selected for their ease of collection and prognostic value rather than the causal structure of their relationship with the outcome.

If we observed $T^{\overline{a}=0}$ for everyone, we could estimate the treatment-naive risk by defining a model $\Pr(T^{\bar{a}=0} \leq t \mid X^*) = g(X^*; \beta)$, such as the Cox proportional hazards model

$$\lambda(t \mid X^*) = \lambda_0(t) \exp(\beta' X^*)$$

where parameters are determined by maximizing the partial likelihood. Using estimated values $\widehat{\beta}$, we can calculate the risk the risk using

$$\Pr(T^{\bar{a}=0} \leq t \mid X^*) = 1 - \exp\{-\widehat{\Lambda}_0(t)\}^{\exp(\widehat{\beta}' X^*)}$$

where $\widehat{\Lambda}_0(t)$ is an estimate of the cumulative baseline hazard using, for instance, the Breslow or Kalbfleisch and Prentice estimator. Unfortunately, instead of $T^{\bar{a}=0}$ we observe T which, under consistency, is equivalent to the potential outcome under an intervention which set the treatment to its natural course value, that is the treatment value actually observed for each individual $T = T^{\overline{A}_k}$ and therefore we say $\Pr(T < t \mid X^*)$ is the risk under the natural course of treatment.

2.2.3 Identification assumptions

To estimate the treatment-naive risk requires additional assumptions. Namely, we require

1. *Sequential Exchangeability:* $T^{\bar{a}=0} \perp\!\!\!\perp A_k \mid \overline{X}_k, \overline{A}_{k-1}, T > k$

2. *Consistency:* $T = T^{\bar{a}=0}$, $\overline{Y}_{k+1} = \overline{Y}_{k+1}^{\bar{a}=0}$, and $\overline{X}_k = \overline{X}_k^{\bar{a}=0}$ if $\overline{A}_k = 0$

and either of

3a. *Positivity:* $\Pr(A_k = 0 \mid \overline{X}_k, \overline{A}_{k-1} = 0, T > k) > 0$

3b. *Known semi-parametric model:* $T^{\bar{a}=0}$ follows a SNAFTM.

for all $k = 0, \ldots, K$. The first condition stipulates that treatment at time k is conditionally independent of the counterfactual failure time under no treatment. This would be ensured by design in a sequentially randomized trial in which, at each time k, participants are randomized to treatment or no treatment conditional on their past treatment and covariate history. The second condition implies that observed outcomes and covariates among the untreated reflect potential outcomes under no treatment. It would be violated if, for instance, there were multiple hidden versions of the therapy under consideration.

When both conditions hold, estimation of the treatment-naive risk is possible, provided that we further assume $T^{a=0}$ is related to T via a known semiparametric model, a SNAFTM, discussed in the next section.

2.2.4 Structural Nested Accelerated Failure Time Models

Structural nested accelerated failure-time models are models for the removal of treatment which assume that the counterfactual time under no treatment $T^{\overline{a}=0}$ is related to observed time via

$$T^{\overline{a}=0} = \int_0^T \exp\{\gamma(t, \overline{A}_t, \overline{X}_t; \psi)\}dt \tag{2.1}$$

where $\gamma(t, \overline{A}_t, \overline{X}_t; \psi)$ is the instantaneous expansion (or contraction) in survival time comparing treatment at time t to no treatment. Often the simplifying assumption is made that this effect of treatment is constant over time, in which case

$$T^{\overline{a}=0} = \int_0^T \exp(\psi A_t)dt \tag{2.2}$$

and parameter ψ is the survival time ratio comparing continuous treatment to no treatment. In either case, under the assumptions of section 2.2.3, ψ can be consistently estimated from the observed data using g-estimation or inverse probability weighting.

Because they are models for the removal of treatment, SNAFTMs are a natural choice for estimating the treatment-naive risk. Unlike alternative methods, such as marginal structural models, they do not require a positivity assumption, i.e. that there is a nonzero possibility of receiving all levels of treatment in every history strata [62]. This is because ψ is only defined for those who are treated during the follow up period. Indeed, note that under the model above for those for whom $\overline{A}_k = 0$ equation 2.2 simplifies to $T^{\overline{a}=0} = T$. Furthermore, estimating the treatment-naive-risk requires a weaker *partial* exchangeability condition. Only the potential outcome under a single regime, $\overline{a} = 0$, must be conditionally independent of past treatment and covariate history, rather than the potential outcome under all possible regimes.

SNAFTM are *nested* in the sense that for any time k we can write

$$T^{(\overline{A}_k,0)} = \int_k^T \exp(\psi A_t)dt$$

where the model is now for the removal of treatment from time k to T, and relate it back to equation 2.2 by partitioning the integral as

$$T^{\bar{a}=0} = \int_0^k \exp(\psi A_t)dt + \int_k^T \exp(\psi A_t)dt$$

Note that from the perspective of someone alive at time k the first integral is a function of past treatment history, $\overline{A}_{k-1}$, only and therefore a fixed constant and uformative for estimating ψ. As suggested in [18, 61, 63], this feature of SNAFTMs permits two conceptualizations of the trial being targeted in an observational study. We can think of the study as a single sequentially randomized trial in which, at each time k, participants are randomly assigned to treatment or no treatment conditional on their past history. Alternatively, we can think of it as sequence of k nested randomized controlled trials, starting at each visit, where eligible participants who are event free are randomized to treatment or no treatment conditional on their baseline history and followed until time K. The latter interpretation is the basis for commonly used methods for emulating nested target trials from observational data.

As written, the structural models above are deterministic or rank-preserving as they assume the counterfactual failure time under no treatment $T^{\bar{a}=0}$ can be computed without error from T and ψ. This implies, for instance, that two participants with same exposure history and failure time would have the exact same failure time in the absence of treatment, which is nonsensical. However, as has been shown previously [59], we can apply the g-estimation and pseudo-outcome regression methods elaborated below unchanged whether or not rank preservation is strictly true. Finally, a particular SNAFTM $\gamma(t, \overline{A}_t, \overline{X}_t; \psi^*)$ may be misspecified, relative to the true model $\gamma(t, \overline{A}_t, \overline{X}_t; \psi)$, when there is time-varying effect modification by past treatment and covariate history that is unnaccounted for by the proposed parameter vector ψ^*. This occurs when overly parsimonious specifications are chosen, either due to omitting particular effect modifiers or misspecifying their functional form.

2.2.5 G-estimation

Here we briefly review g-estimation of the parameters of a SNAFTM, although this is covered in more detail elsewhere [60, 61, 64]. It is often convenient to first define pseudo-outcome $H(\psi^*)$ as

$$H(\psi^*) = \int_0^T \exp\{\psi^* A(t)\} dt$$

such that $H(\psi) = T^{\overline{a}=0}$ when ψ^* is set to the true value ψ, but may, in general, be computed for any value ψ^*. Equivalently, under a nested conceptualization we can define pseudo-outcome $H(k, \psi^*)$ as

$$H(k, \psi^*) = \int_k^T \exp\{\psi^* A(t)\} dt$$

where each subject now contributes K pseudo-outcomes (one for each trial).

Under the sequential exchangeability assumption in section 2.2.3,

$$H(k, \psi) \perp\!\!\!\perp A_k \mid \overline{X}_k, \overline{A}_{k-1}, T > k$$

when evaluated at the true value ψ. This permits two equivalent approaches for estimating ψ. First, note that the conditional independence assumption above implies that $\Pr\{A_k = a_k \mid \overline{X}_k, \overline{A}_{k-1}, T > k\} = \Pr\{A_k = a_k \mid \overline{X}_k, \overline{A}_{k-1}, T > k, H(k, \psi)\}$. Thus, we can find ψ by searching over a grid of possible ψ^* values and determining the value of ψ^* for which A_k is conditionally independent of $H(k, \psi^*)$. For instance, we find the value of ψ^* that makes this $\theta_3 = 0$ in the pooled logistic regression model

$$\text{logit}[\Pr\{A_k = a_k \mid \overline{X}_k, \overline{A}_{k-1}, T > k, H(k, \psi^*)\}] = \theta_0 + \theta_1' \overline{X}_k + \theta_2' \overline{A}_{k-1} + \theta_3 H(k, \psi^*)$$

where we regress treatment at time k on past treatment and covariate history as well as the pseudo-outcome $H(\psi^*)$. Alternatively, ψ can be determined by solving the estimating equation

$$\sum_{i=1}^{N}\sum_{k=0}^{K} H_i(k,\psi^*)[A_{i,k} - \Pr\{A_k = a_k \mid \overline{X}_k, \overline{A}_{k-1}, T > k\}] = 0$$

where $\Pr\{A_k = a_k \mid \overline{X}_k, \overline{A}_{k-1}, T > k\}$ can be estimated using a pooled logistic regression model for treatment at time k conditional on past treatment and covariate history. Both methods require the correct specification of a model for $\Pr\{A_k = a_k \mid \overline{X}_k, \overline{A}_{k-1}, T > k\}$, although doubly-robust estimating equations are also possible [60].

A complication arises when failure times are not observed for all participants as it is not possible to calculate $H(\psi)$ or $H(k,\psi)$ for everyone. In this case, the g-estimation above procedure must be modified to accomodate administrative censoring. Specifically, we replace $H(\psi)$ with a function of $H(\psi)$ and the end of follow up time $K(\psi)$ which is observed for all participants. Examples are

$$\Delta(\psi) = \mathrm{I}\{H(\psi) < K(\psi)\}$$
$$Z(\psi) = \min\{H(\psi), K(\psi)\}$$

where the first is just an indicator that is one if $H(\psi)$ is observed and zero if it is administratively censored, and the second is the minimum of $H(\psi)$ and $K(\psi)$. Because any such function of $H(\psi)$ is also conditionally independent of past treatment and covariate history, we can replace $H(\psi)$ with $\Delta(\psi)$ or $Z(\psi)$ in previous paragraphs and estimate ψ in the same way. In the g-estimation literature, this process is often called *artificial censoring* because treated participants who are uncensored in observed data may be censored in the estimation of ψ.

In the appendix, we describe additional details for g-estimation in the presence of loss to follow up (as opposed to administrative censoring) and competing events. In both cases the steps above to estimate ψ must be slightly modified.

2.2.6 Pseudo-outcome regression

In this section, we describe how to use the estimate of ψ to construct treatment-naive pseudo-outcomes that can then be regressed on the relevant set of clinical predictors using any standard prediction algorithm. Recall, that during the g-estimation procedure we defined treatment-free outcome $H(\psi)$ which could be computed for any value of ψ, including our g-estimate of the true value $\widehat{\psi}$

$$H(\widehat{\psi}) = \int_0^T \exp\{\widehat{\psi}A(t)\}dt$$

Because g-estimators are consistent estimators of the true value and, under our model, $H(\psi)$ evaluated at the true value is the counterfactual failure time under no treatment, it follows that we can replace $T^{a=0}$ with $H(\widehat{\psi})$ in a regression to estimate the treament-naive risk conditional on covariates X^*

$$\Pr(T^{a=0} \leq t \mid X^*) = \Pr_n\{H(\widehat{\psi}) \leq t \mid X^*\}$$

More concretely, if for a particular individual their observed failure time is $T = 8$ but they were treated starting from time four and our estimate is $\widehat{\psi} = 0.5$ then the corresponding pseudo-outcome is

$$\begin{aligned} H(\widehat{\psi}) &= \int_0^T \exp\{-0.5A(t)\}dt \\ &= \int_0^4 \exp(-0.5 \cdot 0)dt + \int_4^T \exp(-0.5)dt \\ &= 4 + 4\exp(-0.5) \\ &\approx 6.43 \end{aligned}$$

implying their treatment-naive failure time is 6.43. This pseudo-outcome can then be used in place of their observed T when modeling the treatment-naive risk using any standard algorithm, for instance in a Cox proportional hazards model.

One of the advantages of this procedure is that, by using treatment-naive pseudo-outcomes, the causal inference and predictive portions of the task are effectively separated. Under the assumptions outlined, this allows analysts to use their preferred algorithm for the prediction task as if they had the true counterfactual values for every individual. To review the complete set of steps are:

1. Define the target trial corresponding to treatment decision of interest.
2. Estimate the parameter(s), ψ, of the SNAFTM using g-estimation.
3. Using the estimated $\widehat{\psi}$, construct pseudo-outcomes $H(\widehat{\psi})$.
4. Estimate $\Pr_n\{H(\widehat{\psi}) \leq t \mid X^*\}$ using any prediction algorithm.

When administrative censoring is present, $H(\widehat{\psi})$ may be replaced by $Z(\widehat{\psi})$ without loss of generality.

2.2.7 Inverse probability of censoring weighting

A limitation of g-estimation of SNATFM is that they require correct specification of the structural model for the removal of treatment. In particular, the analyst must specify any time-varying effect modification by past treatment and covariate history. In priniciple, this could be avoided by using saturated SNATFM with terms for all possible effect modifiers; however, in practice g-estimation of multi-parameter SNATFMs has proven difficult [62, 65]. An alternative approach is to use inverse probability of censoring weighting (IPCW). Returning to our conceptualization of a single-arm target trial in which the regime of interest is to withhold treatment at all time points, starting treatment can be viewed as a form of non-adherence. As in a trial, we can censor participants when they deviate from their "assigned" regime and use inverse probability weighting to adjust for time-varying non-adherence. Practically, this suggests starting with a treatment-naive population who meet the eligibility criteria relevant to the clinical decision in question, censoring participants when they initiate treatment, and then constructing the following stabilized weights

$$W_c = \prod_{k=0}^{K} \frac{I(A_k = 0)\Pr(A_k = 0 \mid X^*, \overline{A}_{k-1} = 0, T > k)}{\Pr(A_k = 0 \mid \overline{X}_k, \overline{A}_{k-1} = 0, T > k)}$$

where the denominator, $\Pr(A_k = 0 \mid \overline{X}_k, \overline{A}_{k-1} = 0, T > k)$, may be estimated from a model for treatment initiation among those who remain treatment free conditional on past covariate history and the numerator, $\Pr(A_k = 0 \mid X^*, \overline{A}_{k-1} = 0, T > k)$ may be estimated from a similar model excluding time-varying covariate history. Unlike traditional adherence adjustment, the numerator and denominator must be conditional on covariates X^* for them to be used in the resulting risk prediction model. Under conditions 1, 2, and 3b in section 2.2.3, the treatment-naive risk is identified and can be estimated using inverse probability weights W_c. Unlike g-estimation of SNAFTM, this approach requires a positive probability of nontreatment at each time point for all untreated individuals. This would be violated if, for instance, within strata of past treatment and covariate history, there were a subset of participants who always receive treatment.

In theory, once estimated, the weights W_c can be used by any prediction algorithm which permits time-varying weighted optimization to predict the treatment-naive risk. Effectively they create a pseudo-population in which treatment is withheld at all time points. As an example, the W_c could be used to fit the pooled logistic regression model

$$\text{logit}\{\Pr(Y_k = 1 \mid X^*, \overline{Y}_{k-1} = 0)\} = \theta_0(k) + \theta_1' X^*$$

using weighted maximum likelihood, which for flexible $\theta_0(k)$ and sufficiently small time steps approximates the Cox proportional hazards model [66].

2.2.8 Sensitivity analysis

In an observational setting, where treatment initiation over the follow up is not strictly controlled by the investigator the exchangeability assumption is likely violated but we often proceed as if it's at least approximately true given a “plausible” set of covariates. However, in the presence of unmeasured confounding, we can conduct a sensitivity analysis as follows. Suppose that the amount of unmeasured confounding were known in the sense that the degree of dependence between $T^{\bar{a}=0}$ and the conditional probability of treatment were known on the log-odds scale. Then we could solve for the parameter ω and function $q(k, \overline{X}_k, \overline{A}_{k-1}, T^{\bar{a}=0})$ in the logistic regression

$$\text{logit}[\Pr\{A_k = a_k \mid \overline{X}_k, \overline{A}_{k-1}, T > k, T^{\bar{a}=0}\}] = \theta_0 + \theta_1' \overline{X}_k + \theta_2' \overline{A}_{k-1} + \omega q(k, \overline{X}_k, \overline{A}_{k-1}, T^{\bar{a}=0})$$

where, if there were no unmeasured confounding, $q(k, \overline{X}_k, \overline{A}_{k-1}, T^{\bar{a}=0}) = 0$. Since ω and $q(k, \overline{X}_k, \overline{A}_{k-1}, T^{\bar{a}=0})$ are unknown, we could instead vary them over a plausible range of values and functional forms and examine the influence of unmeasured confounding on our resulting counterfactual prediction models. That is, for each value of ω and $q(k, \overline{X}_k, \overline{A}_{k-1}, T^{\bar{a}=0})$ we re-estimate the ψ parameters of the SNAFTM and form the pseudo-outcomes or, alternatively, re-estimate the inverse probability of censoring weights, and use the resulting pseudo-outcomes or weights with the desired prediction algorithm. This sensitivity analysis builds on that suggested in [67] for estimating causal effects.

2.3 Application

2.3.1 Study design and data

The Multi-Ethnic Study on Atherosclerosis (MESA) study is a population-based sample of 6,814 men and women aged 45 to 84 drawn from six communities (Baltimore; Chicago; Forsyth County, North Carolina; Los Angeles; New York; and St. Paul, Minnesota) in the United States between 2000 and 2002. The sampling procedure, design, and methods of the study have been described previously [68]. Study teams conducted five examination visits between 2000 and 2011 in 18 to 24 month intervals focused on the prevalence, correlates, and progression of subclinical cardiovascular disease. These examinations included assessments of lipid-lowering (primarily statins) and other medication use as well as cardiovascular risk factors such as systolic blood pressure, serum cholesterol, cigarette smoking, height, weight, and diabetes.

Our goal was to emulate a single-arm trial corresponding to the AHA guidelines on initiation of statin therapy for primary prevention of cardiovascular disease in the MESA cohort and use the emulated trial to develop a prediction model for the treatment-naive risk. The AHA guidelines stipulate that patients aged 40 to 75 with serum LDL cholesterol levels between 70 mg/dL and 190 mg/dL and no history of cardiovascular disease should initiate statins if their risk exceeds 7.5%. Therefore, we considered MESA participants who completed the baseline examination, had no recent history of statin use, no history of cardiovascular disease, and who met the criteria described in the guidelines (excluding the risk threshold) as eligible to participate in the trial. The primary endpoint was time to atherosclerotic cardiovascular disease (ASCVD), defined as nonfatal myocardial infarction, coronary heart disease death, or ischemic stroke.

Follow up began at the second examination cycle to enable a "wash out" period for statin use and to ensure adequate pre-treatment covariates to control confouding. We constructed a sequence of nested trials starting at each examination cycle from exam 2 through exam 5 and pooled the results from all 4 trials into a single analysis and used a robust variance estimator to account for correlation among duplicated participants. In each nested trial, we used the corresponding questionnaire to determine eligibility as well as statin initiators versus non-initiators. Because the exact timing of statin initiation was not known with precision, in each trial, we estimated the start of follow up for initiators and non-initators by drawing a random month between their current and previous examinations. We explored alternative definitions of the start of follow up in sensitivity analyses in the appendix. To mimic the targeted single-arm trial we limited to non-initiators for development of the prediction models.

Our trial emulation is only as good as the strong assumptions which underpin it. While these assumptions cannot be evaluated empirically, we performed a benchmarking exercise to determine whether gross violations of the assumptions were likely. As described in detail in the appendix, we emulated a nested sequence of two-arm trials comparing statin therapy to no therapy and compared the estimated intention to treat and adherence adjusted effects to those from existing randomized trials. Because statins have been extensively evaluated, the range of effect estimates that are "plausible" is available. The estimated hazard ratios for the ITT effects from our trial emulation in MESA fell within range of estimates from meta-analyses (HR = 0.79 vs. HR = 0.75). While these results don't imply that the assumptions required to emulate our single-arm trial for prediction are valid, they at least seem consistent with our prior expectations. Based on this result, we felt confident

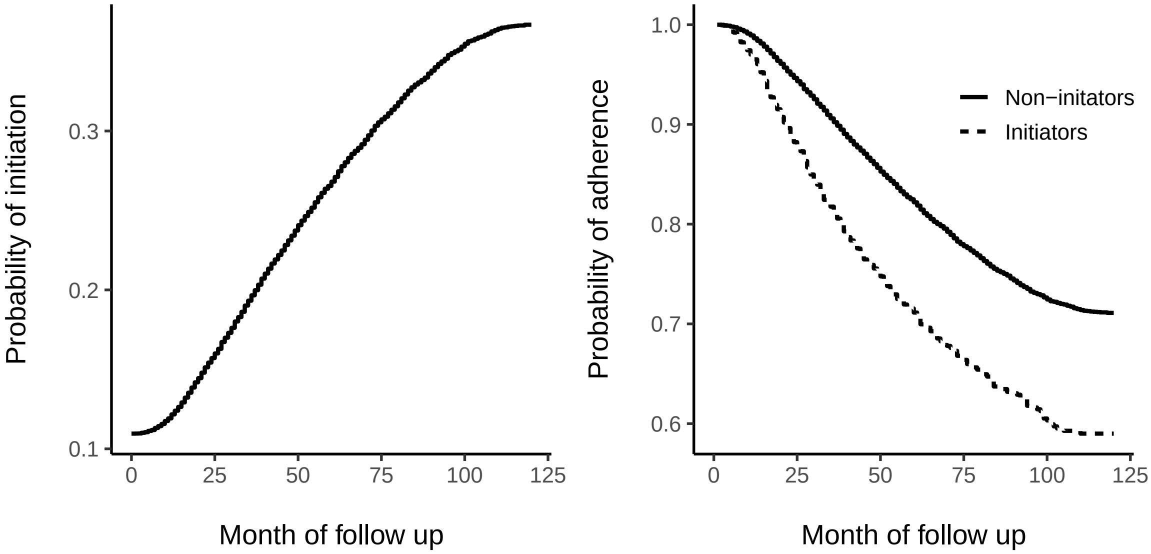


Figure 2.2: *Probability of statin initiation and probability of adherence among initiators and non-initiators in nested target trial emulation, Multi-Ethnic Study of Atherosclerosis, 2000 to 2010.*

proceeding with the development of a model for the statin-naive risk in this setting.

2.3.2 Predicting statin-naive risk

Of the 6,814 MESA participants who completed the baseline examination, 4,149 met the eligibility criteria for our trial emulation. There were 288 ASCVD events, 190 non-ASCVD deaths, and 414 were lost over the 10 year follow up period. In the nested trial dataset, there were 1,592 initiators and 12,767 non-initiators. Table 1 shows the baseline characteristics of initiators and non-initiators of statins in the emulated nested trials. Figure B.2 shows the probability of statin initation over the follow up period. After ten years approximately 40% of MESA participants had initiated statins.

To illustrate the proposed methods, we created prediction models for the statin-naive risk in the emulated single arm trial data based on both g-estimation of SNAFTMs and inverse probability of censoring weighting. For the prediction task, we selected baseline predictors commonly used in development of models for cardiovascular disease including age, sex, smoking status, diabetes history, systolic blood pressure, anti-hypertensive medication use

Table 2.1: *Baseline characteristics of initiators and non-initiators in emulated nested trials*

	Initiators (N = 1,592)	**Non-initiators (N = 12,767)**
Demographics		
Age, years	65.1 (8.2)	62.5 (8.8)
Male, %	45.1	46.8
Married, %	64.6	63.3
Less than high school, %	18.2	15.0
High school graduate, %	48.2	44.9
College or postgraduate, %	33.5	39.8
Non-Hispanic white, %	40.1	37.9
Non-Hispanic black, %	22.9	21.9
Hispanic, %	26.1	27.4
Asian, %	11.0	12.9
Currently employed, %	51.4	59.2
Retired, %	33.3	26.8
No health insurance, %	3.6	8.2
CES Depression scale (0-60)	7.4 (7.5)	7.5 (7.5)
Chronic burden scale (0-5)	1.1 (1.2)	1.1 (1.2)
Perceived discrimination scale (0-4)	0.1 (0.4)	0.1 (0.4)
Emotional support scale (0-30)	24.3 (5.1)	24.1 (5.3)
Everyday hassles scale (0-54)	14.4 (6.0)	15.2 (6.2)
Spielberger trait anger scale (0-40)	15.0 (3.8)	15.0 (3.7)
Spielberger trait anxiety scale (0-40)	15.9 (4.5)	16.0 (4.5)
Neighborhood problems scale (0-28)	10.4 (3.4)	10.5 (3.4)
CVD risk factors		
Systolic blood pressure, mmHg	125.6 (19.7)	122.0 (20.2)
Diastolic bood pressure, mmHg	71.3 (10.2)	71.1 (10.1)
LDL cholesterol, mg/dL	135.4 (31.7)	119.7 (27.6)
HDL cholesterol, mg/dL	50.2 (13.9)	52.3 (15.1)
Triglycerides, mg/dL	147.5 (87.7)	120.5 (66.1)
Baseline ASCVD risk, %	10.1	7.6
Diabetes mellitus, %	38.4	23.0
Hypertension, %	51.9	36.2
Waist circumference, cm	99.6 (14.3)	96.9 (14.7)
Smoked <100 cigarettes in lifetime, %	49.6	48.8
Current smoker, %	10.7	12.9
Drinks per week	2.9 (5.9)	3.4 (7.6)
Exercise, MET/min	1471.6 (2187.1)	1509.3 (2187.7)
Family history of CVD, %	58.9	53.5
Calcium score	124.2 (340.7)	72.4 (256.2)
Left ventricular hypertrophy on ECG, %	1.0	0.8
C-reactive protein, mg/dL	4.3 (6.0)	3.5 (5.0)
Interleukin-6, pg/mL	1.5 (1.2)	1.4 (1.2)
Number of pregnancies	3.1 (2.2)	3.0 (2.3)
Years on birth control pills	3.6 (5.8)	3.7 (5.8)
Age at menopause, years	41.1 (17.5)	37.2 (20.4)
Medications		
Anti-hypertensive medication, %	61.7	34.7
Insulin or oral hypoglycemics, %	22.7	8.6
Daily aspirin use, %	47.2	25.0
Diuretics, %	21.4	12.5
Any anti-depressants, %	11.7	7.8
Any vasodilator, %	3.8	3.3
Any anti-arrhytmic, %	0.6	0.6

and total and HDL serum cholesterol levels.

To build a model for the statin-naive risk using g-estimation, we estimated the SNAFTM parameter $\widehat{\psi}$ for the effect of an instantaneous blip of statin treatment in the full trial population. The estimated value, $\exp(-0.28) = 0.76$, suggests, on average, removing treatment from initiators reduces time to ASCVD by 24%. Next, using our estimates $\widehat{\psi}$ we formed statin-naive pseudo-outcomes $H(\widehat{\psi})$ based on the corresponding SNAFTM and regressed them on our predictors of interest using a Cox proportional hazards model to create a model for the statin-naive risk (Table 2.2). Table **??** in the appendix also includes estimates from SNAFTMs that allow for time-varying effect modification by baseline risk (column 2) and serum LDL cholesterol level (column 3).

We constructed a second model for the statin-naive risk using inverse probability of censoring weights. To calculate the weights, we estimated two pooled logistic regression models: one for the probability of remaining untreated given past covariate history (denominator model) and one for probability of remaining untreated given the selected baseline predictors (numerator model). The mean of the stabilized weights was 1.02. To create a statin-naive prediction model, we used the estimated weights to fit a weighted pooled logistic regression model conditional on the baseline predictors of interest. As mentioned previously, this model approximates a Cox proportional hazards model for sufficiently small time steps and flexible specification of the baseline hazard (we used restricted cubic splines).

For comparison, we also fit a traditional (factual) prediction model by regressing the observed failure times on the same set of baseline predictors, but ignoring treatment initiation over the follow up period. As mentioned previously, this approach targets the natural course risk rather than the statin-naive risk. Model coefficients, p-values, and confidence intervals for both factual and counterfactual prediction models are shown in Table 4. In general, associations between baseline predictors and ASCVD were stronger in the counterfactual models of the statin-naive risk. This makes sense as most predictors in the model are risk factors for cardiovascular disease that are also used to determine eligibility for statins. Therefore, in the absence of statins, we'd expect risk factors associations with

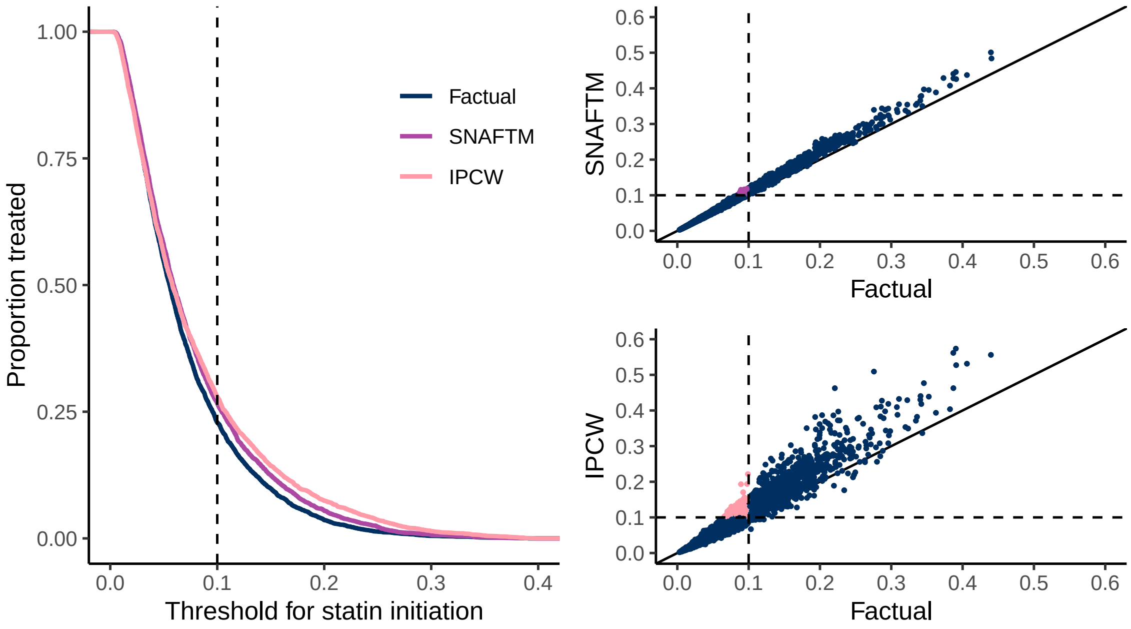


Figure 2.3: *Comparison of predictions using Factual, SNAFTM, and IPCW models.*

cardiovascular disease to be stronger as they are no longer partially moderated by statin use. To determine how the choice of model may affect clinical guidance, in Figure 2.3 we examine the proportion recommended for treatment under natural course and statin-naive prediction models at different risk thresholds. At the 7.5% threshold 53% would be eligible for statins using the statin-naive model compared to 48% using the natural course model, implying for every 1000 patients screened about 50 who would be eligible under the statin-naive model would not be recommended statins using traditional methods. At the 10% threshold this increases to nearly 60 out of every 1000 patients screened.

2.4 Discussion

Treatment initiation over follow up is common in many risk prediction settings, especially when the time horizon for predictions is long. This causes issues particularly when model estimates are used to determine eligibility for the very same treatment. We argue that in this case clinical decision-making is best informed by the treatment-naive risk, i.e. the counterfactual risk if no one were treated. However, this risk requires additional untestable assumptions and appropriate causal inference methods to estimate. In this paper, we proposed methods for building a model for the treatment-naive risk based on g-estimation of SNAFTMs and inverse probability of censoring weighting. We also applied the target trial framework to better clarify the clinical decision that is to be informed by the model and to guide analytic choices for emulating the trial in observational data. We used these methods to emulate a trial corresponding to the AHA guidelines on risk-based treatment eligibility for statins and then develop models for the statin-naive risk. Compared to traditional approaches which don't account for treatment initiation, statin-naive models produced estimates that were systematically higher and suggested that traditional approaches may lead to underallocation of treatment among those eligible.

Our work builds on a growing literature at the intersection of the "two cultures" of statistical modeling: prediction and causal inference [5, 6]. While one might be tempted to prefer a clean separation between the two, like previous approaches, we emphasize that many important prediction questions can be recast as counterfactual questions, especially when data are imperfect, clinical settings change, or predictions under hypothetical interventions are required. In practice, these situations are quite common. However, the strength of assumptions required may exceed our ability to effectively answer them using the data assembled.

For example, both the g-estimation and IPCW approaches suggested here to estimate the counterfactual treatment-naive risk require sequential exchangeability of treatment intitiation. Only subject matter expertise can inform whether this assumption is plausible in a given analysis. Therefore, it must be evaluated on a case-by-case basis. In general, counterfactual prediction will likely require higher quality data than is routinely collected for prediction including rich data on time-varying predictors of treatment initiation.

An alternative for well studied treatments might be to use treatment effect estimates from randomized trials. For instance if we had "trusted" estimates of ψ based on a meta-analysis of RCTs, we could forgo steps 1 and 2 in section 2.2.6 and constuct treatment naive pseudo-outcomes under the SNAFTM and then a treatment-naive model could be created using the pseudo-outcomes. This is similar to the approach taken in [57]. However, this exchanges one set of untestable assumptions for another, in this case, the assumption that treatment initiation within the observational study is unconfounded is exchanged for the assumption that the treatment effects from RCTs are transportable to the observational setting [69, 70]. In particular, the latter requires that either treatment effects are constant or the distribution of effect modifiers is the same between populations and all effect modification by time-varying covariates is properly modeled in the SNAFTM. Attempts to transport results from RCT are also hampered by the fact that, unless compliance with assigned treatment is 100%, effect estimates are generally intent to treat and therefore may understate the effect of removing treatment even whe effects are constant.

In the prediction of the treatment naive risk, the positivity assumption may require more thoughtful evaluation than traditionally appreciated, as there is often imperfect overlap between the population used to train the model and the population eligible for treatment. Furthermore, there may be subsets for whom treatment is always recommended. In some instances conceptualizing the single-armed target trial of interest may be helpful here.

Througout this paper, our focus has been on estimating the treatment-naive risk specifically, as opposed to other possible counterfactual estimands, and our choice of methods reflect the desire to make minimal assumptions. Both g-estimation and IPCW require fewer assumptions than alternatives such as marginal structural models or the g-formula that model risk under all possible regimes. However, they may be less efficient choices when the additional assumptions are valid and are less flexible in their ability to target other estimands of interest. Where possible we have also focused on approaches that are agnostic to the final prediction algorithm chosen, allowing the analyst to chose the most suitable to the task from among a proven toolkit of prediction algorithms.

The g-estimation procedure for estimating the parameter of a SNAFTM has a number of known limitations in the presence of administrative censoring. Because the estimating function is non-smooth the algorithm may fail to converge or multiple optimal solutions may be possible. This is most likely to occur when the number of observed failure times is low and the blip function $\gamma(t, \overline{A}_t, \overline{X}_t; \psi)$ is complex. Alternatives such as structural nested cumulative failure time models (SNCFTM) [71] and structural nested cumulative survival time models (SNCSTM) [72] have been suggested previously. Under certain parameterizations of the blip function they could still be used to generate pseudo-outcomes for use in the approach described here.

We have taken for granted that treatment decisions should be informed by risk under no treatment. However, other rules for allocating treatment are possible such as modeling the conditional average treatment effect [73–75] and assigning treatment to anyone with net benefit (i.e. estimated treatment effect minus costs or side effects). Indeed, a major limitation of a risk-based approach is that once treatment coverage is close to 100%, i.e. when nearly everyone who becomes eligible initiates treatment immediately, it's unclear how to continue to update a treatment-naive model without strong assumptions such as assuming that effects are stable over time.

Finally, we have ignored other possible uses of prediction modeling such as public health planning, where a descriptive summary of the risk across different groups is needed and treatment initiation over follow up is not an issue (in fact it's desirable to know risk under the natural course). Ideally, the same models shouldn't be used to serve dual roles without proper adjustment.

Table 2.2: *Cox proportional hazards model using observed survival times compared with counterfactual models (SNAFTM and IPCW).*

	Factual			Counterfactual (SNAFTM)			Counterfactual (IPCW)		
Characteristic	**HR[1]**	**95% CI[1]**	**p-value**	**HR[1]**	**95% CI[1]**	**p-value**	**HR[1]**	**95% CI[1]**	**p-value**
bl_age	1.27	(1.18, 1.37)	<0.001	1.28	(1.19, 1.38)	<0.001	1.20	(1.11, 1.30)	<0.001
gender	1.64	(1.27, 2.13)	<0.001	1.66	(1.28, 2.15)	<0.001	1.59	(1.21, 2.11)	0.001
bl_cursmk	1.86	(1.41, 2.46)	<0.001	1.86	(1.41, 2.46)	<0.001	1.62	(1.19, 2.16)	0.002
bl_dm03	1.28	(1.00, 1.63)	0.051	1.32	(1.03, 1.69)	0.026	1.52	(1.17, 1.98)	0.002
bl_sbp	1.25	(1.15, 1.36)	<0.001	1.25	(1.15, 1.36)	<0.001	1.27	(1.16, 1.39)	<0.001
bl_hdl	0.81	(0.73, 0.89)	<0.001	0.79	(0.72, 0.87)	<0.001	0.75	(0.67, 0.84)	<0.001
bl_chol	1.03	(1.00, 1.06)	0.034	1.05	(1.02, 1.08)	<0.001	1.09	(1.06, 1.13)	<0.001
bl_htnmed	1.35	(1.04, 1.74)	0.025	1.47	(1.13, 1.90)	0.004	1.57	(1.16, 2.11)	0.003
bl_sbp * bl_htnmed	0.83	(0.75, 0.93)	0.002	0.83	(0.74, 0.93)	0.001	0.88	(0.78, 0.99)	0.039

[1] HR = Hazard Ratio, CI = Confidence Interval

Chapter 3

Validating counterfactual predictions[1]

Counterfactual prediction methods may be required when treatment policies differ between model training and deployment settings or when the prediction target is explicity counterfactual. However, validating counterfactual predictions is challenging as typically one does not observe the full set of potential outcomes for all individuals. We consider methods for validating a prediction model under counterfactual shifts in treatment policy. We discuss how to tailor a model for use in the same population under a counterfactual shift in treatment, how to assess its performance, and how to perform model and tuning parameter selection. We also provide identifiability results for measures of counterfactual performance for a potentially misspecified prediction model based on training and test data from the (factual) source population only. We illustrate the methods using simulation and apply them to the task of developing a statin-naive risk prediction model for cardiovascular disease.

3.1 Introduction

Prediction models are often deployed in settings that are different from those in which they are trained. One of the ways settings may differ is that the natural course of treatment after

[1] Co-authored with James Robins, Andrew Beam, and Goodarz Danaei

baseline may vary, particularly for models with a longer time horizon [11]. For example, a prediction model fit in a population where 5% are treated over the follow up period may not produce valid predictions in one where 50% are treated and vice versa. Even when models are deployed in the same population, treatment policies may change over time, affecting who is likely to be treated and leading to problems of "domain adaption" or "dataset shift" [9, 55]. These differences between the training and deployment environments can cause the performance of models to degrade, particularly when, as is often the case, model predictors are themselves correlated with, or direct determinants of, treatment [36].

Ideally, when faced with such a change in the treatment environment one would simply re-train the model. However, collecting the necessary data in the new setting may be inordinately expensive or time consuming. Absent sufficient resources or as a stop gap, one might consider tailoring the original model to target the expected outcome that would be observed were treatment administered to everyone as in the deployment setting but using only training data. Alternatively, one might simply wish to estimate how poorly the existing model is likely to perform in the deployment setting, to determine whether data collection efforts are worthwhile. In either case, the implicit inquiries are counterfactual.

Beyond accounting for descrepancies between training and deployment, there are also instances in which the target prediction estimand is explicitly counterfactual. For instance, a model may be used to inform clinical decisions about whether to initiate treatment or to compare outcomes under alternative treatment strategies [16, 76, 77]. This could involve risk-based rules for treatment adoption or the transportation or direct estimation of treatment effects. In some circumstances, models may be built and evaluated without explicit appeal to counterfactuals, such as when effects are modeled in a randomized trial and used in the same population. However, as often is the case, when training data are obtained in an observational setting where treatment initiation over follow up is not strictly controlled by the investigator, the predictions most relevant to decision-making are counterfactual [15, 77].

In both instances, we need methods for tailoring models to target counterfactual queries, even when data on the full set of potential outcomes is not available. We also need performance metrics that agnostically evaluate model performance in these new environments independent of whether the prediction model itself is correctly specified. In this paper, we examine the conditions under which tailoring a prediction model to counterfactual outcomes is possible using training data alone. Under similar conditions, we also show that the counterfactual performance of the model may be estimated independently from the method used to fit the model and may be evaluated even if the model is misspecified or does not target the counterfactual estimand directly. This is a key result as it implies the counterfactual performance of a model can be identified and estimated even for models that are "wrong". Absent better data or in the meantime while such data are being collected, performance metrics may therefore be used to differentiate between better and worse-performing models or to quantify how badly a model is likely to perform in a hypothetical environment.

3.2 Set up and notation

Let Y be the outcome of interest, X a baseline covariate vector, and A an indicator of treatment over the follow up period. We assume all are obtained via a simple random sample from a population $\{(X_i, A_i, Y_i)\}_{i=1}^{n}$ in which the initiation of treatment follows it's natural course. Covariates in X include possible predictors of the outcome which do not act as confounders (P), as well as joint determinants of the outcome and treatment which act as confounders (L). We would like to build a prediction model for Y using covariates X^* which are a subset of X, i.e. $X^* \subset X$, and are chosen on the basis of availability and prediction potential rather than necessarily for their causal relationship to the outcome. To fix concepts, we assume for now A is a point treatment, i.e. that either treatment is always initiated immediately after baseline or there is no effect of duration of treatment on the outcome. However, we extend this to the case that treaments are time-varying in the appendix. We also assume for now that there is no loss to follow up. An example directed acyclic graph for this process is shown in Figure 3.1.

The data are randomly split into a training set and a test set with $n = n_{train} + n_{test}$. Let D_{train} and D_{test} be indicators of whether an observation is in the training set or test

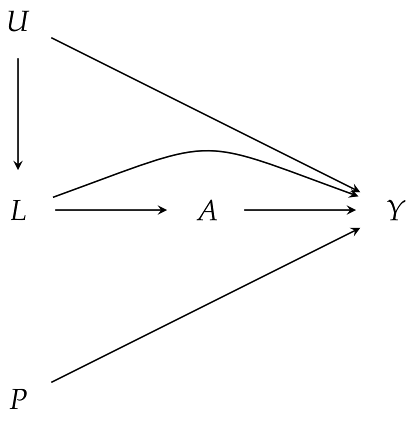


Figure 3.1: *Example directed acyclic graph for prediction in a setting with a single time fixed treatment A over follow up.*

set respectively. As is customary, we use the training set to build a prediction model for the expected outcome conditional on covariates $E[Y|X^*]$ and the test set to evaluate model performance. Let $\mu_\beta(X^*)$ be a parametric model indexed by parameter β and $\mu_{\widehat{\beta}}(X^*)$ be the "fitted" model using parameter estimates $\widehat{\beta}$. We allow for the possibility that model $\mu_\beta(X^*)$ is misspecified. For a particular estimand such as $E[Y|X^*]$, a model is correctly specified if there exists $\beta_0 \in \mathcal{B}$, where $\mathcal{B}$ is the parameter space of β, such that $\mu_{\beta_0}(X^*) = E[Y|X^*]$ and the model is misspecified if no such β_0 exists. In several places, we use $f(\cdot)$ generically to denote a density.

To define counterfactual estimands of interest, let Y^a be the potential outcome under an intervention which sets treatment A to a. To keep our notation simple, here we limit our focus to so-called *static* and *deterministic* interventions, in which the potential outcome desired is the outcome under a fixed value of A, but extend to *random* and *dynamic* regimes, such as those mentioned in the introduction, in the appendix.

3.3 Training and performance targets

Our goal is to make and assess predictions in a counterfactual version of the source population in which treatment policies differ, for instance if no one in source population took treatment, if everyone did, or if specific guidelines changed. To make predictions, we posit a parametric model $\mu_\beta(X^*)$ for the expected potential outcome conditional on covariates $E[Y^a|X^*]$, which we wish to estimate from the training dataset. The model may be tailored to the counterfactual outcome Y^a, in the sense that it was trained to target $E[Y^a|X^*]$ directly, or it may be a model for another target such as the expected (factual) outcome in the source population $E[Y|X^*]$ and we would like to know how it might perform in a counterfactual setting.

To determine the performance of the model, one generally relates its fitted predictions $\mu_{\widehat{\beta}}(X^*)$ to the observed outcomes Y^a using any of a number of metrics from the prediction literature [1, 21, 78]. However, for counterfactual predictions, this is not as simple as the potential outcome Y^a is not observed for all individuals. Yet, as we will show, under certain conditions the expected value of the metric may still be identified from the observed data in the test set. An example target performance metric of interest is

$$\psi = E[(Y^a - \mu_{\widehat{\beta}}(X^*))^2]$$

where the squared error loss $(Y^a - \mu_{\widehat{\beta}}(X^*))^2$ quantifies the discrepancy between the potential outcome under treatment level $A = a$ and the model prediction $\mu_{\widehat{\beta}}(X^*)$ in terms of the squared difference. In the main text, we focus on the mean squared error as the metric ψ for assessing performance of the model. However, in the appendix we extend our results to that case that ψ is any member of a generic class of loss functions $L(Y^a, \mu_{\widehat{\beta}}(X^*))$ as well as common metrics such as model discrimination and risk calibration. Importantly, ψ is always defined without assuming $\mu_{\widehat{\beta}}(X^*)$ is correctly specified.

3.4 Identifiability conditions

We will assume the following identifiability assumptions which have been described in more detail elsewhere [12, 22, 79].

1. *Exchangeability.* $Y^a \perp\!\!\!\perp A \mid X$

2. *Consistency.* $Y^a = Y$ if $A = a$

3. *Positivity.* For all x, $\Pr(A = a \mid X = x) > 0$

The first condition stipulates that treatment initiation over follow up is conditionally independent of the potential outcome given covariates X. This would be ensured by design in a randomized trial in which, participants are randomized to treatment or no treatment conditional covariates X. The second condition implies that observed outcomes among those with $A = a$ reflect potential outcomes under corresponding level of treatment. It would be violated if, for instance, there were multiple hidden versions of the therapy under consideration. Finally, the third positivity condition implies that there is a positive probability of observed treatment level $A = a$ in all strata of X.

3.5 Tailoring a model for counterfactual predictions

As we show in section C.1.1 of the appendix, under the conditions above the expected potential outcome conditional on covariates X^* is identified by the expression

$$E[Y^a \mid X^*] = E[E[Y \mid X, A = a, D_{train} = 1] \mid X^*, D_{train} = 1] \tag{3.1}$$

or, equivalently

$$E[Y^a \mid X^*] = E\left[\frac{I(A = a)}{\Pr(A = a \mid X, D_{train} = 1)} Y \mid X^*, D_{train} = 1\right] \tag{3.2}$$

in the training dataset. Both suggest possible targets for tailoring the model for counterfactual predictions using only the training data.

For simplicity, assume for a moment that $X = X^*$, that is the predictors included in the model are also those necessary to ensure exchangeability. Note that in this case the right hand side of equation 3.1 above reduces to $E[Y \mid X, A = a, D_{train} = 1]$ which suggests tailoring the model for the counterfactual prediction target $E[Y^a \mid X]$ using the training data could be accomplished by subsetting to participants with corresponding treatment level $A = a$ and fitting model $\mu_\beta(X)$ for the observed Y conditional X. Such a model will be consistent for $E[Y^a \mid X]$ provided it is correctly specified. Generally though there will not be perfect overlap between the covariates necessary to ensure exchangeability and those available for prediction. When X^* is a subset of X, tailoring a model for counterfactual prediction will require some method of marginalizing over the covariates in X that are not in X^*, either analytically or using Monte Carlo methods. This will also generally be true for random and dynamic interventions.

Under the same identifiability conditions, equation 3.2 suggests an alternative approach to targeting $E[Y^a \mid X]$ using the training data is to fit a weighted model $\mu_\beta(X^*)$, using for instance weighted maximum likelihood, with weights equal to the probability of receiving treatment level $A = a$ conditional on covariates X necessary to ensure exchangeability, i.e. $W = \frac{I(A=a)}{\Pr(A=a|X,D_{train}=1)}$. This is the basis for several previously proposed methods for counterfactual prediction based on inverse probability of treatment weighting [17]. Note that, unlike the first approach, it is possible to specify a subset of predictors X^* used in the prediction model $\mu_\beta(X^*)$ as compared to the full set of covariates X required for exchangeability which are only necessary for defining the weights W. This means tailoring the model for counterfactual predictions using this approach can be accomplished using off-the-shelf software.

3.6 Assessing model performance

Using the same conditions, in section C.1.2 of the appendix we show the model performance metric ψ is identifiable using data from the test set through the expression

$$\psi = E\left[E[(Y - \mu_{\widehat{\beta}}(X^*))^2 \mid X, A = a, D_{test} = 1] \mid D_{test} = 1\right] \tag{3.3}$$

or, equivalently using an inverse probability weighted expression

$$\psi = E\left[\frac{I(A = a)}{\Pr(A = a \mid X, D_{test} = 1)}(Y - \mu_{\widehat{\beta}}(X^*))^2 \mid D_{test} = 1\right] \tag{3.4}$$

regardless of whether the model $\mu_{\widehat{\beta}}(X^*)$ has been tailored to target $E[Y^a \mid X]$ or is correctly specified in general. As previously the two expression suggest two different approaches for the estimation of model performance using the test data alone.

First, using the sample analog of expression (3.3), an estimator of the target MSE is

$$\widehat{\psi}_{CL} = \frac{1}{n_{test}} \sum_{i=1}^{n} I(D_{test,i} = 1)\widehat{h}_a(X_i) \tag{3.5}$$

where $\widehat{h}_a(X)$ is an estimator for the conditional loss $E[(Y - \mu_{\widehat{\beta}}(X^*))^2 \mid X, A = a, D_{test} = 1]$. To keep notation simple, we supress the dependency of $\widehat{h}_a(X)$ on $\mu_{\widehat{\beta}}$. When the dimension of X is small it may be possible to use the sample analog of $\widehat{h}_a(X)$ as an estimator as well. In practice, though, some form of modeling will often be required. In this case, $\widehat{\psi}_{CL}$ is a consistent estimator for ψ as long as $\widehat{h}_a(X)$ is correctly specified.

Next, using the sample analog of expression (3.4), an alternative weight-based estimator of the target MSE is

$$\widehat{\psi}_{IPW} = \frac{1}{n_{test}} \sum_{i=1}^{n} \frac{I(A_i = a, D_{test,i} = 1)}{\widehat{e}_a(X_i)} (Y_i - \mu_{\widehat{\beta}}(X_i^*))^2 \tag{3.6}$$

where $\widehat{e}_a(X)$ is an estimator of the probability of receiving treatment level $A = a$ conditional on X, i.e. $\Pr(A = a \mid X, D_{test} = 1)$. Again, when the dimension of X is small it may be possible to use the sample analog of $\widehat{e}_a(X)$ as an estimator, but in practice, it will have to be modeled. The weighting estimator $\widehat{\psi}_{IPW}$ is a consistent estimator of ψ as long as $\widehat{e}_a(X)$ is correctly specified.

The conditional loss estimator (3.5) relies on correctly specifying the model for the conditional loss and the weighting estimator (3.6) relies on correctly specifying the model for the probability of treatment. In some settings, one estimator may be preferred over the other when more is known about one process, such as when the algorithm for administering treatment is clearly defined. In practice though, both may be difficult to specify correctly. Using data-adaptive and more flexible machine learning estimators for estimation of these nuisance models offers the possibility of capturing arbitrarily complex data generation processes. However, these estimators generally have slower rates of convergence than the $\sqrt{n}$ rates of parametric models and therefore will not yield asymptotically valid confidence intervals [38]. An alternative is to use a doubly-robust estimator which combines models for $\widehat{h}_a(X)$ and $\widehat{e}_a(X)$, such as

$$\widehat{\psi}_{DR} = \frac{1}{n_{test}} \sum_{i=1}^{n} I(D_{test,i} = 1) \left[\widehat{h}_a(X_i) + \frac{I(A_i = a)}{\widehat{e}_a(X_i)} \left\{ (Y - \mu_{\widehat{\beta}}(X_i^*))^2 - \widehat{h}_a(X_i) \right\} \right] \tag{3.7}$$

As we show in the Appendix, under mild regularity conditions [80], this estimator will be consistent if one of $\widehat{h}_a(X)$ and $\widehat{e}_a(X)$ is correctly specified. They also permit the use of machine learning or data-adaptive estimators that are not $\sqrt{n}$-covergent allowing for more flexible estimation of the nuisance functions. This is due to the fact that the empirical process terms governing the convergence of $\widehat{\psi}_{DR}$ involve a product of the errors for $\widehat{h}_a(X)$ and $\widehat{e}_a(X)$ which converge under the weaker condition that only the *combined* rate of convergence for both nuisance functions is at least $\sqrt{n}$.

3.7 Model and tuning parameter selection

To this point, we have assumed that $\mu_\beta(X^*)$ is a pre-specified parametric model and ignored any form of model selection (e.g. variable or other specification search) or data-adaptive tuning parameter selection. However, in reality analysts often select between multiple models or perform a data-adaptive search through a parameter space for tuning parameter selection when developing a prediction model [1]. When done rigorously, analysts typically use methods such as cross-validation or the bootstrap to perform selection. These techniques rely on optimizing some measure of model performance, such as the MSE.

When performing model or tuning parameter selection for counterfactual prediction, the results from the previous sections suggest that the model performance measure should be targeted to the counterfactual performance in a population in which the hypothetical intervention were universally applied. For example, when using cross-validation for model selection the analyst splits the data into K mutually exclusive "folds" and estimates the candidate models using $K - 1$ of the folds and estimates the performance of each in the held out fold. This process is repeated K times where each fold is left out once. The final estimate of performance is the average of the K estimates and the model with best overall performance is selected (or, alternatively, the tuning parameter with the best performance).

When targeting counterfactual predictions, at each stage in the procedure the analyst should use modified performance measures such as those in section 3.6 above. Failure to do so, can lead to sub-optimal selection with respect to the counterfactual prediction of interest.

3.8 Simulation

In this section we perform two simulation experiments to illustrate (i) the benefits of tailoring models to the correct counterfactual estimand of interest, (ii) the potential for bias when using naive estimators of model performance such as the MSE, (iii) the importance of correct specification of the nuisance models when estimating counterfactual performance, and (iv) the properties of the doubly-robust estimator under misspecification of the nuisance models. We adapt data generation processes previously used for transporting models between settings under covariate shift [10, 81].

3.8.1 Experiment 1

We simulated treatment initiation over the follow up period based on the logistic model $\Pr(A = 1 \mid X) = \text{expit}(1.5 - 0.3X)$, where predictors X are drawn from $X \sim \text{Uniform } (0, 10)$. Under this model, about 50% initiate treatment over follow up but those with higher values of X are less likely to start treatment than those with lower values of X. We then simulated the outcome using the linear model $Y = 1 + X + 0.5X^2 - 3A + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, X)$. We set the total sample size to 1000 and the data were randomly split in a 1:1 ratio into a training and a test set. The full process may be written:

$$\begin{aligned}
&X \sim \text{Unif}(0, 10) \\
&A \sim \text{Bernoulli}\{\text{expit}(-1.5 + 0.3 \cdot X)\} \\
&Y \sim \text{Normal}(1 + X + 0.5X^2 - 3A, X)
\end{aligned}$$

Our goal was to estimate a model in a counterfactual population in which no one initiated treatment over follow up, i.e. we targeted $E[Y^{a=0} \mid X]$. Under this data generating mechanism, the MSE under no treatment is larger than the MSE under the natural course and identifiability conditions 1-3 are satisfied. We considered two specifications of prediction models $\mu_\beta(X^*)$:

1. a correctly specified linear regression model that included the main effects of X and X^2, i.e. $\mu_\beta(X^*) = \beta_0 + \beta_1 X + \beta_2 X^2$.

2. a misspecified linear regression model that only included the main effect of X, i.e. $\mu_\beta(X^*) = \beta_0 + \beta_1 X$.

For each specification, we also considered two estimation strategies: one using ordinary least squares regression (OLS) and ignoring treatment initiation and the other using weighted least squares regression (WLS) where the weights were equal to the inverse of the probability of being untreated. As discussed above the latter specifically targets the counterfactual estimand under no treatment. Finally, we considered two approaches for estimating the performance of the models in the test set: a naive estimate of the MSE using observed outcome values, i.e.

$$\widehat{\psi}_{Naive} = \frac{1}{n_{test}} \sum_{i=1}^{n} I(D_{test,i} = 1)(Y_i - \mu_{\widehat{\beta}}(X^*))^2,$$

and the inverse-probability weighted estimator $\widehat{\psi}_{IPW}$ from section 3.6. For the latter, we fit a correctly specified logistic regression model for $e_a(X)$, i.e. $e_a(X) = \text{expit}(\alpha_0 + \alpha_1 X)$, in the test set to estimate the weights. Lastly, we also calculated the true counterfactual MSE if we had access to the potential outcomes $Y^{a=0}$ by generating test data under same process as above but forcing $A = 0$ for everyone and then estimating counterfactual MSE and averaging across simulations.

Table 1 shows the results of the experiment based on 10,000 monte carlo simulations. In general, correctly specified models yielded smaller average MSE than misspecified models. Comparing the performance of OLS and WLS estimation, when using $\widehat{\psi}_{Naive}$, the naive estimator of the MSE, OLS seemed to produce better predictions than WLS when correctly specified (average MSE of 2.9 vs. 5.5) as well as when misspecified (average MSE of 16.8 vs.

Model $\mu_{\beta}(X)$	$\widehat{\psi}_{Naive}$	$\widehat{\psi}_{IPW}$	Truth
Correct			
OLS	2.9	3.6	3.6
WLS	5.5	1.0	1.0
Misspecified			
OLS	16.8	17.5	17.5
WLS	19.5	15.0	15.0

Correct and misspecified refers to the specification of the prediction model $\mu_{\beta}(X)$. OLS = model estimation using ordinary least squares regression (unweighted); WLS = model estimation using weighted least squares regression with weights equal to the inverse probability of being untreated. Results were averaged over 10,000 simulations. The true counterfactual MSE was obtained using numerical methods.

19.5). In contrast, when using $\widehat{\psi}_{IPW}$, the inverse-probability weighted estimate of the MSE, WLS performed better than OLS both when the model was correctly specified (average MSE of 1.0 vs. 3.6) and when misspecified (average MSE of 15.0 vs. 17.5). For reference, in the last column we show the true counterfactual MSE that would be obtained if one had access to the potential outcomes (obtained via numerical methods). We find that the average of the inverse probability weighted estimator across the simulations was equivalent to this quantity for all specifications and for both OLS and WLS estimation. This suggests that only the modified estimators of model performance in section 3.6 are able to accurately estimate the counterfactual performance of the model. Indeed, under this data generation process, if one were to use the naive estimator one might erroneously conclude that the OLS model is the better choice.

3.8.2 Experiment 2

In the previous experiment we assumed the nuisance models for the MSE were correctly specified. Here we consider estimation of the MSE in the more likely case that nuisance models are misspecified. This time, we simulated treatment initation over follow up A

based on the logistic model $\Pr[A = 1 \mid X] = \text{expit}\left(-0.3 + 0.2\sum_{i=1}^{3} X_{(i)} + 0.3\sum_{i=1}^{3}\left(X_{(i)}\right)^2\right)$, where X is now a vector of predictors drawn from a 10-dimensional mean zero multivariate normal and $X_{(i)}$ is the ith component of the vector X. This resulted in expected treatment initiation over follow up of 61%. We also simulated a binary outcome from a Bernoulli distribution with mean $\text{expit}\left(-0.3 + 0.2\sum_{i=1}^{3} X_{(i)} + 0.3\sum_{i=1}^{3}\left(X_{(i)}\right)^2 - 0.5A\right)$. Again, we set the total sample size to 1000, but this time we randomly split the data in a 2:1 ratio into a training and a test set.

$$\begin{aligned}
&X \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma}) \\
&A \sim \text{Bernoulli}\left\{\text{expit}\left(-0.3 + 0.2\sum_{i=1}^{3} X_{(i)} + 0.3\sum_{i=1}^{3} X_{(i)}^2\right)\right\} \\
&Y \sim \text{Bernoulli}\left\{\text{expit}\left(-0.3 + 0.2\sum_{i=1}^{3} X_{(i)} + 0.3\sum_{i=1}^{3} X_{(i)}^2 - 0.5A\right)\right\}
\end{aligned}$$

Our prediction model was a main effects logistic regression model fit in the training data, i.e. $\mu\left(X^*\right) = \text{expit}(\beta_0 + \sum_{i=1}^{10} \beta_i X_{(i)})$. This model was misspecified with respect to the true generation process. As previously, we assessed the counterfactual performance of the model in an untreated population using the MSE, which for a binary outcome is equivalent to the Brier score [82]. In general, positing a parametric model for $h_0(X) = \text{E}[(Y - g\left(X^*\right))^2 \mid X, A = 0]$ may be difficult as the outcome is the squared difference. However, for binary outcomes, by expanding the square we can show it is enough to estimate $\Pr[Y = 1 \mid X, A = 0]$, which is what we did in practice. To determine the effect of the specification of nuisance models $e_a(X)$ and $h_a(X)$ on performance estimates, we compared four MSE estimators (ψ_{Naive}, ψ_{IPW}, ψ_{CL}, and ψ_{DR}) using different combinations of correctly specified and misspecified models for $e_a(X)$ and $h_a(X)$:

1. Correct $e_a(X)$ - main effects logistic regression model with linear and quadratic terms, i.e. $e_a(X) = \text{expit}(\alpha_0 + \sum_{i=1}^{10} \alpha_{1,i} X_{(i)} + \sum_{i=1}^{10} \alpha_{2,i} X_{(i)}^2)$.

2. Misspecified $e_a(X)$ - main effects logistic regression model with linear terms only terms, i.e. $e_a(X) = \text{expit}(\alpha_0 + \sum_{i=1}^{10} \alpha_{1,i} X_{(i)})$.

Estimator $\widehat{\psi}$	Mean	Bias ($\times 10^2$)	Bias (%)
Naive	0.244	0.603	2.5
Correct			
CL	0.238	0.058	0.2
IPW	0.238	0.095	0.4
DR	0.238	0.045	0.2
$e_a(X)$ misspecified			
CL	0.238	0.058	0.2
IPW	0.245	0.770	3.2
DR	0.238	0.059	0.2
$h_a(X)$ misspecified			
CL	0.246	0.867	3.6
IPW	0.238	0.095	0.4
DR	0.238	0.076	0.3
both misspecified			
CL, gam	0.240	0.227	1.0
IPW, gam	0.240	0.275	1.2
DR, gam	0.238	0.095	0.4
Truth	0.238	0.000	0.0

Correct and misspecified refers to the specification of the nuisance models ($e_a(X)$ or $h_a(X)$) for the MSE. Results were averaged over 10,000 simulations.

3. Correct $h_a(X)$ - main effects logistic regression model with linear and quadratic terms, i.e. $h_a(X) = \text{expit}(\gamma_0 + \sum_{i=1}^{10} \gamma_{1,i} X_{(i)} + \sum_{i=1}^{10} \gamma_{2,i} X_{(i)}^2)$.

4. Misspecified $h_a(X)$ - main effects logistic regression model with linear terms only terms, i.e. $h_a(X) = \text{expit}(\gamma_0 + \sum_{i=1}^{10} \gamma_{1,i} X_{(i)})$.

Finally, we also considered using more flexible estimation techniques for nuisance terms $e_a(X)$ and $h_a(X)$. Specifically, we fit generalized additive models for both using the `mgcv` package in R entering all covariates as splines using the default options in the `gam` function.

Table 2 shows the results. As in the previous experiment, the naive empirical estimator of the MSE was biased relative to the true counterfactual MSE with a relative bias of 2.5%. When all models were correctly specified, the weighting, conditional loss, and doubly robust estimators were all unbiased (relative bias between 0.2% to 0.4%). When $e_a(X)$ was misspecified, the weighting estimator was biased (relative bias of 3.2%) but the conditional loss and doubly robust estimator were unbiased (relative bias of 0.2%). Under misspecification of $h_a(X)$ (relative bias of 3.6%), the conditional loss estimator was biased, but the weighting estimator and the doubly robust estimator were unbiased (relative bias of 0.4% and 0.3%). When both models $e_a(X)$ and $h_a(X)$ were misspecified all estimators, including the doubly robust estimator, were biased. Finally, when a generalized additive model was used to estimate both $e_a(X)$ and $h_a(X)$, only the doubly robust estimator was approximately unbiased (relative bias of 0.4%). Across all scenarios, the weighting estimator generally had the largest standard errors and widest confidence intervals and the conditional loss estimator had the smallest standard errors and the shortest confidence intervals.

3.9 Application to prediction of statin-naive risk

Here we apply our proposed methods to evaluate the counterfactual performance of two prediction models targeting the statin-naive risk of cardiovascular disease: one that was explicitly tailored for the counterfactual estimand of interest and a second that was not.

3.9.1 Study design and data

The Multi-Ethnic Study on Atherosclerosis (MESA) study is a population-based sample of 6,814 men and women aged 45 to 84 drawn from six communities (Baltimore; Chicago; Forsyth County, North Carolina; Los Angeles; New York; and St. Paul, Minnesota) in the United States between 2000 and 2002. The sampling procedure, design, and methods of the study have been described previously [68]. Study teams conducted five examination visits between 2000 and 2011 in 18 to 24 month intervals focused on the prevalence, correlates, and progression of subclinical cardiovascular disease. These examinations included assessments of lipid-lowering (primarily statins) medication use as well as cardiovascular risk factors such as systolic blood pressure, serum cholesterol, cigarette smoking, height, weight, and diabetes.

In a previous analysis, we used MESA data to emulate a statin trial and benchmarked our results against those from published randomized trials. To construct a model of the statin-naive risk, we then emulated a single arm trial in which no one started statins over a 10-year follow up period. To determine trial eligibility, we followed the AHA guidelines [3] on statin use which stipulate that patients aged 40 to 75 with serum LDL cholesterol levels between 70 mg/dL and 190 mg/dL and no history of cardiovascular disease should initiate statins if their (statin-free) risk exceeds 7.5%. Therefore, we considered MESA participants who completed the baseline examination, had no previous history of statin use, no history of cardiovascular disease, and who met the criteria described in the guidelines (excluding the risk threshold) as eligible to participate in the trial. The primary endpoint was time to atherosclerotic cardiovascular disease (ASCVD), defined as nonfatal myocardial infarction, coronary heart disease death, or ischemic stroke.

Follow up began at the second examination cycle to enable a "wash out" period for statin use and to ensure adequate pre-treatment covariates to control confouding. In the original analysis, we constructed a sequence of nested trials, however here for simplicity we limited our attention to the first trial. We used the questionnaire in examinations three through five to determine statin initiation over the follow up period. Because the exact timing of statin initiation was not known with precision, we estimated it by drawing a random month between the current and previous examinations.

Of the 6,814 MESA participants who completed the baseline examination, 4,149 met the eligibility criteria for our trial emulation. There were 288 ASCVD events and 190 non-ASCVD deaths. For the sake of clarity, here we dropped those lost to follow up and ignored competing risks although in practice both can be accommodated in our framework for evaluting the performance of a counterfactual prediction model. For model training and evaluation we further split the dataset into training and test sets of equal size.

3.9.2 Model estimation and performance

We compared two prediction models: one that was explicitly tailored to the statin-naive risk and a second that was not. Both models used the same specification with baseline predictors commonly used in cardiovascular risk prediction: age, sex, smoking status, diabetes history, systolic blood pressure, anti-hypertensive medication use and total and HDL serum cholesterol levels. In the main text, to be consistent with our initial set up we assume the effect of statins is independent of duration and therefore may be viewed as a time-fixed intervention. Both trial evidence and subject matter knowledge suggest this is implausible, and we consider time-varying effects in the appendix.

We tailored the first model for the statin-naive risk using inverse probability of censoring weights. In the emulated single arm trial, statin initiation can be viewed as "non-adherence" which can be adjusted for by inverse probability weighting, therefore we censored participants when they initiated statins. To calculate the weights, we estimated two logistic regression models: one for the probability of remaining untreated given past covariate history (denominator model) and one for probability of remaining untreated given the selected baseline predictors (numerator model). The list of covariates in the weight models are given in the appendix. To create a prediction model for the statin-naive risk, we used the estimated weights to fit a weighted logistic regression model conditional on the baseline predictors of interest.

For comparison, we fit a second traditional (factual) prediction model by regressing the observed ASCVD event indicator on the same set of baseline predictors, but ignoring treatment initiation over the follow up period. This approach targets the natural course risk rather than the statin-naive risk. We fit the model using standard logistic regression based on maximum likelihood.

To assess the performance of the models, we estimated the naive and counterfactual MSE in the test set. For the latter we used the conditional loss, inverse probability weighting, and doubly robust estimators of the MSE. Models for the initiation of treatment $e_a(X)$ and for the conditional loss $h_a(X)$ were implemented as main effects logistic regression models. As in the simulation example, to estimate the conditional loss it is sufficient to model $\Pr[Y = 1 \mid X, A = 0]$ alone. To quantify uncertainty, we used the non-parametric bootstrap with 1000 bootstrap replicates.

3.9.3 Results

Table 3 shows estimates of the MSE and the associated standard errors in a hypothetical statin-naive population for both prediction models using the naive empirical, conditional

Table 3.1: *Estimated MSE in a statin-naive population for two prediction models using emulated trial data from MESA.*

Model $\mu_\beta(X)$	$\widehat{\psi}_{Naive}$	$\widehat{\psi}_{CL}$	$\widehat{\psi}_{IPW}$	$\widehat{\psi}_{DR}$
Logistic	0.066	0.091	0.111	0.095
	(0.003)	(0.006)	(0.012)	(0.007)
Weighted Logistic	0.070	0.090	0.102	0.091
	(0.003)	(0.004)	(0.008)	(0.005)

The first column refers to the posited prediction model: the first model is an (unweighted) logistic regression model and the second is a logistic regression model with inverse probability weights for remaining statin-free. $\widehat{\psi}_{Naive}$ is the empirical estimator of the MSE using factual outcomes, $\widehat{\psi}_{CL}$ is the conditional loss estimator, $\widehat{\psi}_{IPW}$ is the inverse probability weighting estimator, $\widehat{\psi}_{DR}$ is the doubly-robust estimator. Standard error estimates are shown in parentheses obtained via 1000 bootstrap replicates.

loss, weighting, and doubly robust estimators. The conditional loss, weighting, and doubly robust estimators of the MSE yielded estimates that were substantially (30-50%) greater than those of the naive empirical estimator, suggesting performance of both models in statin-naive population is worse than in the source population. Of the three estimators of the statin-naive MSE, the weighting estimator had greater standard errors than the doubly robust estimator (by 50-70%) as well as the conditional loss estimator (by 100%). Consistent with the first simulation experiment, the inverse probability weighted logistic model, which was tailored to target the statin-naive risk, performed worse in the source population, but had lower MSE in the counterfactual statin-naive population.

3.10 Discussion

Many practical problems in prediction modeling involve counterfactuals, such as when treatment varies between training and deployment or when predictions are meant to inform treatment initiation. Here, we considered cases where predictions under hypothetical interventions were desired but only training data from observational sources were available.

We described how to tailor models to target counterfactual estimands and the identification conditions necessary to unbiasedly estimate them. Separately, we also discussed how to adjust common measures of model performance to estimate the counterfactual performance of the model under the same hypothetical interventions. Importantly, our performance results were valid even when the prediction model is misspecified. A key insight was that for counterfactual prediction standard performance measures will be biased, but performance could be assessed independently from the method used to tailor the model. For loss-based metrics of performance, we proposed three estimators based on modeling the conditional loss, the probability of treatment, and a doubly robust estimator that can be used with data-adaptive estimators of either nuisance function.

In this paper, we have focused on measures of performance under a particular treatment regime. However, prediction models may instead target the estimation treatment effects, i.e. the comparison between treatment regimes. In some cases, effects may be easier to communicate to end users or may be desirable to evaluate benefits versus harms of treatment initiation [53]. Several authors have proposed model performance metrics which are similar to our own [83–87].

Throughout, we did not assume that the covariates needed to satisfy the exchangeability assumption were the same covariates used in the prediction model. This is important as, in practice, predictors are often chosen subject to clinical contraints in the data available to end users rather than what would be optimal from a causal perspective [1]. However, we did assume that a sufficient set of covariates could be identified at the time of training to ensure exchanageability. Alternative identification conditions are beyond the scope of this study, but it is possible that counterfactual performance metrics could also be identified, for instance, if an instrumental variable [88] were available or under a more general proximal inference framework [89]. It's also possible to develop sensitivity analyses for exploring how violations of this assumption might affect model performance estimates [67].

In this work, we have also implicitly assumed that the distribution of predictors are the same in the training and deployment setting. However, in many cases the covariate distributions are likely to differ [90, 91]. Like differences in treatment initiation, this may cause the performance of the prediction model to degrade, particularly when the model is misspecified. Methods for transporting prediction models from source to target populations which mirror our own have previously been proposed [10, 81, 92, 93]. In future work, it's possible that our results could be integrated with those to allow for both sources of difference between training and deployment.

References

1. Steyerberg, E. W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* doi:10.1007/978-3-030-16399-0 (Cham, 2019).

2. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2009).

3. Grundy Scott M. *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS /APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **139,** e1082–e1143. doi:10.1161/CIR.0000000000000625 (2019).

4. Damen, J. A. A. G. *et al.* Prediction Models for Cardiovascular Disease Risk in the General Population: Systematic Review. *BMJ* **353.** doi:10.1136/bmj.i2416 (2016).

5. Breiman, L. Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). *Statistical Science* **16,** 199–231. doi:10.1214/ss/1009213726 (2001).

6. Shmueli, G. To Explain or to Predict? *Statistical Science* **25,** 289–310 (2010).

7. Kennedy, E. H., Bonvini, M. & Mishler, A. Comment on "Statistical Modeling: The Two Cultures" by Leo Breiman. *Observational Studies* **7,** 145–156. doi:10.1353/obs.2021.0001 (2021).

8. Bühlmann, P. One Modern Culture of Statistics: Comments on Statistical Modeling: The Two Cultures (Breiman, 2001b). *Observational Studies* **7,** 33–40. doi:10.1353/obs.2021.0020 (2021).

9. Subbaswamy, A. & Saria, S. From Development to Deployment: Dataset Shift, Causality, and Shift-Stable Models in Health AI. *Biostatistics* **21,** 345–352. doi:10.1093/biostatistics/kxz041 (2020).

10. Steingrimsson, J. A., Gatsonis, C. & Dahabreh, I. J. *Transporting a Prediction Model for Use in a New Target Population* 2021.

11. van Geloven, N. *et al.* Prediction Meets Causal Inference: The Role of Treatment in Clinical Prediction Models. *Eur J Epidemiol* **35,** 619–630. doi:10.1007/s10654-020-00636-1 (2020).

12. Hernán, M. A. & Robins, J. M. *Causal Inference: What If* (Boca Raton, 2020).

13. Pearl, J. *Causality* (2009).

14. Imbens, G. W. & Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (2015).

15. Dickerman, B. A. & Hernán, M. A. Counterfactual Prediction Is Not Only for Causal Inference. *Eur J Epidemiol* **35,** 615–617. doi:`10.1007/s10654-020-00659-8` (2020).

16. Dickerman, B. A. *et al.* Predicting Counterfactual Risks under Hypothetical Treatment Strategies: An Application to HIV. *Eur J Epidemiol* **37,** 367–376. doi:`10.1007/s10654-022-00855-8` (2022).

17. Sperrin, M. *et al.* Using Marginal Structural Models to Adjust for Treatment Drop-in When Developing Clinical Prediction Models. *Statistics in Medicine* **37,** 4142–4154. doi:`10.1002/sim.7913` (2018).

18. Hernán, M. A. *et al.* Observational Studies Analyzed Like Randomized Experiments: An Application to Postmenopausal Hormone Therapy and Coronary Heart Disease. *Epidemiology* **19,** 766–779 (2008).

19. Hernán, M. A. & Robins, J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol* **183,** 758–764. doi:`10.1093/aje/kwv254` (2016).

20. Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. & Rosati, R. A. Regression Modelling Strategies for Improved Prognostic Prediction. *Statistics in Medicine* **3,** 143–152. doi:`10.1002/sim.4780030207` (1984).

21. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine* **15,** 361–387. doi:`10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4` (1996).

22. 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. *Mathematical Modelling* **7,** 1393–1512. doi:`10.1016/0270-0255(86)90088-6` (1986).

23. Hernán, M. A., Hsu, J. & Healy, B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE* **32,** 42–49. doi:`10.1080/09332480.2019.1579578` (2019).

24. Jewell, N. P. & Nielsen, J. P. A Framework for Consistent Prediction Rules Based on Markers. *Biometrika* **80,** 153–164. doi:`10.1093/biomet/80.1.153` (1993).

25. Wolbers, M., Koller, M. T., Witteman, J. C. M. & Steyerberg, E. W. Prognostic Models With Competing Risks: Methods and Application to Coronary Risk Prediction. *Epidemiology* **20,** 555–561 (2009).

26. Sperrin, M., Martin, G. P., Sisk, R. & Peek, N. Missing Data Should Be Handled Differently for Prediction than for Description or Causal Explanation. *Journal of Clinical Epidemiology* **125,** 183–187. doi:10.1016/j.jclinepi.2020.03.028 (2020).

27. Young, J. G., Stensrud, M. J., Tchetgen, E. J. T. & Hernán, M. A. A Causal Framework for Classical Statistical Estimands in Failure-Time Settings with Competing Events. *Statistics in Medicine* **39,** 1199–1236. doi:10.1002/sim.8471 (2020).

28. Taubman, S. L., Robins, J. M., Mittleman, M. A. & Hernán, M. A. Intervening on Risk Factors for Coronary Heart Disease: An Application of the Parametric g-Formula. *Int J Epidemiol* **38,** 1599–1611. doi:10.1093/ije/dyp192 (2009).

29. Young, J. G., Hernán, M. A. & Robins, J. M. Identification, Estimation and Approximation of Risk under Interventions That Depend on the Natural Value of Treatment Using Observational Data. *Epidemiologic Methods* **3,** 1–19. doi:10.1515/em-2012-0001 (2014).

30. Young, J. G., Cain, L. E., Robins, J. M., O'Reilly, E. J. & Hernán, M. A. Comparative Effectiveness of Dynamic Treatment Regimes: An Application of the Parametric G-Formula. *Stat Biosci* **3,** 119. doi:10.1007/s12561-011-9040-7 (2011).

31. Schoop, R., Graf, E. & Schumacher, M. Quantifying the Predictive Performance of Prognostic Models for Censored Survival Data with Time-Dependent Covariates. *Biometrics* **64,** 603–610. doi:10.1111/j.1541-0420.2007.00889.x (2008).

32. Schoop, R., Beyersmann, J., Schumacher, M. & Binder, H. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal* **53,** 88–112. doi:10.1002/bimj.201000073 (2011).

33. Wilson, P. W. F. *et al.* Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* **97,** 1837–1847. doi:10.1161/01.CIR.97.18.1837 (1998).

34. Goff, D. C. *et al.* 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation* **129,** S49–S73. doi:10.1161/01.cir.0000437741.48606.98 (2014).

35. Wen, L., Young, J. G., Robins, J. M. & Hernán, M. A. Parametric G-Formula Implementations for Causal Survival Analyses. *Biometrics* **n/a.** doi:10.1111/biom.13321.

36. Pajouheshnia, R., Peelen, L. M., Moons, K. G. M., Reitsma, J. B. & Groenwold, R. H. H. Accounting for Treatment Use When Validating a Prognostic Model: A Simulation Study. *BMC Med Res Methodol* **17,** 103. doi:10.1186/s12874-017-0375-8 (2017).

37. Bang, H. & Robins, J. M. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics* **61,** 962–973. doi:10.1111/j.1541-0420.2005.00377.x (2005).

38. Chernozhukov, V. *et al.* Double/Debiased Machine Learning for Treatment and Structural Parameters. *Econom J* **21,** C1–C68. doi:10.1111/ectj.12097 (2018).

39. Robins, J. M., Rotnitzky, A. & Zhao, L. P. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association* **89,** 846–866. doi:10.2307/2290910 (1994).

40. Díaz, I., Williams, N., Hoffman, K. L. & Schenck, E. J. Non-Parametric Causal Effects Based on Longitudinal Modified Treatment Policies. *arXiv:2006.01366 [stat]* (2021).

41. Papageorgiou, G., Mauff, K., Tomer, A. & Rizopoulos, D. An Overview of Joint Modeling of Time-to-Event and Longitudinal Outcomes. *Annu. Rev. Stat. Appl.* **6,** 223–240. doi:10.1146/annurev-statistics-030718-105048 (2019).

42. Rizopoulos, D. Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. *Biometrics* **67,** 819–829 (2011).

43. Paige, E. *et al.* Use of Repeated Blood Pressure and Cholesterol Measurements to Improve Cardiovascular Disease Risk Prediction: An Individual-Participant-Data Meta-Analysis. *Am J Epidemiol* **186,** 899–907. doi:10.1093/aje/kwx149 (2017).

44. Lindbohm, J. V. *et al.* Association between Change in Cardiovascular Risk Scores and Future Cardiovascular Disease: Analyses of Data from the Whitehall II Longitudinal, Prospective Cohort Study. *The Lancet Digital Health* **3,** e434–e444. doi:10.1016/S2589-7500(21)00079-0 (2021).

45. Hernán, M. A. & Taubman, S. L. Does Obesity Shorten Life? The Importance of Well-Defined Interventions to Answer Causal Questions. *Int J Obes* **32,** S8–S14. doi:10.1038/ijo.2008.82 (2008).

46. Westreich, D. & Cole, S. R. Invited Commentary: Positivity in Practice. *American Journal of Epidemiology* **171,** 674–677. doi:10.1093/aje/kwp436 (2010).

47. McGrath, S., Young, J. G. & Hernán, M. A. Revisiting the G-Null Paradox. *Epidemiology* **33,** 114–120. doi:10.1097/EDE.0000000000001431 (2022).

48. Robins, J. M. & Wasserman, L. A. Estimation of Effects of Sequential Treatments by Reparameterizing Directed Acyclic Graphs. *arXiv:1302.1566 [stat]* (2013).

49. Lin, V. *et al.* gfoRmula: An R Package for Estimating Effects of General Time-Varying Treatment Interventions via the Parametric g-Formula. *arXiv:1908.07072 [stat]* (2019).

50. Zivich, P. N. & Breskin, A. Machine Learning for Causal Inference: On the Use of Cross-fit Estimators. *Epidemiology* **32,** 393–401. doi:10.1097/EDE.0000000000001332 (2021).

51. Dahabreh, I. J., Robins, J. M., Haneuse, S. J.-P. A. & Hernán, M. A. Generalizing Causal Inferences from Randomized Trials: Counterfactual and Graphical Identification. *arXiv:1906.10792 [stat]* (2019).

52. Dahabreh, I. J., Robins, J. M. & Hernán, M. A. Benchmarking Observational Methods by Comparing Randomized Trials and Their Emulations. *Epidemiology* **31,** 614–619. doi:10.1097/EDE.0000000000001231 (2020).

53. Kent, D. M. *et al.* The Predictive Approaches to Treatment Effect Heterogeneity (PATH) Statement. *Ann Intern Med* **172,** 35–45. doi:10.7326/M18-3667 (2020).

54. Djulbegovic, B. & Paul, A. From Efficacy to Effectiveness in the Face of Uncertainty: Indication Creep and Prevention Creep. *JAMA* **305,** 2005–2006. doi:10.1001/jama.2011.650 (2011).

55. Finlayson, S. G. *et al.* The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med* **385,** 283–286. doi:10.1056/NEJMc2104626 (2021).

56. Liew, S. M., Doust, J. & Glasziou, P. Cardiovascular Risk Scores Do Not Account for the Effect of Treatment: A Review. *Heart* **97,** 689–697. doi:10.1136/hrt.2010.220442 (2011).

57. Xu, Z. *et al.* Prediction of Cardiovascular Disease Risk Accounting for Future Initiation of Statin Treatment. *American Journal of Epidemiology* **190,** 2000–2014. doi:10.1093/aje/kwab031 (2021).

58. Robins, J. M., Blevins, D., Ritter, G. & Wulfsohn, M. G-Estimation of the Effect of Prophylaxis Therapy for Pneumocystis Carinii Pneumonia on the Survival of AIDS Patients. *Epidemiology* **3,** 319–336 (1992).

59. Robins, J. M. in *Methodological Issues in AIDS Behavioral Research* (eds Ostrow, D. G., Kelly, J. A., Ostrow, D. G. & Kessler, R. C.) 213–288 (Boston, MA, 2002). doi:10.1007/0-306-47137-X_12.

60. Robins, J. M. Structural Nested Failure Time Models. *Encyclopedia of Biostatistics* **6,** 4372–4389 (1998).

61. Hernán, M. A., Robins, J. M. & García Rodríguez, L. A. Discussion on 'Statistical Issues Arising in the Women's Health Initiative'. *Biometrics* **61,** 922–930. doi:10.1111/j.0006-341X.2005.454_7.x (2005).

62. Vansteelandt, S. & Joffe, M. Structural Nested Models and G-estimation: The Partially Realized Promise. *Statistical Science* **29,** 707–731 (2014).

63. Guo, F. R., Richardson, T. S. & Robins, J. M. Discussion of 'Estimating Time-Varying Causal Excursion Effects in Mobile Health with Binary Outcomes'. *Biometrika* **108,** 541–550. doi:10.1093/biomet/asab029 (2021).

64. Witteman, J. C. M. *et al.* G-Estimation of Causal Effects: Isolated Systolic Hypertension and Cardiovascular Death in the Framingham Heart Study. *American Journal of Epidemiology* **148,** 390–401. doi:10.1093/oxfordjournals.aje.a009658 (1998).

65. Joffe, M. M., Yang, W. P. & Feldman, H. G-Estimation and Artificial Censoring: Problems, Challenges, and Applications. *Biometrics* **68,** 275–286. doi:10.1111/j.1541-0420.2011.01656.x (2012).

66. D'Agostino, R. B. *et al.* Relation of Pooled Logistic Regression to Time Dependent Cox Regression Analysis: The Framingham Heart Study. *Statistics in Medicine* **9,** 1501–1515. doi:10.1002/sim.4780091214 (1990).

67. Robins, J. M., Rotnitzky, A. & Scharfstein, D. O. *Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models* in *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (eds Halloran, M. E. & Berry, D.) (New York, NY, 2000), 1–94. doi:10.1007/978-1-4612-1284-3_1.

68. Bild, D. E. *et al.* Multi-Ethnic Study of Atherosclerosis: Objectives and Design. *American Journal of Epidemiology* **156,** 871–881. doi:10.1093/aje/kwf113 (2002).

69. Dahabreh, I. J. & Hernán, M. A. Extending Inferences from a Randomized Trial to a Target Population. *Eur J Epidemiol* **34,** 719–722. doi:10.1007/s10654-019-00533-2 (2019).

70. Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A. & Hernán, M. A. Generalizing Causal Inferences from Individuals in Randomized Trials to All Trial-Eligible Individuals. *Biometrics* **75,** 685–694. doi:10.1111/biom.13009 (2019).

71. Picciotto, S., Hernán, M. A., Page, J. H., Young, J. G. & Robins, J. M. Structural Nested Cumulative Failure Time Models to Estimate the Effects of Interventions. *Journal of the American Statistical Association* **107,** 886–900. doi:10.1080/01621459.2012.682532 (2012).

72. Seaman, S., Dukes, O., Keogh, R. & Vansteelandt, S. Adjusting for Time-Varying Confounders in Survival Analysis Using Structural Nested Cumulative Survival Time Models. *Biometrics* **76,** 472–483. doi:10.1111/biom.13158 (2020).

73. Wager, S. & Athey, S. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association* **113,** 1228–1242. doi:10.1080/01621459.2017.1319839 (2018).

74. Kennedy, E. H. *Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects* 2022.

75. Ferreira, J. P. *et al.* Individualizing Treatment Choices in the Systolic Blood Pressure Intervention Trial. *Journal of Hypertension* **36,** 428–435. doi:10.1097/HJH.0000000000001535 (2018).

76. Lin, L., Sperrin, M., Jenkins, D. A., Martin, G. P. & Peek, N. A Scoping Review of Causal Methods Enabling Predictions under Hypothetical Interventions. *Diagn Progn Res* **5,** 3. doi:10.1186/s41512-021-00092-9 (2021).

77. Schulam, P. & Saria, S. *Reliable Decision Support Using Counterfactual Models* in *Advances in Neural Information Processing Systems* **30** (2017).

78. Altman, D. G. & Royston, P. What Do We Mean by Validating a Prognostic Model? *Statistics in Medicine* **19,** 453–473. doi:10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5 (2000).

79. Robins, J. A Graphical Approach to the Identification and Estimation of Causal Parameters in Mortality Studies with Sustained Exposure Periods. *Journal of Chronic Diseases* **40,** 139S–161S. doi:10.1016/S0021-9681(87)80018-8 (1987).

80. Robins, J., Li, L., Tchetgen, E. & van der Vaart, A. Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals. *Probability and Statistics: Essays in Honor of David A. Freedman* **2,** 335–422. doi:10.1214/193940307000000527 (2008).

81. Morrison, S., Gatsonis, C., Dahabreh, I. J., Li, B. & Steingrimsson, J. A. *Robust Estimation of Loss-Based Measures of Model Performance under Covariate Shift* 2022.

82. Brier, G. W. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Mon. Wea. Rev.* **78,** 1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2 (1950).

83. Schuler, A., Baiocchi, M., Tibshirani, R. & Shah, N. *A Comparison of Methods for Model Selection When Estimating Individual Treatment Effects* 2018.

84. Rolling, C. A. & Yang, Y. Model Selection for Estimating Treatment Effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76,** 749–769 (2014).

85. Xu, Y. & Yadlowsky, S. *Calibration Error for Heterogeneous Treatment Effects* 2022.

86. Van der Laan, M. J. & Robins, J. M. *Unified Methods for Censored Longitudinal Data and Causality* (2003).

87. Alaa, A. & Schaar, M. V. D. *Validating Causal Inference Models via Influence Functions* in *Proceedings of the 36th International Conference on Machine Learning* (2019), 191–201.

88. Hernan, M. A. & Robins, J. M. Instruments for Causal Inference. **17** (2006).

89. Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X. & Miao, W. *An Introduction to Proximal Causal Learning* 2020.

90. Bickel, S., Brückner, M. & Scheffer, T. Discriminative Learning Under Covariate Shift. *J. Mach. Learn. Res.* **10,** 2137–2155 (2009).

91. Sugiyama, M., Krauledat, M. & Müller, K.-R. Covariate Shift Adaptation by Importance Weighted Cross Validation. *Journal of Machine Learning Research* **8,** 985–1005 (2007).

92. Steingrimsson, J. A. Extending Prediction Models for Use in a New Target Population with Failure Time Outcomes. *Biostatistics,* kxac011. doi:10.1093/biostatistics/kxac011 (2022).

93. Li, B., Gatsonis, C., Dahabreh, I. J. & Steingrimsson, J. A. Estimating the Area under the ROC Curve When Transporting a Prediction Model to a Target Population. *Biometrics,* biom.13796. doi:10.1111/biom.13796 (2022).

Appendix A

Appendix to Chapter 1

A.1 Identification results

A.1.1 End of follow up outcome

Here we apply the proofs in Robins [22] to our conditional g-formula estimator for any deterministic regime g. Assume the data structure outlined in section 1.2 and the following conditions hold:

1. *Exchangeability:* $Y^g_{K+1} \perp\!\!\!\perp A_k \mid \overline{L}_k, \overline{P}_k, \overline{A}_{k-1}$
2. *Consistency:* $Y_{K+1} = Y^g_{K+1},\ \overline{L}_k = \overline{L}^g_k,\ \text{and}\ \overline{P}_k = \overline{P}^g_k\ \text{ if }\ \overline{A}_k = \overline{a}^g_k$
3. *Positivity:* $\Pr(A_k = a_k \mid \overline{L}_k = \overline{l}_k, \overline{P}_k = \overline{p}_k, \overline{A}_{k-1} = \overline{a}_{k-1}) > 0$

By the rules of probability

$$\begin{aligned}
&\Pr(Y^g_{K+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k) = \\
&\qquad \sum_{l_{k+1}} \sum_{p_{k+1}} \Pr(Y^g_{k+1} = 1 \mid \overline{L}_k, \overline{A}_k, \overline{P}_k, L_{k+1} = l_{k+1}, P_{k+1} = p_{k+1}) \times \\
&\qquad\qquad f(l_{k+1} \mid \overline{L}_k, \overline{P}_k, \overline{A}_k) \times f(p_{k+1} \mid l_{k+1}, \overline{L}_k, \overline{P}_k, \overline{A}_k)
\end{aligned}$$

By exchangeability condition (1)

$$\begin{aligned}&\Pr(Y^g_{K+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k) = \\ &\qquad \sum_{l_{k+1}} \sum_{p_{k+1}} \Pr(Y^g_{k+1} = 1 \mid \overline{L}_k, \overline{A}_k, \overline{P}_k, L_{k+1} = l_{k+1}, P_{k+1} = p_{k+1}, A_{k+1} = a^g_{k+1}) \times \\ &\qquad\qquad f(l_{k+1} \mid \overline{L}_k, \overline{P}_k, \overline{A}_k) \times f(p_{k+1} \mid l_{k+1}, \overline{L}_k, \overline{P}_k, \overline{A}_k)\end{aligned}$$

Arguing recursively from k to K for the complete regime $g(\underline{a}_k) = \underline{a}^g_k$

$$\begin{aligned}&\Pr(Y^g_{K+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k) = \\ &\quad \sum_{\underline{l}_k} \sum_{\underline{p}_k} \sum_{\underline{a}_k} \Pr(Y^g_{k+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, \underline{L}_k = \underline{l}_k, \underline{P}_k = \underline{p}_k, \underline{A}_k = \underline{a}^g_k) \times \\ &\qquad \prod_{j=k}^{K} \left\{ f(l_j \mid \bar{l}_{j-1}, \overline{p}_{j-1}, \overline{a}^g_{j-1}) \times f(p_j \mid \bar{l}_j, \overline{p}_{j-1}, \overline{a}^g_{j-1}) \right\}\end{aligned}$$

By consistency (2) and positivity (3)

$$\begin{aligned}&\Pr(Y^g_{K+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k) = \\ &\quad \sum_{\underline{l}_k} \sum_{\underline{p}_k} \sum_{\underline{a}_k} \Pr(Y_{k+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, \underline{L}_k = \underline{l}_k, \underline{P}_k = \underline{p}_k, \underline{A}_k = \underline{a}^g_k) \times \\ &\qquad \prod_{j=k}^{K} \left\{ f(l_j \mid \bar{l}_{j-1}, \overline{p}_{j-1}, \overline{a}^g_{j-1}) \times f(p_j \mid \bar{l}_j, \overline{p}_{j-1}, \overline{a}^g_{j-1}) \right\}\end{aligned}$$

For random and dynamic regimes g, following the arguments in Young *et al.* [30] and Lemma 4.2 of Robins [22], we can show that the generalized g-formula expression in 1.3 is a particular weighted average of the g-formula for a deterministic regime for the set of all deterministic regimes that satisfy positivity when $f^{obs}(a_j \mid \bar{l}_j, \overline{p}_j, \overline{a}_{j-1})$ is replaced with $f^g(a_j \mid \bar{l}_j, \overline{p}_j, \overline{a}_{j-1})$ and only depend on past covariate history.

For continuous outcomes, the conditional g-formula estimator for an end of follow up outcome is given by

$$\begin{aligned}
&\mathrm{E}(Y^{g}_{K+1} \mid \overline{L}_k = \bar{l}_k, \overline{P}_k = \overline{p}_k, \overline{A}_k = \bar{a}_k) = \\
&\qquad \sum_{\underline{l}_k}\sum_{\underline{p}_k} \mathrm{E}(Y_{k+1} \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, \underline{L}_k = \underline{l}_k, \underline{P}_k = \underline{p}_k, \underline{A}_k = \underline{a}^{g}_{k}) \times \\
&\qquad\qquad \prod_{j=k}^{K} \left\{ f(l_j \mid \bar{l}_{j-1}, \overline{p}_{j-1}, \bar{a}^{g}_{j-1}) \times f(p_j \mid \bar{l}_j, \overline{p}_{j-1}, \bar{a}_{j-1}) \right\}
\end{aligned}$$

noting that when Y_{K+1} is binary $\mathrm{E}(Y_{K+1} \mid \overline{X}_k = \bar{x}_k) = \Pr(Y_{K+1} = 1 \mid \overline{X}_k = \bar{x}_k)$.

A.1.2 Survival outcome

For survival outcomes, we introduce the possibility that subjects are lost to follow up and the possibility of competing events. In section, 1.2.5 we distinguished between estimands involving elimination of competing events and those that do not. We start by identifying estimands with elimination of competing events. The modified eligibility criteria are:

1. *Exchangeability:*

$$\underline{Y}^{g,\bar{c}=0,\bar{d}=0}_{k+1} \perp\!\!\!\perp (A_k, C_{k+1}, D_{k+1}) \mid \overline{L}_k, \overline{P}_k, \overline{A}_{k-1}, \overline{Y}_k = \overline{C}_k = \overline{D}_k = 0$$

2. *Consistency:*

$$\overline{Y}_{k+1} = \overline{Y}^{g,\bar{c}=0,\bar{d}=0}_{k+1},\ \overline{L}_{k+1} = \overline{L}^{g,\bar{c}=0,\bar{d}=0}_{k+1},\ \text{and}\ \overline{P}_{k+1} = \overline{P}^{g,\bar{c}=0,\bar{d}=0}_{k+1}$$
$$\text{if}\ \overline{A}_k = \bar{a}^{g}_{k}, \overline{D}_k = 0,\ \text{and}\ \overline{C}_k = 0$$

3. *Positivity:*

$$\Pr(A_k = \bar{a}^{g}_{k}, C_{k+1} = 0 \mid \overline{L}_k = \bar{l}_k, \overline{P}_k = \overline{p}_k, \overline{A}_{k-1} = \bar{a}^{g}_{k-1}, \overline{Y}_k = \overline{C}_k = \overline{D}_k = 0) > 0$$

For simplicity of exposition let $Y^{g,\bar{c}=0,\overline{d}=0} = Y^{g^*}$By definition for a survival outcome $Y^{g,\bar{c}=0,\overline{d}=0}_{K+1}$ we have

$$\Pr(Y^{g^*}_{K+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k) = \\ \sum_{k=k}^{K} \Pr(Y^{g^*}_{k+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, \overline{Y}^{g^*}_k = 0) \times \prod_{j=k}^{K} \Pr(Y^{g^*}_j = 0 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, \overline{Y}^{g^*}_{j-1} = 0)$$

By the rules of probability

$$\Pr(Y^{g^*}_{K+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k) = \\ \sum_{l_{k+1}} \sum_{p_{k+1}} \sum_{k=k}^{K} \Pr(Y^{g^*}_{k+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, L_{k+1} = l_{k+1}, P_{k+1} = p_{k+1}, \overline{Y}^{g^*}_k = 0) \times \\ \prod_{j=k}^{K} \Pr(Y^{g^*}_j = 0 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, L_{k+1} = l_{k+1}, P_{k+1} = p_{k+1}, \overline{Y}^{g^*}_{j-1} = 0) \times \\ f(l_{k+1} \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, \overline{Y}^{g^*}_k = 0) \times f(p_{k+1} \mid l_{k+1}, \overline{L}_k, \overline{P}_k, \overline{A}_k, \overline{Y}^{g^*}_k = 0)$$

By exchangeability condition (1)

$$\Pr(Y^{g^*}_{K+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k) = \\ \sum_{l_{k+1}} \sum_{p_{k+1}} \sum_{k=k}^{K} \Pr(Y^{g^*}_{k+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, l_{k+1}, p_{k+1}, a^{g^*}_{k+1}, \overline{Y}^{g^*}_k = D_{k+1} = C_{k+1} = 0) \times \\ \prod_{j=k}^{K} \Pr(Y^{g^*}_j = 0 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, l_{k+1}, p_{k+1}, a^{g^*}_{k+1}, \overline{Y}^{g^*}_{j-1} = 0) \times \\ f(l_{k+1} \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, \overline{Y}^{g^*}_k = 0) \times f(p_{k+1} \mid l_{k+1}, \overline{L}_k, \overline{P}_k, \overline{A}_k, \overline{Y}^{g^*}_k = 0)$$

Arguing recursively from k to K for the complete regime $g(\underline{a}_k) = \underline{a}^g_k$

$$\Pr(Y^{g^*}_{K+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k) = \\ \sum_{\underline{l}_k} \sum_{\underline{p}_k} \sum_{\underline{a}_k} \sum_{k=k}^{K} \Pr(Y^{g^*}_{k+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, \underline{L}_k = \underline{l}_k, \underline{P}_k = \underline{p}_k, \underline{A}_k = \underline{a}^g_k, \overline{Y}^{g^*}_k = \overline{C}_k = \overline{D}_k = 0) \times \\ \prod_{j=k}^{K} \Big\{ \Pr(Y^{g^*}_j = 0 \mid \overline{L}_{j-1} = \bar{l}_{j-1}, \overline{P}_{j-1} = \overline{p}_{j-1}, \overline{A}_{j-1} = \overline{a}^g_{j-1}, \overline{Y}^{g^*}_{j-1} = \overline{D}_j = \overline{C}_j = 0) \times \\ f(l_j \mid \bar{l}_{j-1}, \overline{p}_{j-1}, \overline{a}^g_{j-1}, \overline{Y}^{g^*}_j = \overline{D}_j = \overline{C}_j = 0) \times f(p_j \mid \bar{l}_j, \overline{p}_j, \overline{a}^g_{j-1}, \overline{Y}^{g^*}_j = \overline{D}_j = \overline{C}_j = 0) \Big\}$$

By consistency (2) and positivity (3)

$$
\begin{aligned}
&\Pr(Y_{K+1}^{g,\bar{c}=0,\bar{d}=0} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k) = \\
&\quad \sum_{\underline{l}_k} \sum_{\underline{p}_k} \sum_{\underline{a}_k} \sum_{k=k}^{K} \Pr(Y_{k+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, \underline{L}_k = \underline{l}_k, \underline{P}_k = \underline{p}_k, \underline{A}_k = \underline{a}_k, \overline{Y}_k = \overline{C}_k = \overline{D}_k = 0) \times \\
&\quad\quad \prod_{j=k}^{K} \Big\{ \Pr(Y_j = 0 \mid \overline{L}_{j-1} = \overline{l}_{j-1}, \overline{P}_{j-1} = \overline{p}_{j-1}, \overline{A}_{j-1} = \overline{a}_{j-1}, \overline{Y}_{j-1} = \overline{D}_j = \overline{C}_j = 0) \times \\
&\quad\quad f(l_j \mid \overline{l}_{j-1}, \overline{p}_{j-1}, \overline{a}^g_{j-1}, \overline{Y}_j = \overline{D}_j = \overline{C}_j = 0) \times f(p_j \mid \overline{l}_j, \overline{p}_{j-1}, \overline{a}_{j-1}, \overline{Y}_j = \overline{D}_j = \overline{C}_j = 0) \Big\}
\end{aligned}
\tag{A.1}
$$

Applying similar logic we can also derive the estimator for the risk without elimination of competing events under slightly weaker conditions, i.e.

1. *Exchangeability:*

$$\underline{Y}_{k+1}^{g,\bar{c}=0} \perp\!\!\!\perp (A_k, C_{k+1}) \mid \overline{L}_k, \overline{P}_k, \overline{A}_{k-1}, \overline{D}_k = \overline{d}_k, \overline{Y}_k = \overline{C}_k = 0$$

2. *Consistency:*

$$\overline{Y}_{k+1} = \overline{Y}_{k+1}^{g,\bar{c}=0}, \overline{D}_{k+1} = \overline{D}_{k+1}^{g,\bar{c}=0}, \overline{L}_{k+1} = \overline{L}_{k+1}^{g,\bar{c}=0}, \text{ and } \overline{P}_{k+1} = \overline{P}_{k+1}^{g,\bar{c}=0}$$
$$\text{if } \overline{A}_k = \overline{a}^g_k, \text{ and } \overline{C}_k = 0$$

3. *Positivity:*

$$\Pr(A_k = \overline{a}^g_k, C_{k+1} = 0 \mid \overline{L}_k = \overline{l}_k, \overline{P}_k = \overline{p}_k, \overline{A}_{k-1} = \overline{a}^g_{k-1}, \overline{D}_k = \overline{d}_k, \overline{Y}_k = \overline{C}_k = 0) > 0$$

We omit the steps for brevity but the resulting expression is

$$\begin{aligned}
&\Pr(Y_{K+1}^{g,\bar{c}=0} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k) = \\
&\quad \sum_{\underline{l}_k} \sum_{\underline{p}_k} \sum_{k=k}^{K} \Pr(Y_{k+1} = 1 \mid \overline{L}_k, \overline{P}_k, \overline{A}_k, \underline{L}_k = \underline{l}_k, \underline{P}_k = \underline{p}_k, \underline{A}_k = \underline{a}_k^g, \overline{Y}_k = \overline{C}_k = \overline{D}_k = 0) \times \\
&\quad\quad \prod_{j=k}^{K} \Big\{ \Pr(Y_j = 0 \mid \overline{L}_{j-1} = \bar{l}_{j-1}, \overline{P}_{j-1} = \overline{p}_{j-1}, \overline{A}_{j-1} = \overline{a}_{j-1}^g, \overline{Y}_{j-1} = \overline{C}_j = \overline{D}_j = 0) \times \\
&\quad\quad\quad \Pr(D_{j+1} = 0 \mid \overline{L}_{j-1} = \bar{l}_{j-1}, \overline{P}_{j-1} = \overline{p}_{j-1}, \overline{A}_{j-1} = \overline{a}_{j-1}^g, \overline{Y}_j = \overline{C}_{j+1} = \overline{D}_j = 0) \times \\
&\quad\quad\quad f(l_j \mid \bar{l}_{j-1}, \overline{p}_{j-1}, \overline{a}_{j-1}^g, \overline{Y}_j = \overline{D}_j = \overline{C}_j = 0) \times f(p_j \mid \bar{l}_j, \overline{p}_{j-1}, \bar{a}_{j-1}, \overline{Y}_j = \overline{D}_j = \overline{C}_j = 0) \Big\}
\end{aligned} \tag{A.2}$$

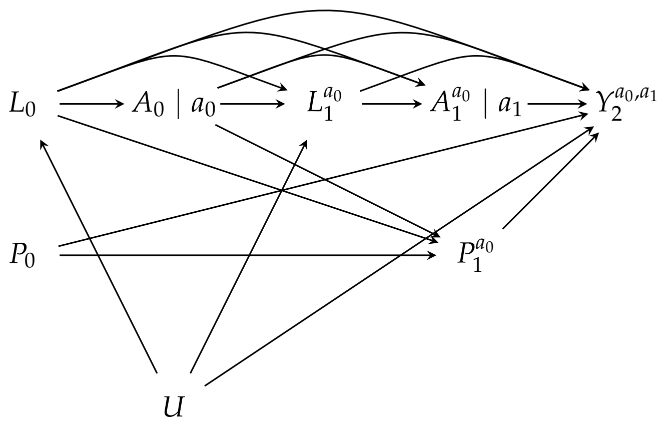


Figure A.1: *Example single world intervention graph for end of follow up outcome.*

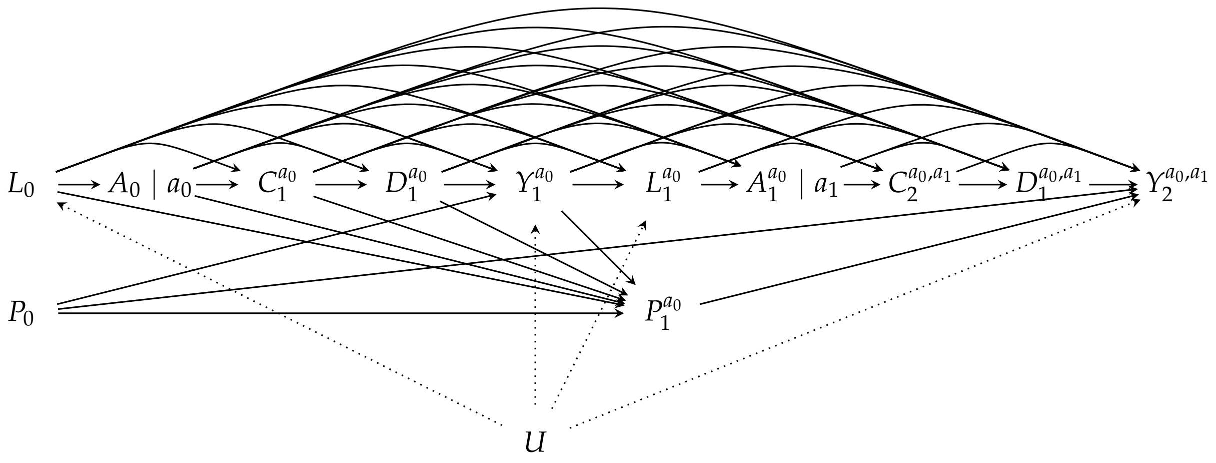


Figure A.2: *Example single world intervention graph for survival outcome with competing event D_k.*

A.2 What if not all covariates are available during counseling?

Often, in clinical prediction tasks, the predictors in X_k are determined by the information available to the decision-maker rather than what might be optimal from a statistical or theoretical point-of-view. For instance, certain laboratory values may be cost prohibitive or may take too long to collect relative to the decision timeline. For factual prediction tasks, this is not an issue and the covariates in X_k can be determined by the operating constraints of the decision-maker. However, for counterfactual prediction tasks, during training X_k must include all L_k sufficient to ensure exchangeability of potential outcomes with respect to treatments A_k in order to yield unbiased predictions. In practice, this may necessitate selecting training data where either (1) exchangeability is assured by design, such as in a randomized controlled trial, or (2) covariate data for X_k is sufficiently rich to make identification plausible. Once accomplished, the g-formula in the main text can be modified to produce predictions based on a subset $V_k \subset X_k$ of covariates available to the decision-maker by summing/integrating out the covariates that are not available $V_k^* = X_k - V_k$, e.g.

$$
\begin{aligned}
&\Pr(Y^g_{K+1} = 1 \mid \overline{V}_k = \overline{v}_k) = \\
&\qquad \sum_{\overline{v}^*_k} \sum_{\underline{l}_k} \sum_{\underline{a}_k} \sum_{k=k}^{K} \Pr(Y_{k+1} = 1 \mid \overline{V}_k = \overline{v}_k, \overline{V}^*_k = \overline{v}^*_k, \underline{A}_k = \underline{a}^g_k, \underline{L}_k = \underline{l}_k, \overline{Y}_k = 0) \times \\
&\qquad \prod_{j=k}^{K} \Bigg\{ \Pr(Y_j = 0 \mid \overline{A}_{j-1} = \overline{a}^g_{j-1}, \overline{L}_{j-1} = \overline{l}_{j-1}, \overline{Y}_{j-1} = 0) \times \\
&\qquad\qquad f(l_j \mid \overline{a}^g_{j-1}, \overline{l}_{j-1}, \overline{Y}_j = 0) \times f^g(a_j \mid \overline{a}_{j-1}, \overline{l}_j, \overline{Y}_j = 0) \Bigg\}
\end{aligned}
\tag{A.3}
$$

An alternative approach, as in [16], is to use the g-formula based on X_k to simulate datasets $(\overline{X}_k, \overline{Y}^g)$ under the regime of interest g and then fit a new model on the simulated data using traditional approaches such as logistic or Cox regression.

A.3 Simulation study results

Below we include additional results not featured in the main text.

Table A.1: *Monte carlo simulation results comparing g-formula and landmark approaches under misspecified covariate models.*

	MSE(Δk, k^*)			AUC(Δk, k^*)		
k^*	g-formula	landmark	landmark (lags)	g-formula	landmark	landmark (lags)
Scenario 1: Factual prediction						
0	**0.126**	**0.126**	**0.126**	**0.834**	**0.834**	**0.834**
	(0.007)	(0.007)	(0.007)	(0.013)	(0.013)	(0.013)
3	**0.115**	0.117	0.116	0.850	**0.851**	**0.851**
	(0.006)	(0.007)	(0.007)	(0.012)	(0.012)	(0.012)
6	**0.102**	0.104	0.103	**0.871**	0.870	0.870
	(0.006)	(0.006)	(0.006)	(0.012)	(0.012)	(0.012)
Scenario 2: Competing risk prediction						
0	**0.111**	0.112	0.112	**0.838**	**0.838**	**0.838**
	(0.007)	(0.007)	(0.007)	(0.017)	(0.017)	(0.017)
3	**0.107**	**0.107**	**0.107**	0.854	**0.855**	0.853
	(0.006)	(0.007)	(0.007)	(0.016)	(0.016)	(0.016)
6	**0.099**	**0.099**	**0.099**	**0.870**	0.869	0.869
	(0.006)	(0.006)	(0.006)	(0.015)	(0.015)	(0.016)
Scenario 3: Counterfactual prediction						
0	**0.233**	0.543	0.543	**0.930**	0.906	0.906
	(0.012)	(0.023)	(0.023)	(0.007)	(0.010)	(0.010)
3	**0.172**	0.418	0.411	**0.939**	0.932	0.933
	(0.010)	(0.017)	(0.017)	(0.006)	(0.008)	(0.008)
6	**0.128**	0.227	0.209	**0.943**	**0.943**	**0.943**
	(0.008)	(0.017)	(0.020)	(0.006)	(0.006)	(0.006)

Note:

All results based on 500 Monte Carlo simulations using data generation process described in section 3. Standard deviations of Monte Carlo estimates are provided in parentheses. The best performing estimator is shown in **bold**. All simulations use correctly specified models. For results under misspecification see the appendix.

Table A.2: *Monte carlo simulation results comparing g-formula and landmark approaches under misspecified outcome models.*

	MSE(Δk, k^*)			AUC(Δk, k^*)		
k^*	g-formula	landmark	landmark (lags)	g-formula	landmark	landmark (lags)
Scenario 1: Factual prediction						
0	**0.106**	0.115	0.115	**0.911**	**0.911**	**0.911**
	(0.006)	(0.007)	(0.007)	(0.010)	(0.010)	(0.010)
3	**0.091**	0.100	0.100	**0.930**	**0.930**	0.929
	(0.005)	(0.006)	(0.006)	(0.008)	(0.008)	(0.008)
6	**0.082**	0.088	0.088	**0.939**	0.938	0.938
	(0.005)	(0.005)	(0.005)	(0.008)	(0.008)	(0.008)
Scenario 2: Competing risk prediction						
0	**0.102**	0.106	0.106	**0.929**	0.928	0.928
	(0.006)	(0.007)	(0.007)	(0.013)	(0.013)	(0.013)
3	**0.096**	0.098	0.099	**0.947**	**0.947**	0.946
	(0.006)	(0.006)	(0.006)	(0.012)	(0.012)	(0.012)
6	**0.088**	0.091	0.091	**0.948**	**0.948**	0.946
	(0.005)	(0.006)	(0.006)	(0.011)	(0.011)	(0.011)
Scenario 3: Counterfactual prediction						
0	**0.148**	0.257	0.257	**0.958**	0.948	0.948
	(0.011)	(0.014)	(0.014)	(0.006)	(0.007)	(0.007)
3	**0.111**	0.187	0.185	**0.969**	0.968	0.967
	(0.009)	(0.010)	(0.010)	(0.005)	(0.005)	(0.005)
6	**0.088**	0.113	0.129	**0.971**	**0.971**	**0.971**
	(0.008)	(0.010)	(0.014)	(0.004)	(0.004)	(0.004)

Note:

All results based on 500 Monte Carlo simulations using data generation process described in section 3. Standard deviations of Monte Carlo estimates are provided in parentheses. The best performing estimator is shown in **bold**. All simulations use correctly specified models. For results under misspecification see the appendix.

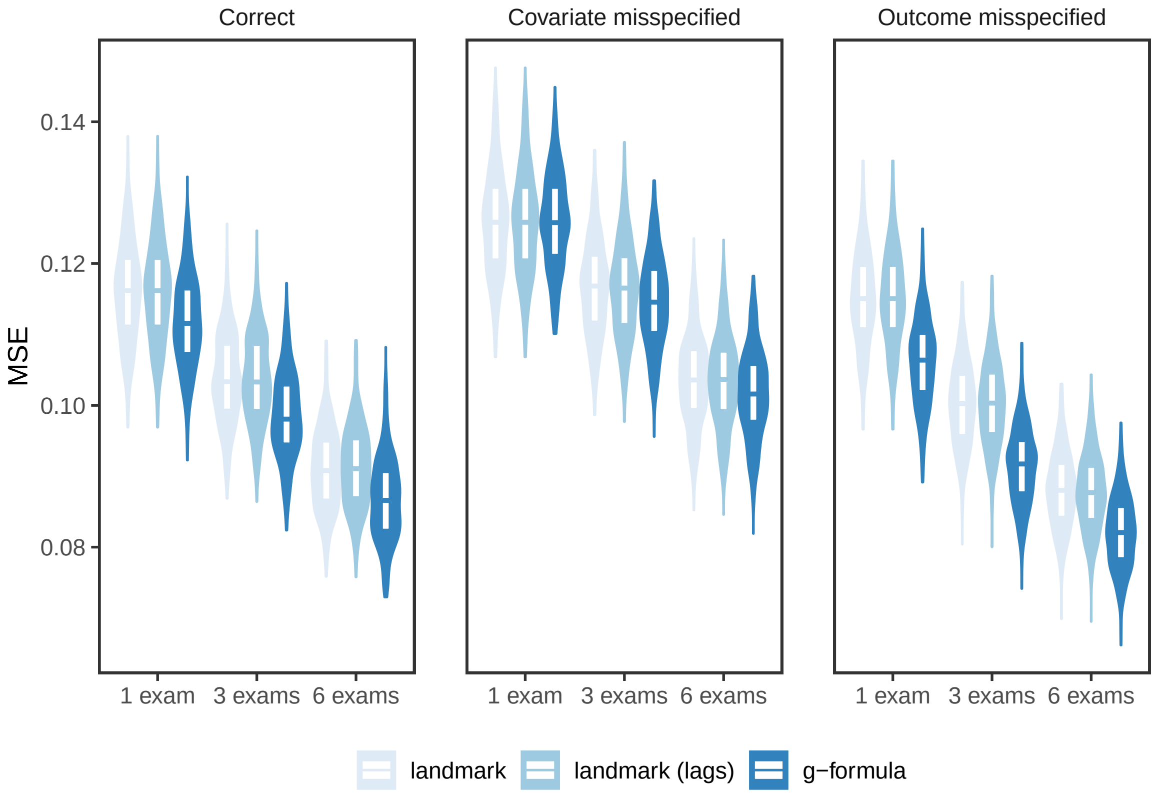


Figure A.3: *Simulation results - factual prediction MSE*

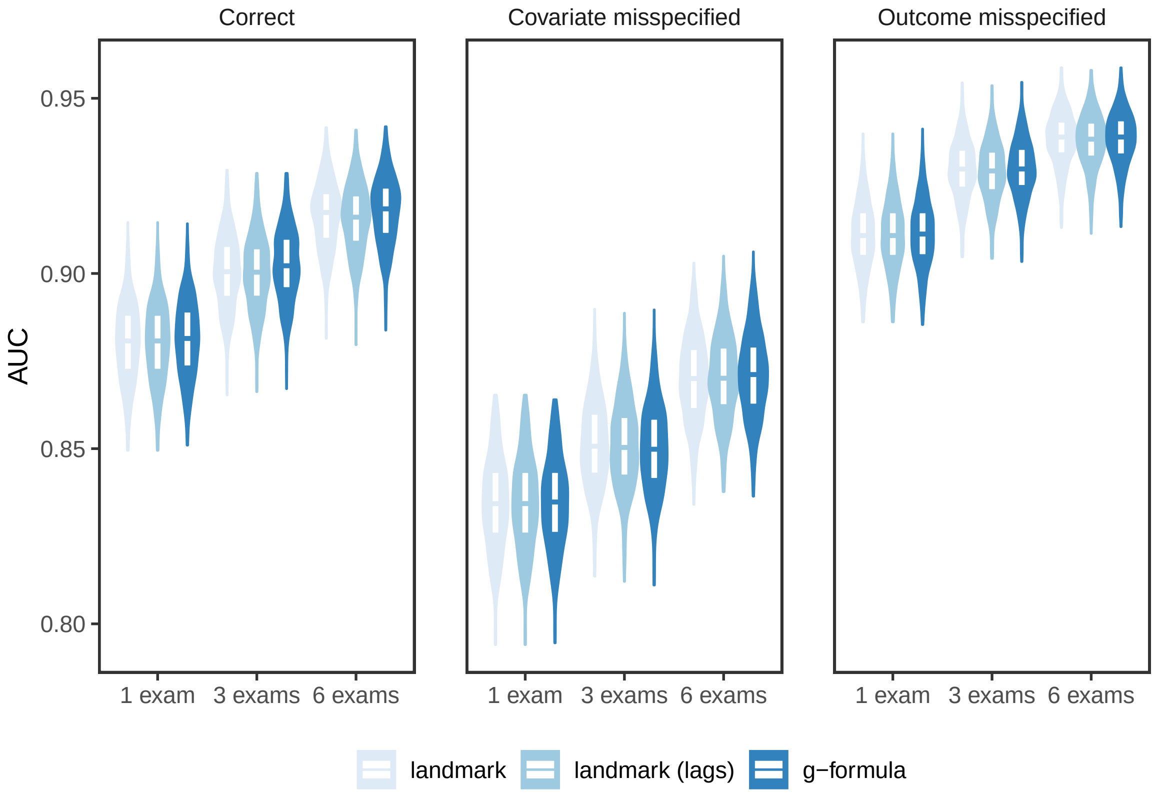


Figure A.4: *Simulation results - factual prediction AUC*

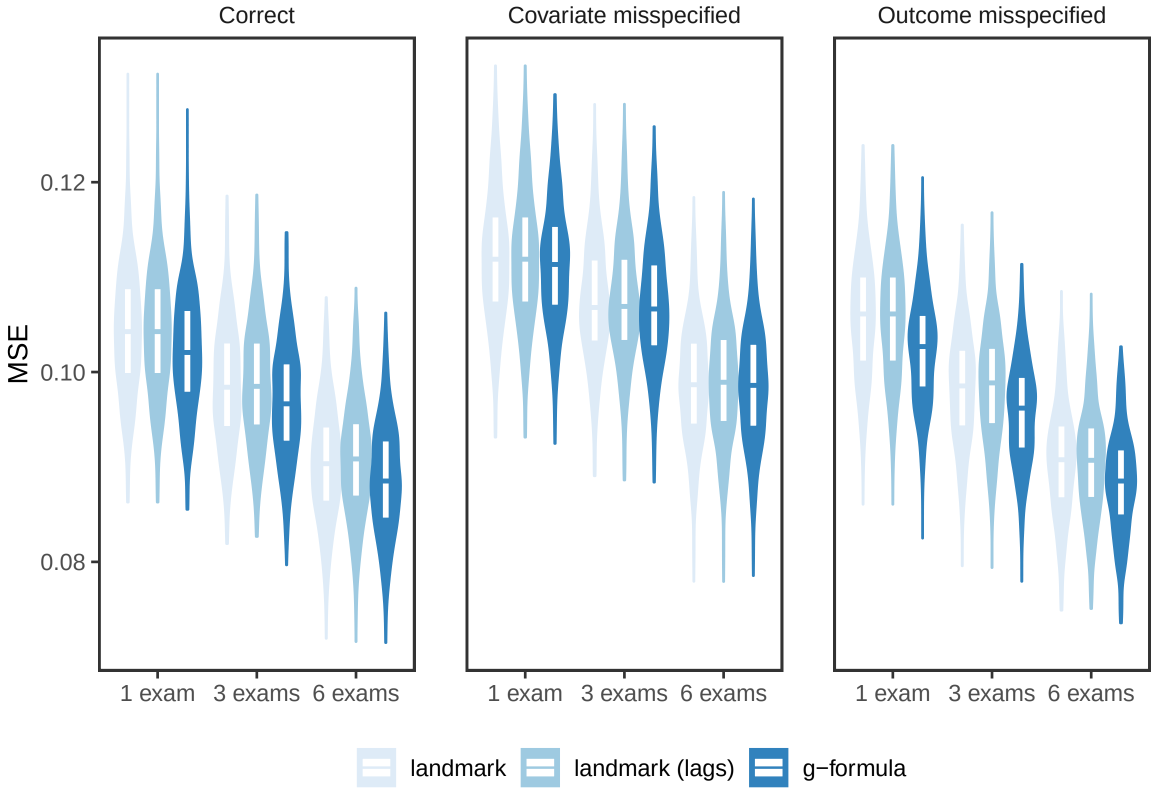


Figure A.5: *Simulation results - competing risk prediction MSE*

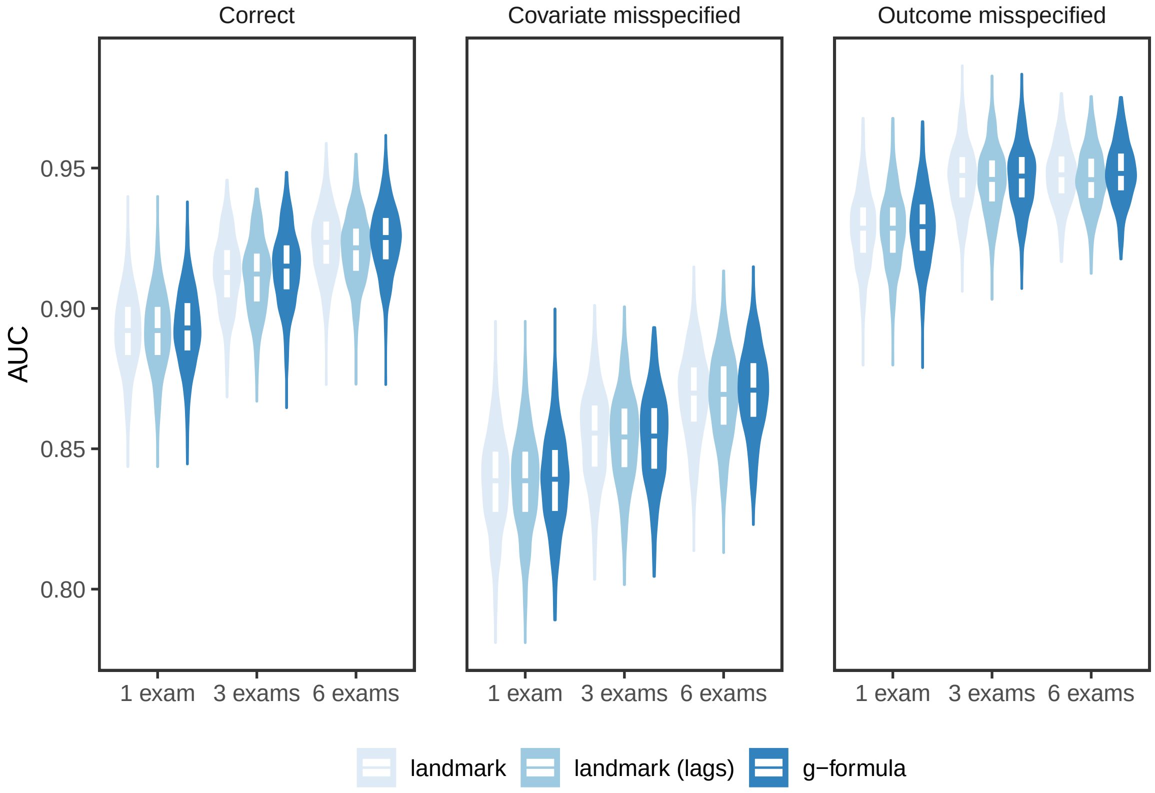


Figure A.6: *Simulation results - competing risk prediction AUC*

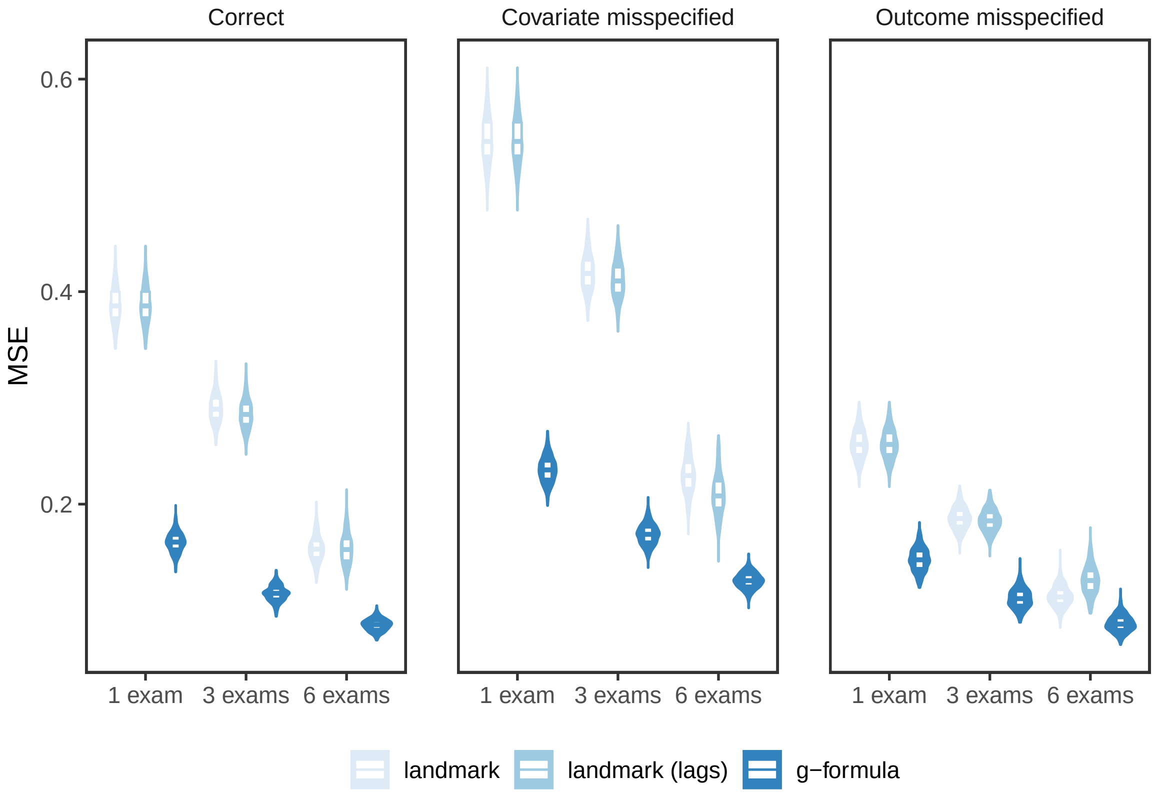


Figure A.7: *Simulation results - counterfactual prediction MSE*

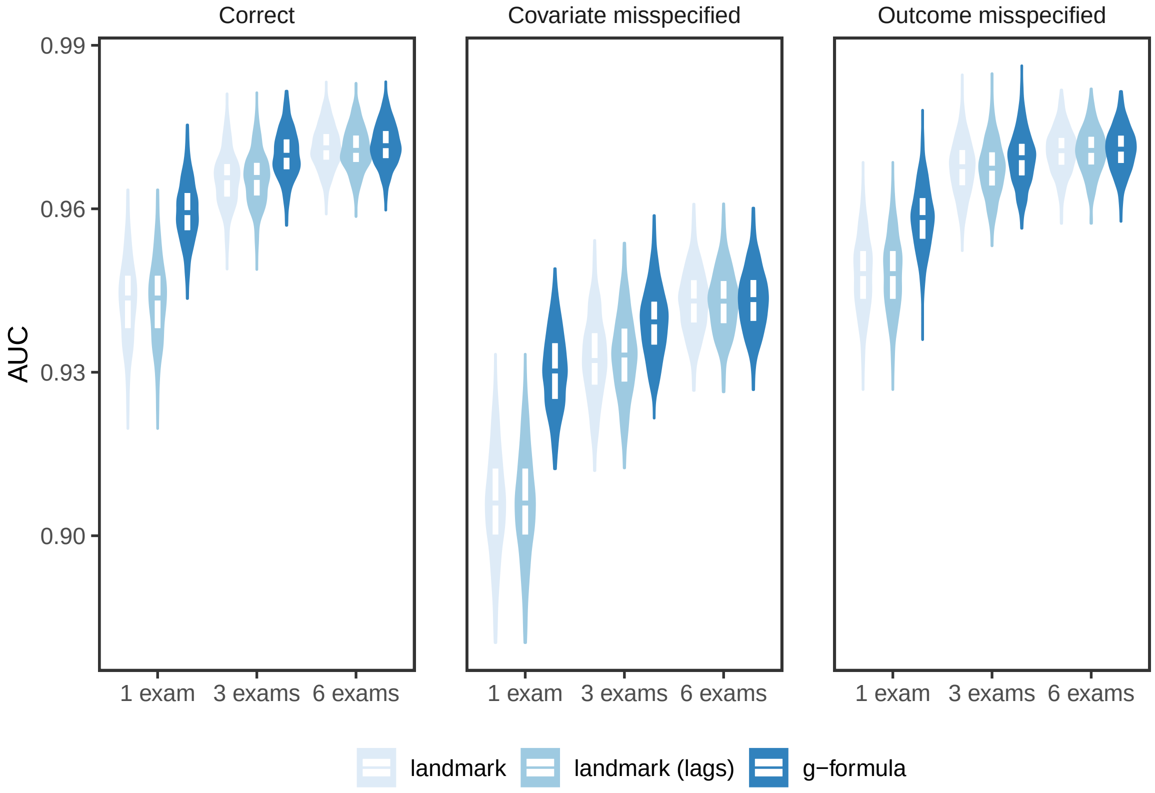


Figure A.8: *Simulation results - counterfactual prediction AUC*

A.4 Description of the Framingham Offspring Cohort data

Table A.3: *Descriptive summary of outcomes and time-varying covariates across examination cycles.*

Characteristic (Z)	Variable	5th exam (1991–1994)	6th exam (1994–1998)	7th exam (1998–2001)
Age, mean (SD)	age	55.2 (8.8)	59.0 (8.7)	61.7 (8.6)
Current smoker, (%)	smk	18.2	14.6	12.1
Body mass index (kg/m^2), mean (SD)	bmi	27.2 (4.8)	27.7 (5.0)	27.9 (5.1)
Diabetes mellitus, (%)	dm	5.6	7.7	9.0
Blood pressure medication, (%)	hrx	16.8	25.2	31.5
Lipid lowering medication, (%)	liprx	5.3	9.5	16.2
Total cholesterol (mg/dL), mean (SD)	tc	205.5 (35.7)	206.7 (36.4)	202.9 (35.0)
LDL cholesterol (mg/dL), mean (SD)	ldl	127.5 (32.9)	128.3 (32.8)	122.3 (31.7)
HDL cholesterol (mg/dL), mean (SD)	hdl	51.1 (14.9)	52.3 (15.9)	54.9 (16.6)
Systolic blood pressure (mmHg), mean (SD)	sbp	126.2 (18.3)	128.2 (18.1)	127.2 (18.1)
Coronary heart disease events (Y)	event_chd	54	57	55
Atherosclerotic cardiovascular disease events (Y)	event_ascvd	68	69	64
non-CHD deaths (D)	event_dth	42	38	49
non-ASCVD deaths (D)	event_dth_ascvd	39	37	46
Lost to follow up (C)	event_cen	0	73	2341

A.5 Parametric g-formula

A.5.1 Model definitions

To implement the g-formula, we used separate regressions to model:

- ASCVD (Y_k)
- non-ASCVD death (D_k)

and each of the following time-varying risk factors (X_k):

- cigarette smoking
- BMI
- diabetes
- anti-hyertension medication use
- lipid-lowering medication use
- serum LDL cholesterol
- serum HDL cholesterol
- systolic blood pressure

Specifications relating to parametric model choice and functional form are provided in Table A.4. We used pooled discrete-time logistic regression to model the probability of ASCVD and the probability of non-ASCVD death in each year.

Each time-varying risk factor was classed as binary, binary-to-failure, or continuous, and then modelled using a generalized linear model as specified in Table A.4. To increase efficiency all models were pooled over all examination cycles. All models included, as predictors, age, the two most recent values of all time-varying risk factors, and the fixed covariates baseline age and sex. We included product terms between lipid-lowering drugs and serum LDL and HDL cholesterol as well as between anti-hypertensive medications and

SBP. Binary predictors were entered into the models as indicators; continuous predictors were entered as polynomials (linear, quadratic and cubic) and restricted cubic splines in sensitivity analyses. tables A.5 and A.6 provide estimated model coefficient values and fit statistics for the outcome and covariate models respectively.

Because values of time-varying risk factors are assessed contemporaneously during the examination, we need to specify an ordering of the covariates to correctly model the joint distribution. Based on substantive knowledge of the biology and the examination process we elected to use the following ordering (also given in Table A.4): cigarette smoking, BMI, diabetes, anti-hypertension medication use, lipid-lowering medication use, serum total cholesterol, serum HDL cholesterol, and systolic blood pressure.

Based on the estimation procedure outlined in section 1.2.7, we estimate the 10-year cumulative risk using the Monte Carlo procedure outline in section 1.2.7 with 500 simulations per individual, or 1,297,500 total. We then took the mean of the probability of ASCVD across all simulations for each observation as its predicted probability of ASCVD.

Table A.4: *Model specifications for the parametric g-formula.*

	Variable	Order	Dependent variable: parametric model	Independent variable: functional form(s)
Time-fixed covariates (X_0)				
Sex	sex	-	Not predicted	Indicator
Age, (years)	age	-	Not predicted	Linear, cubic spline
Time-varying covariates (X_k)				
Current smoker, (%)	smk	1	Logistic	Indicator
Body mass index (kg/m^2)	bmi	2	Linear	Linear, cubic spline
Diabetes mellitus, (%)	dm	3	Logistic to failure	Indicator
Blood pressure medication, (%)	hrx	4	Logistic	Indicator
Lipid lowering medication, (%)	liprx	5	Logistic	Indicator
LDL cholesterol, (mg / dL)	ldl	6	Linear	Linear, cubic spline
HDL cholesterol, (mg / dL)	hdl	7	Linear	Linear, cubic spline
Systolic blood pressure, (mmHg)	sbp	8	Linear	Linear, cubic spline
Outcomes				
Coronary heart disease events (Y)	event_ascvd	-	Logistic to failure	-
Deaths due to other causes (D)	event_dth	-	Logistic to failure	-
Censored due to loss to follow up (C)	event_cens	-	Logistic to failure	-

A.5.2 Model fits

Table A.5: *Outcome, competing event, and censoring event regression models*

	ASCVD	Death	Censor
(Intercept)	−9.147***	−10.191***	−7.736***
	(1.035)	(1.269)	(1.586)
sex	−0.561***	−0.758***	0.644**
	(0.165)	(0.204)	(0.233)
age	0.044***	0.106***	−0.001
	(0.010)	(0.012)	(0.013)
dm	0.545	−0.304	0.161
	(0.549)	(0.871)	(1.534)
hrx	0.344	0.564	−1.578
	(1.123)	(1.512)	(1.154)
liprx	−0.517	0.673	−13.741
	(1.581)	(2.081)	(557.843)
smk	1.147*	0.576	1.490
	(0.537)	(0.682)	(1.650)
bmi	−0.070	−0.010	−0.225
	(0.044)	(0.056)	(0.169)
ldl	0.005	−0.009	0.006
	(0.005)	(0.007)	(0.014)
hdl	−0.018	0.003	−0.011
	(0.014)	(0.015)	(0.029)
sbp	0.023**	−0.004	0.008
	(0.008)	(0.011)	(0.021)
log(year + 1)	0.217	0.381+	
	(0.159)	(0.200)	
lag1_dm	0.471	1.016	−0.825
	(0.565)	(0.877)	(1.563)
lag1_hrx	0.353	0.428	1.084
	(0.405)	(0.561)	(1.154)
lag1_liprx	1.036	1.874	−12.358
	(0.816)	(1.160)	(597.519)
lag1_smk	−0.678	0.368	−0.760
	(0.561)	(0.695)	(1.650)
lag1_bmi	0.067	−0.042	0.190
	(0.044)	(0.056)	(0.168)
lag1_ldl	0.001	0.005	−0.009
	(0.006)	(0.007)	(0.014)
lag1_hdl	−0.003	0.002	−0.027
	(0.015)	(0.015)	(0.029)
lag1_sbp	−0.011	0.002	0.004
	(0.008)	(0.011)	(0.022)
liprx × ldl	−0.014	−0.023	
	(0.010)	(0.015)	
liprx × hdl	0.024	−0.004	
	(0.021)	(0.027)	
hrx × sbp	−0.001	−0.006	
	(0.007)	(0.011)	
I(year == 8)			5.295***
			(0.825)
year			0.099
			(0.200)
Num.Obs.	26 767	26 890	26 890
RMSE	0.08	0.07	0.06

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.6: *Covariate history regression models*

	dm	smk	bmi	hrx	liprx	ldl	hdl	sbp
(Intercept)	−9.763***	−4.965***	0.580***	−13.700***	−8.246***	9.640***	3.396***	5.307***
	(1.074)	(0.849)	(0.061)	(1.379)	(0.853)	(0.803)	(0.313)	(0.509)
sex	0.013	0.094	0.023+	0.133	0.140	0.347*	0.610***	−0.016
	(0.195)	(0.156)	(0.012)	(0.100)	(0.123)	(0.155)	(0.060)	(0.098)
age	0.015	−0.028**	−0.003***	0.005	0.014*	−0.033***	−0.004	0.066***
	(0.011)	(0.009)	(0.001)	(0.006)	(0.007)	(0.009)	(0.003)	(0.006)
lag1_smk	0.292	9.560***	1.221***	1.358***	−0.871+	−3.698***	1.781***	−0.073
	(0.231)	(0.202)	(0.057)	(0.282)	(0.494)	(0.756)	(0.294)	(0.477)
lag1_bmi	0.100***	0.006	0.995***	−0.256***	−0.335***	−2.177***	0.714***	−1.546***
	(0.014)	(0.016)	(0.001)	(0.039)	(0.048)	(0.085)	(0.034)	(0.055)
lag1_hrx	0.131	0.059	0.014	12.764***	−1.924***	−1.126*	−0.891***	7.877***
	(0.205)	(0.220)	(0.014)	(0.965)	(0.177)	(0.488)	(0.190)	(0.879)
lag1_liprx	−0.573	−0.014	−0.021	−0.421+	15.072***	29.254***	−2.206+	0.406
	(0.469)	(0.377)	(0.026)	(0.239)	(1.187)	(2.967)	(1.155)	(0.448)
lag1_ldl	0.002	−0.001	0.000**	−0.001	0.034***	0.948***	−0.007**	0.002
	(0.003)	(0.002)	(0.000)	(0.001)	(0.004)	(0.002)	(0.003)	(0.004)
lag1_hdl	−0.041***	0.002	0.000	−0.014***	−0.034***	−0.025***	0.966***	−0.162***
	(0.008)	(0.005)	(0.000)	(0.004)	(0.005)	(0.005)	(0.002)	(0.011)
lag1_sbp	0.012*	0.000	−0.001**	0.075***	−0.006+	0.010*	−0.003	0.924***
	(0.005)	(0.004)	(0.000)	(0.011)	(0.003)	(0.004)	(0.002)	(0.003)
log(year + 1)	0.578**							
	(0.193)							
dm		−1.627**	0.500***	1.668***	0.612	−0.573	−1.973***	0.369
		(0.504)	(0.065)	(0.283)	(0.375)	(0.853)	(0.331)	(0.538)
year		−0.050+	−0.001	0.061***	0.134***	−0.113***	0.057***	−0.027
		(0.029)	(0.002)	(0.018)	(0.024)	(0.028)	(0.011)	(0.018)
lag1_dm		1.675**	−0.602***	−0.962**	−0.193	0.416	1.828***	−0.027
		(0.570)	(0.067)	(0.307)	(0.401)	(0.879)	(0.341)	(0.554)
smk			−1.240***	−1.640***	0.571	3.189***	−1.850***	−0.065
			(0.058)	(0.303)	(0.506)	(0.767)	(0.298)	(0.484)
bmi				0.265***	0.317***	2.154***	−0.744***	1.568***
				(0.038)	(0.046)	(0.084)	(0.033)	(0.054)
I(lag1_sbp ≥ 130)				6.622***				
				(1.403)				
lag1_sbp × I(lag1_sbp ≥ 130)				−0.048***				
				(0.011)				
lag1_hrx × lag1_sbp				−0.034***				
				(0.007)				
hrx					2.903***	0.977*	0.890***	−6.262***
					(0.177)	(0.479)	(0.186)	(0.322)
I(lag1_ldl ≥ 160)					3.013***			
					(0.857)			
lag1_ldl × I(lag1_ldl ≥ 160)					−0.017**			

A.5.3 Model Diagnostics

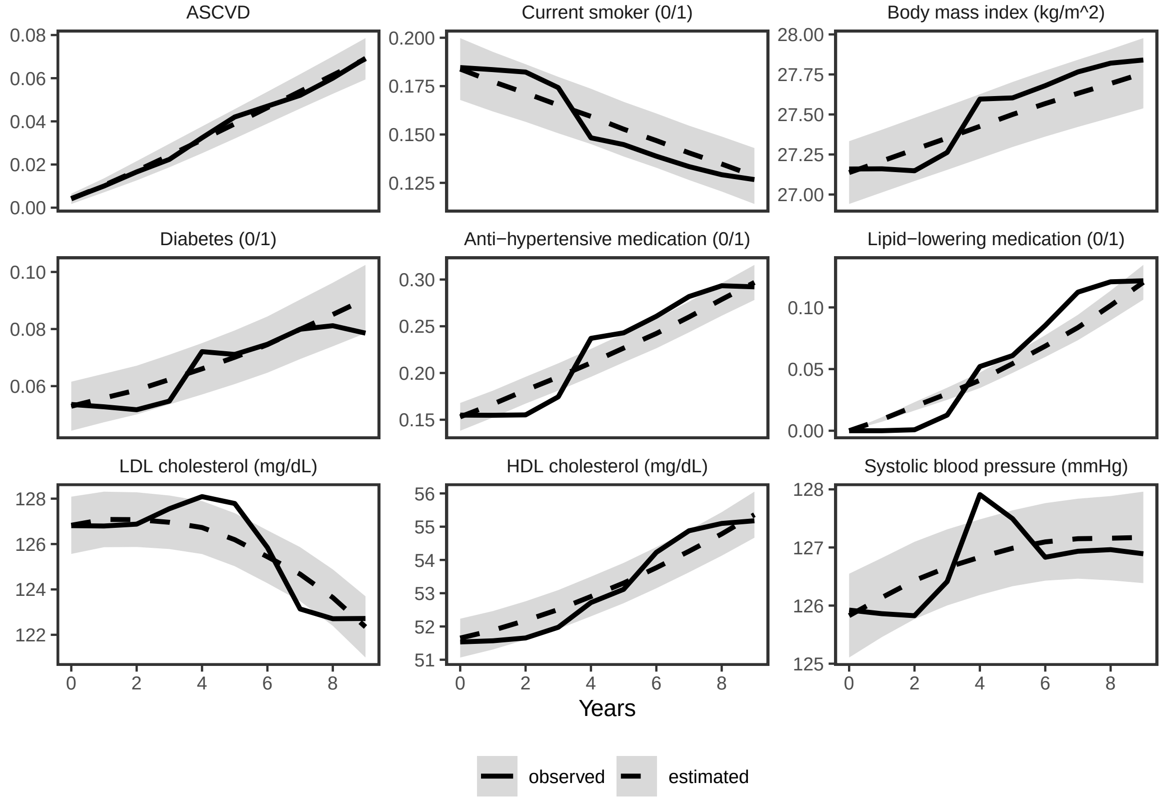


Figure A.9: *Model specification tests for the parametric g-formula - one lag specification. A comparison of estimated (dotted line) and observed (solid line) means under the natural course for outcome and covariate models with 90% confidence intervals obtained using the nonparametric bootstrap.*

A.6 Comparison models

As comparators we considered de-novo developed landmark Cox proportional hazards models at baseline ($k^* = 0$), after four years ($k^* = 3$), and after seven years ($k^* = 6$), i.e.

$$h(t \mid X_{k^*}) = h_0(t) \exp\{\beta^t X_{k^*}\},$$

where X_{k^*} include same covariate set as used in fitting the g-formula. We included both version with lagged values of covariates and without. Models included the following risk factors (X_{k^*}) measured at the landmark time:

- cigarette smoking
- BMI
- diabetes
- anti-hyertension medication use
- lipid-lowering medication use
- serum LDL cholesterol
- serum HDL cholesterol
- systolic blood pressure

Lagged models included previous values before time k^*. As in the g-formula models we included product terms between lipid-lowering drugs and serum LDL and HDL cholesterol as well as between anti-hypertensive medications and SBP. The 10-year cumulative incidence was calculated using Breslow estimates of the baseline hazard. Fitted model coefficients are shown in Table A.7.

Table A.7: *Model fits for landmark Cox regressions.*

	landmark			landmark (lags)		
	k* = 0	k* = 3	k* = 6	k* = 0	k* = 3	k* = 6
sex	−0.546***	−0.486**	−0.329	−0.546***	−0.499**	−0.310
	(0.163)	(0.188)	(0.250)	(0.163)	(0.188)	(0.251)
age	0.045***	0.046***	0.036*	0.045***	0.047***	0.036*
	(0.009)	(0.011)	(0.015)	(0.009)	(0.011)	(0.015)
dm	0.889***	0.726**	0.820**	0.889***	−1.006	−0.296
	(0.203)	(0.239)	(0.295)	(0.203)	(0.971)	(0.988)
hrx	0.769	0.368	0.695	0.769	0.194	0.853
	(1.144)	(1.284)	(1.631)	(1.144)	(1.381)	(1.694)
smk	0.407*	0.364+	0.628*	0.407*	−0.661	0.133
	(0.180)	(0.214)	(0.300)	(0.180)	(0.769)	(1.032)
bmi	0.006	0.005	−0.010	0.006	−0.013	0.030
	(0.016)	(0.018)	(0.024)	(0.016)	(0.106)	(0.117)
ldl	0.004+	0.005+	0.005	0.004+	0.001	0.005
	(0.002)	(0.003)	(0.004)	(0.002)	(0.007)	(0.004)
hdl	−0.018**	−0.019**	−0.021*	−0.018**	−0.026	−0.021*
	(0.007)	(0.007)	(0.010)	(0.007)	(0.019)	(0.010)
sbp	0.013**	0.016**	0.006	0.013**	0.028*	0.007
	(0.005)	(0.005)	(0.009)	(0.005)	(0.011)	(0.009)
hrx × sbp	−0.003	−0.002	0.000	−0.003	−0.003	0.000
	(0.008)	(0.009)	(0.012)	(0.008)	(0.009)	(0.012)
liprx		−0.373	0.535		−0.595	0.284
		(0.721)	(1.880)		(0.779)	(1.977)
liprx × ldl			−0.010			−0.009
			(0.012)			(0.012)
liprx × hdl			0.013			0.015
			(0.028)			(0.029)
lag1_dm					1.841+	1.207
					(0.979)	(0.985)
lag1_hrx					0.403	−0.147
					(0.512)	(0.564)
lag1_smk					1.028	0.525
					(0.758)	(1.007)
lag1_bmi					0.020	−0.042
					(0.108)	(0.120)
lag1_ldl					0.004	
					(0.007)	
lag1_hdl					0.007	
					(0.019)	
lag1_sbp					−0.014	
					(0.011)	
lag1_liprx						0.164
						(0.746)
Num.Obs.	2828	2750	2648	2828	2750	2648
AIC	2915.3	2230.0	1197.2	2915.3	2237.3	1205.0
RMSE	0.26	0.23	0.17	0.26	0.23	0.17

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

A.7 Counterfactual prediction results

Table A.8: *Validation of model performance for counterfactual prediction of 'treatment-naive' risk in Framingham Heart Study.*

k^*	Model	MSE(Δk, k^*)	AUC(Δk, k^*)
0	g-formula	**0.1062**	**0.734**
	landmark	0.1088	0.714
	landmark (lags)	0.1088	0.714
3	g-formula	**0.0950**	**0.725**
	landmark	0.0965	0.701
	landmark (lags)	0.0970	0.700
6	g-formula	**0.0625**	**0.689**
	landmark	0.0633	0.655
	landmark (lags)	0.0637	0.653

A.8 Code testing and validation

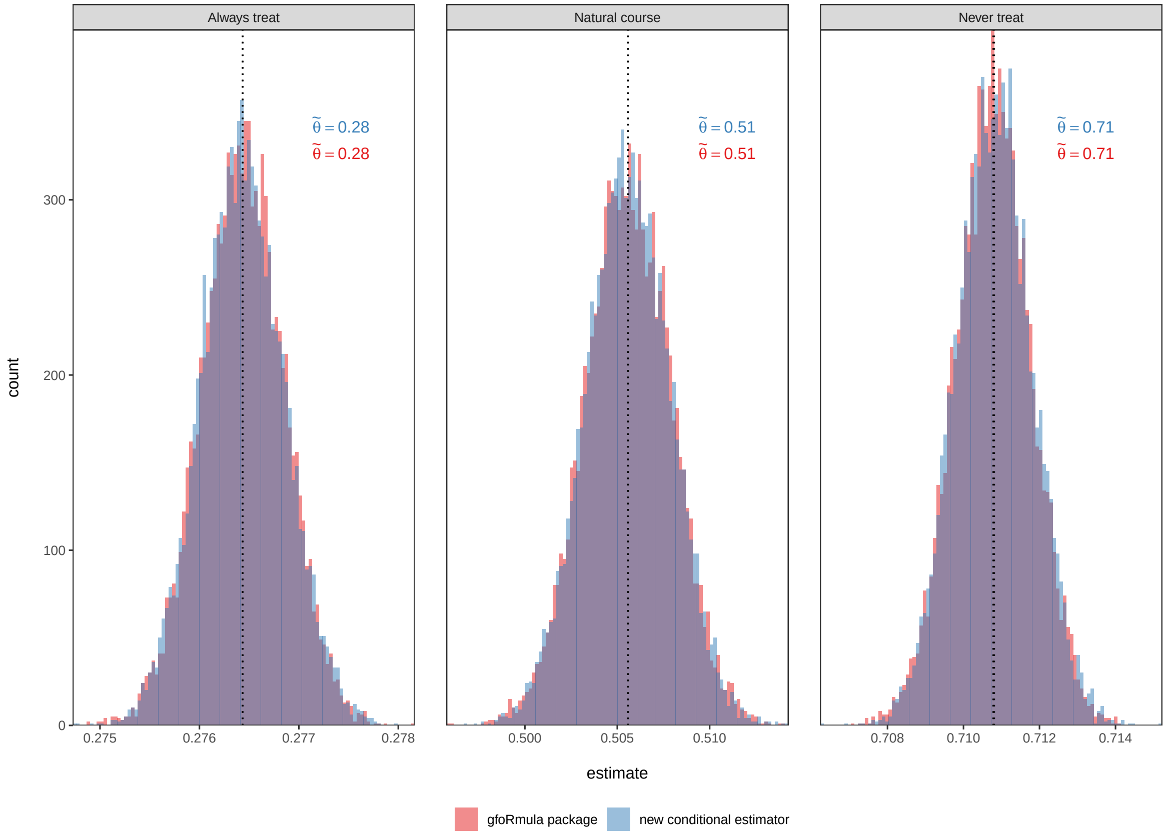


Figure A.10: *Comparison of new conditional g-formula and* `gfoRmula` *R package to validate coding.*

Appendix B

Appendix to Chapter 2

B.1 Loss to follow up

When there is censoring due to loss to follow up, a natural prediction target is the counterfactual risk if everyone remained untreated and no one was lost to follow up, i.e.

$$\Pr(T^{\bar{a}=0,\bar{c}=0} \leq t \mid X^*).$$

In this case, the treatment-naive risk is identified only if censoring is non-informative given past treatment and covariate history. Therefore, we require the following modified set of identification conditions

1. *Sequential Exchangeability:*

$$T^{\bar{a}=0,\bar{c}=0} \perp\!\!\!\perp (A_k, C_{k+1}) \mid \overline{X}_k, \overline{A}_{k-1}, \overline{Y}_k = \overline{C}_k = 0$$

2. *Consistency:*

$$T = T^{\bar{a}=0,\bar{c}=0},\ \overline{Y}_{k+1} = \overline{Y}_{k+1}^{\bar{a}=0,\bar{c}=0},\ \text{and } \overline{X}_k = \overline{X}_k^{\bar{a}=0,\bar{c}=0} \text{ if } \overline{A}_k = 0 \text{ and } \overline{C}_{k+1} = 0$$

and one of

3a. *Positivity:* $\Pr(A_k = 0, C_{k+1} = 0 \mid \overline{X}_k, \overline{A}_{k-1} = 0, \overline{Y}_k = \overline{C}_k = 0) > 0$

3b. *Known semi-parametric model:* $T^{\overline{a}=0,\overline{c}=0}$ follows a SNAFTM.

We also make the following modifications to the g-estimation and IPCW procedures described in the main text. When estimating ψ using estimating equations we solve

$$\sum_{i=1}^{N}\sum_{k=0}^{K} W_i(k)H_i(k,\psi^*)[A_{i,k} - \Pr\{A_k = a_k \mid \overline{X}_k, \overline{A}_{k-1}, \overline{Y}_k = \overline{C}_k = 0\}] = 0$$

where $W_i(k)$ are inverse probability weights for the probability of being uncensored through time k, i.e.

$$W_i(k) = \prod_{k=0}^{K} \frac{I(C_{k+1} = 0)}{\Pr(C_{k+1} = 0 \mid \overline{X}_k, \overline{A}_{k-1} = 0, \overline{Y}_k = \overline{C}_k = 0)}.$$

Similarly, when ψ is estimated using a manual grid search based on score statistic of a term from a pooled logistic regression with $H(k,\psi^*)$ as a covariate, we can use weighted logistic regression with weights $W_i(k)$ above.

For the IPCW method, we modify the existing weights to be

$$W_c = \prod_{k=0}^{K} \frac{I(A_k = 0, C_k = 0)\Pr(A_k = 0 \mid X^*, \overline{A}_{k-1} = 0, \overline{Y}_k = \overline{C}_k = 0)}{\Pr(A_k = 0 \mid \overline{X}_k, \overline{A}_{k-1} = 0, \overline{Y}_k = \overline{C}_k = 0)\Pr(C_{k+1} = 0 \mid \overline{X}_k, \overline{A}_{k-1} = 0, \overline{Y}_k = \overline{C}_k = 0)}$$

where the denominator is now the product of the probability of remaining untreated through time k and the probability of remaning uncensored through time $k+1$ conditional on past treatment and covariate history.

B.2 Competing events

Here we modify our original set up slightly, we now observe i.i.d. samples of

$$O_i = (\overline{X}_k, \overline{A}_k, \overline{C}_{k+1}, \overline{D}_{k+1}, \overline{Y}_{k+1}, T)$$

where D_{k+1} is an indicator of a competing event at time $k+1$ and the other variables are defined as previously. By definition $C_0 \equiv 0$, $D_0 \equiv 0$, and $Y_0 \equiv 0$ as we restrict to those who are uncensored and event-free at the start of follow up. By convention, all subsequent variables are zero when when any of $Y_{k+1} = 1$, $C_{k+1} = 1$, or $D_{k+1} = 1$. An example directed acyclic graph for a two time point process with the addition of competing events is shown

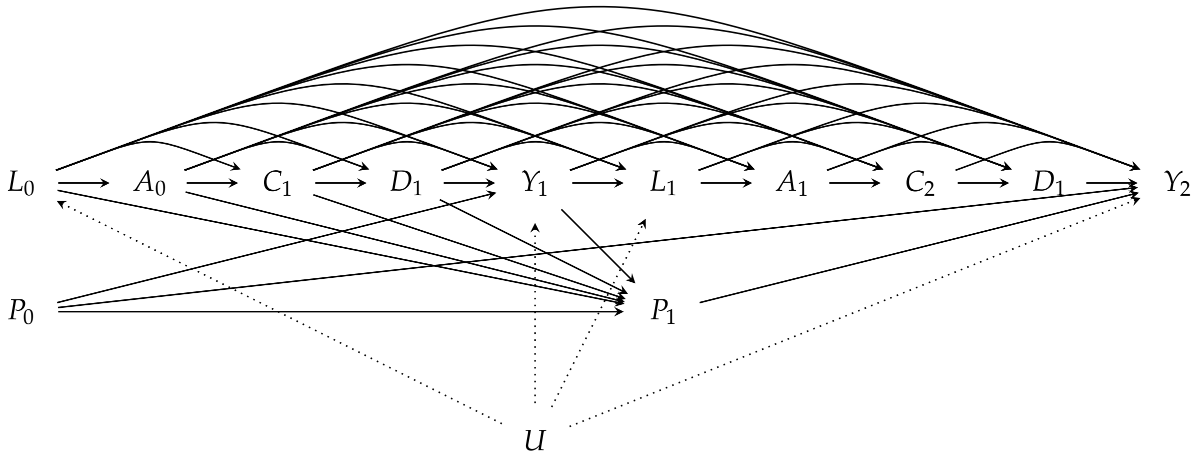


Figure B.1: *Example directed acyclic graph for survival outcome with competing event D_k.*

in Figure C.1.

In the presence of competing events, there are a number of different prediction estimands of interest. On the one hand, we might be interested in the treatment-naive risk under the elimination of competing events, e.g.

$$\Pr(T^{\overline{a}=\overline{c}=\overline{d}=0} \leq t \mid X^*).$$

However, in many cases, it's unclear whether such an intervention to remove the competing event is feasible or desirable, such as in our application to cardiovascular disease in section 4 where the competing event is non-ASCVD death. Alternatively, we might instead be interested in which non-ASCVD deaths were not eliminated, but rather occurred at the counterfactual rate if no one were treated, i.e.

$$\Pr(T^{\overline{a}=0,\overline{c}=0} \leq t \mid X^*).$$

In the main text, we target the latter, which requires the following identification assumptions

1. *Sequential Exchangeability:*

$$T^{\bar{a}=0,\bar{c}=0} \perp\!\!\!\perp (A_k, C_{k+1}) \mid \overline{X}_k, \overline{A}_{k-1}, \overline{Y}_k = \overline{C}_k = \overline{D}_k = 0$$

2. *Consistency:*

$$T = T^{\bar{a}=0,\bar{c}=0},\ \overline{Y}_{k+1} = \overline{Y}_{k+1}^{\bar{a}=0,\bar{c}=0},\ \overline{D}_{k+1} = \overline{D}_{k+1}^{\bar{a}=0,\bar{c}=0},\text{ and } \overline{X}_k = \overline{X}_k^{\bar{a}=0,\bar{c}=0}$$
$$\text{if } \overline{A}_k = 0 \text{ and } \overline{C}_{k+1} = 0$$

and one of

3a. *Positivity:* $\Pr(A_k = 0, C_{k+1} = 0 \mid \overline{X}_k, \overline{A}_{k-1} = 0, \overline{Y}_k = \overline{C}_k = \overline{D}_k = 0) > 0$

3b. *Known semi-parametric model:* $T^{\bar{a}=0,\bar{c}=0}$ follows a SNAFTM.

To estimate the treatment-naive risk under competing events, we modify our g-estimation and IPCW procedures to target the subdistribution hazard as follows. In our dataset, whenever a competing event occurs we artificially set all future values of $\overline{X}_k$, $\overline{A}_k$, $\overline{C}_{k+1}$, and $\overline{Y}_{k+1}$ to zero and then adjust the inverse probability weights for both approaches described in section A.1. to reflect the deterministic fact that when $D_{k+1} = 1$

$$\Pr(C_{j+1} = 0 \mid \overline{X}_j, \overline{A}_{j-1} = 0, \overline{Y}_j = \overline{C}_j = 0) = 1$$

and

$$\Pr(A_j = 0 \mid \overline{X}_j, \overline{A}_{j-1} = 0, \overline{Y}_j = \overline{C}_j = 0) = 1$$

for $j > k$ where the first applies to both approaches and the latter to only the IPCW approach.

B.3 Benchmarking emulation of a statin trial

In the main text, our goal was to emulate a single arm trial in the MESA cohort and then use the proposed methods to develop a prediction model for the statin-naive risk. These methods crucially rely on the identification conditions given in section 2.2.3 which may not hold in many observational settings. While these assumptions cannot be empirically evaluated, here we describe a benchmarking exercise which may at least help us determine whether they are massively violated.

Statin therapy has been extensively studied across dozens of large randomized trials. We emulated a standard two arm trial comparing statin initiation to control and benchmark our findings against those from previous trials. While average effects may vary from trial to trial for reasons such as changes in distribution of effect modifiers, in theory if our emulation fell well outside the bounds of previous trials it may suggest residual confounding or other violations are present which would inhibit the development of a statin-naive model from these data.

We emulated a randomized trial corresponding to the AHA guidelines on initiation of statin therapy for primary prevention of cardiovascular disease in the MESA cohort. An example protocol is shown in Table X. The AHA guidelines stipulate that patients aged 40 to 75 with serum LDL cholesterol levels between 70 mg/dL and 190 mg/dL and no history of cardiovascular disease should initiate statins if their risk exceeds 7.5%. Therefore, we considered MESA participants who completed the baseline examination, had no recent history of statin use, no history of cardiovascular disease, and who met the criteria described in the guidelines (excluding the risk threshold) as eligible to participate in the trial. The primary endpoint was time to atherosclerotic cardiovascular disease (ASCVD), defined as nonfatal myocardial infarction, coronary heart disease death, or ischemic stroke.

Follow up began at the second examination cycle to enable a "wash out" period for statin use and to ensure adequate pre-treatment covariates to control confouding. We constructed a sequence of nested trials starting at each examination cycle from exam 2 through exam 5 and pooled the results from all 4 trials into a single analysis and used a robust variance estimator to account for correlation among duplicated participants. In each nested trial, we used the corresponding questionnaire to determine eligibility as well as statin initiators versus non-initiators. Because the exact timing of statin initiation was not known with precision, in each trial, we estimated the start of follow up for initiators and non-initators by drawing a random month between their current and previous examinations. We explored

Table B.1: *Protocol for the specification and emulation of a target trial of statin therapy initiation strategies in the MESA cohort.*

Protocol component	Target trial specification	Emulation
Eligibility	Age 40 to 75 years No prior statin use No history of ASCVD LDL-C $>$ 70 mg/dL LDL-C $<$ 190 mg/dL	same
Treatment strategies	(1) initiation of statins within 3 months of baseline randomization (2) no initiation of statins over follow up	same
Treatment assignment	non-blinded random assignment to either (1) or (2) at baseline	same but randomization is emulated conditional on covariates necessary to control confounding
Outcomes	cumulative incidence of ASCVD defined as nonfatal myocardial infarction, coronary heart disease death, or ischemic stroke	same
Follow up	Start at baseline and follow until ASCVD event, non-ASCVD death, or until 10 years have elapsed, whichever happens first	same but exact starting time was estimated from time of questionnaire return
Statistical analysis	*ITT* - compare cumulative incidence of ASCVD under each strategy, adjusting for prognostic factors to increase efficiency *Per protocol* - Use IPW/g-estimation to account for time-varying non-adherence.	same but additionally emulating baseline randomization conditional on covariates

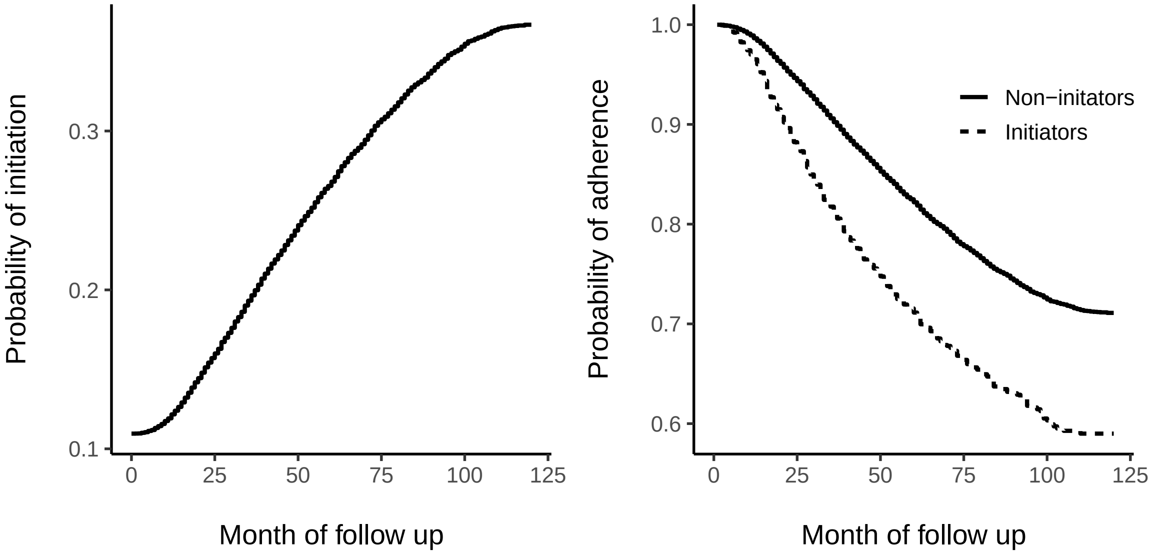


Figure B.2: *Probability of statin initiation and probability of adherence among initiators and non-initiators in nested target trial emulation, Multi-Ethnic Study of Atherosclerosis, 2000 to 2010.*

alternative definitions of the start of follow up in sensitivity analyses in section B.4.

For intention-to-treat (ITT) analyses, we estimated the effects of statin initiation by comparing initiators versus non-initiators as defined at baseline, regardless of whether they adhered to therapy or no therapy throughout follow up. We used pooled logistic regression in nested trial data to estimate ITT hazard ratios conditional on covariates listed in X.

We also considered adherence-adjusted analyses comparing always treat versus never treat. To do so we censored initiators and non-initiators when they deviated from their baseline regime and adjusted for time-varying confounding using inverse probability of censoring weighting and g-estimation.

Of the 6,814 MESA participants who completed the baseline examination, 4,149 met the eligibility criteria for our trial emulation. There were 288 ASCVD events, 190 non-ASCVD deaths, and 414 were lost over the 10 year follow up period. In the nested trial dataset, there were 1,592 initiators and 12,767 non-initiators. Table 1 shows the baseline characteristics of initiators and non-initiators of statins in the emulated nested trials.

Figure B.2 shows the probability of statin initation over the follow up period. After ten years approximately 40% of MESA participants had initiated statins. The 5-year estimated

Table B.2: *Baseline characteristics of initiators and non-initiators in emulated nested trials*

	Initiators (N = 1,592)	**Non-initiators (N = 12,767)**
Demographics		
Age, years	65.1 (8.2)	62.5 (8.8)
Male, %	45.1	46.8
Married, %	64.6	63.3
Less than high school, %	18.2	15.0
High school graduate, %	48.2	44.9
College or postgraduate, %	33.5	39.8
Non-Hispanic white, %	40.1	37.9
Non-Hispanic black, %	22.9	21.9
Hispanic, %	26.1	27.4
Asian, %	11.0	12.9
Currently employed, %	51.4	59.2
Retired, %	33.3	26.8
No health insurance, %	3.6	8.2
CES Depression scale (0-60)	7.4 (7.5)	7.5 (7.5)
Chronic burden scale (0-5)	1.1 (1.2)	1.1 (1.2)
Perceived discrimination scale (0-4)	0.1 (0.4)	0.1 (0.4)
Emotional support scale (0-30)	24.3 (5.1)	24.1 (5.3)
Everyday hassles scale (0-54)	14.4 (6.0)	15.2 (6.2)
Spielberger trait anger scale (0-40)	15.0 (3.8)	15.0 (3.7)
Spielberger trait anxiety scale (0-40)	15.9 (4.5)	16.0 (4.5)
Neighborhood problems scale (0-28)	10.4 (3.4)	10.5 (3.4)
CVD risk factors		
Systolic blood pressure, mmHg	125.6 (19.7)	122.0 (20.2)
Diastolic bood pressure, mmHg	71.3 (10.2)	71.1 (10.1)
LDL cholesterol, mg/dL	135.4 (31.7)	119.7 (27.6)
HDL cholesterol, mg/dL	50.2 (13.9)	52.3 (15.1)
Triglycerides, mg/dL	147.5 (87.7)	120.5 (66.1)
Baseline ASCVD risk, %	10.1	7.6
Diabetes mellitus, %	38.4	23.0
Hypertension, %	51.9	36.2
Waist circumference, cm	99.6 (14.3)	96.9 (14.7)
Smoked <100 cigarettes in lifetime, %	49.6	48.8
Current smoker, %	10.7	12.9
Drinks per week	2.9 (5.9)	3.4 (7.6)
Exercise, MET/min	1471.6 (2187.1)	1509.3 (2187.7)
Family history of CVD, %	58.9	53.5
Calcium score	124.2 (340.7)	72.4 (256.2)
Left ventricular hypertrophy on ECG, %	1.0	0.8
C-reactive protein, mg/dL	4.3 (6.0)	3.5 (5.0)
Interleukin-6, pg/mL	1.5 (1.2)	1.4 (1.2)
Number of pregnancies	3.1 (2.2)	3.0 (2.3)
Years on birth control pills	3.6 (5.8)	3.7 (5.8)
Age at menopause, years	41.1 (17.5)	37.2 (20.4)
Medications		
Anti-hypertensive medication, %	61.7	34.7
Insulin or oral hypoglycemics, %	22.7	8.6
Daily aspirin use, %	47.2	25.0
Diuretics, %	21.4	12.5
Any anti-depressants, %	11.7	7.8
Any vasodilator, %	3.8	3.3
Any anti-arrhytmic, %	0.6	0.6

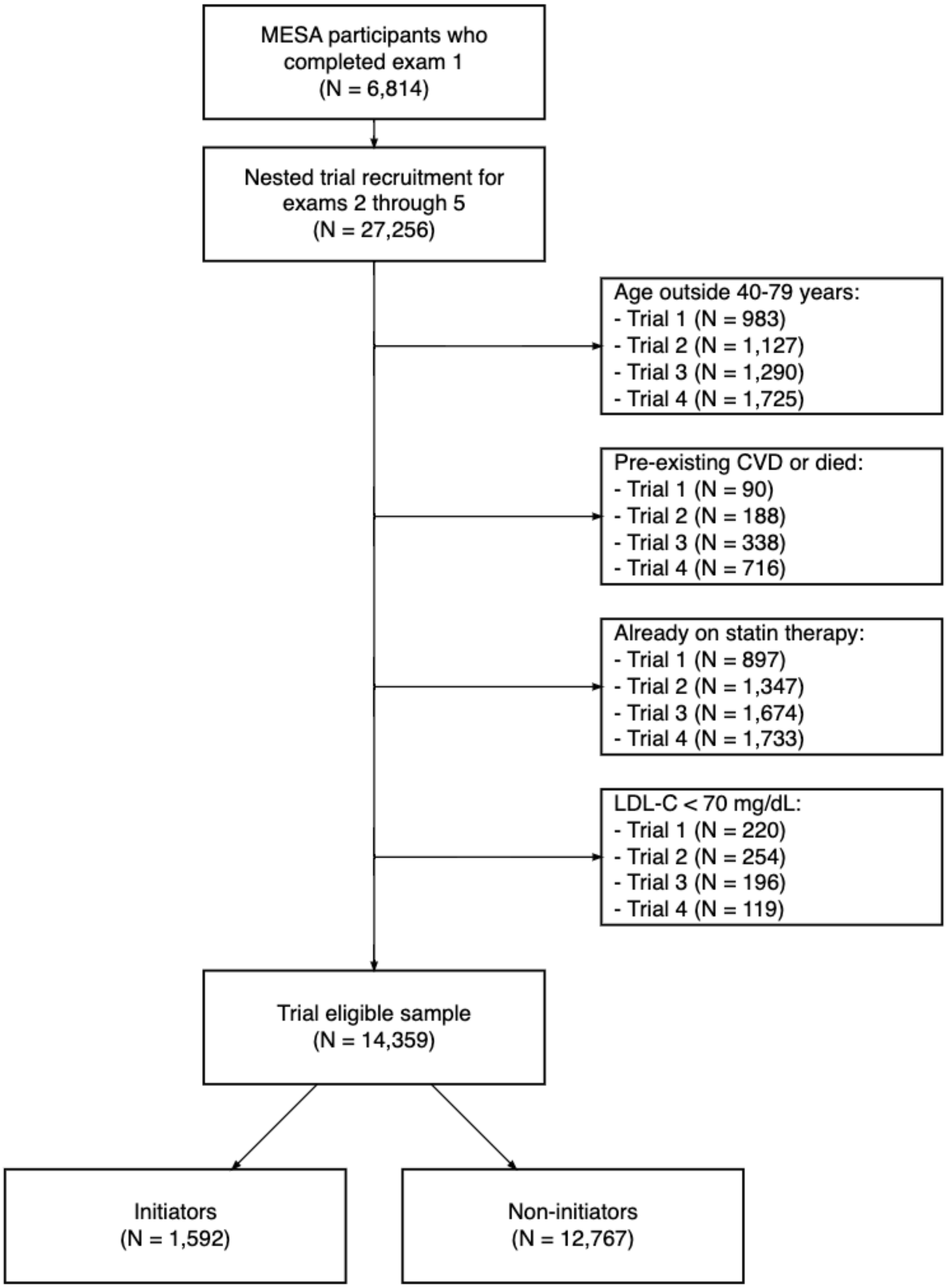


Figure B.3: *Exclusion criteria.*

Table B.3: *Benchmarking intention to treat and per protocol effects in emulated nested target trial of statin therapy, Multi-Ethnic Study of Atherosclerosis, 2000 to 2010*

	5-year[a]		**10-year**	
	HR	**95% CI**	**HR**	**95% CI**
ITT				
Pooled logit	0.79	(0.65, 0.93)	0.70	(0.56, 0.88)
g-estimation	0.77	(0.56, 0.98)	0.69	(0.47, 1.06)
Weibull κ	1.7		1.7	
Adherence-adjusted				
IPCW	0.68	(0.48, 0.94)	0.60	(0.39, 0.92)
g-estimation	0.66	(0.44, 0.98)	0.59	(0.37, 0.94)
Weibull κ	1.7		1.7	

HR = Hazard Ratio, CI = Confidence Interval

[a] 5-year estimate from HPS: HR = 0.75

ITT hazard ratio (HR) for statin initiation was 0.79 (95% CI: 0.65, 0.93) using pooled logistic regression in the nested trial data and 0.77 (95% CI: 0.56, 0.98) using g-estimation (Table B.3) after transforming from the survival time ratio using a Weibull distribution with scale parameter value 1.7. Both compared favorably with published 5-year ITT estimates from meta-analyses of statin treatment trials (HR = 0.75).

The probability of adherence among initators and non-initiators in the the nested trial emulation is shown in Figure B.2. The estimated HR comparing the always treat and never treat strategies was 0.68 (95% CI: 0.48, 0.94) using inverse probability of censoring weighting and 0.66 (95% CI: 0.44, 0.98) using g-estimation (Table B.3), suggesting stronger benefit of statins under full adherence. The estimated stabilized inverse probability weights had mean 1.02 in initiators and 0.99 in non-initiators.

B.4 Sensitivity Analyses

Below we provide sensitivity analyses in which we vary:

- The covariate adjustment sets.

- The estimated time zero.

To properly estimate the effect of statin initation in our benchmark two-arm trial, we assume we have conditioned on a sufficient set of covariates to make the exchangeability assumption plausible. In the first analysis, we show how our estimate changes as we include additional classes of covariates.

Because the timing of statin initiation was not known with certainty for all participants in MESA, we estimated it by drawing a random start month between the current and previous exam. Crucially, we applied this definition equally to initiators and noninitiators. Here we consider alternative definitions in which we use the earliest possible time (the last exam plus one month) and the latest possible time (the current exam minus one month) to see how sensitive our estimates are to the definition.

Table B.4: *Estimates of intention to treat effect of statin initation under different adjustment sets in emulated target trial for benchmarking, Multi-Ethnic Study of Atherosclerosis, 2000 to 2010.*

Model	HR	95% CI	P-value
unadjusted	1.03	(0.81, 1.31)	0.8
demographics[a]	0.90	(0.71, 1.14)	0.4
demographics[a] and risk factors[b]	0.81	(0.65, 0.91)	0.007
demographics[a], risk factors[b], and medications[c]	0.79	(0.64, 0.89)	0.004

CI = Confidence Interval; HR = Hazard Ratio

[a] Age, gender, marital status, education, race/ethnicity, employment, health insurance status, depression, perceived discrimination, emotional support, anger and anxiety scales, and neighborhood score

[b] Systolic and diastolic blood pressure, serum cholesterol levels (LDL, HDL, Triglycerides), hypertension, diabetes, waist circumference, smoking, alcohol consumption, exercise, family history of CVD, calcium score, hypertrophy on ECG, CRP, IL-6, number of pregnancies, oral contraceptive use, age of menopause

[c] Anti-hypertensive use, insulin use, daily aspirin use, anti-depressant use, vasodilator use, anti-arryhtmic use

Table B.5: *Estimates of intention to treat effect of statin initation under different estimated trial start times in emulated target trial for benchmarking, Multi-Ethnic Study of Atherosclerosis, 2000 to 2010.*

Model	HR	95% CI	P-value
randomly selected mo.	0.79	(0.64, 0.89)	0.004
last exam + 1 mo.	0.72	(0.62, 0.81)	<0.001
current exam - 1 mo.	0.76	(0.65, 0.84)	<0.001

CI = Confidence Interval; HR = Hazard Ratio

Appendix C

Appendix to Chapter 3

C.1 Time-fixed treatment initiation

C.1.1 Tailoring models for counterfactual predictions

Our goal is to build a model that targets the expected potential outcome under a hypothetical intervention, e.g. the parametric model

$$E[Y^a \mid X^*] = \mu_\beta(X^*).$$

However, we do not observe Y^a for all individuals. Here we show there are alternative targets written only in terms of observables in the training set that are identified under the conditions in section 3.4, namely

$$E[Y^a \mid X^*] = E[E[Y \mid X, A = a, D_{train} = 1] \mid X^*, D_{train} = 1] \tag{C.1}$$

and

$$E[Y^a \mid X^*] = E\left[\frac{I(A = a)}{\Pr(A = a \mid X, D_{train} = 1)} Y \mid X^*, D_{train} = 1\right] \tag{C.2}$$

in which case we can build a model for $E[Y^a \mid X^*]$ by targeting either estimand in the training dataset.

Proof. For the first representation we have

$$\begin{aligned} E[Y^a \mid X^*] &= E[Y^a \mid X^*, D_{train} = 1] \\ &= E(E[Y^a \mid X, D_{train} = 1] \mid X^*, D_{train} = 1) \\ &= E(E[Y^a \mid X, A = a, D_{train} = 1] \mid X^*, D_{train} = 1) \\ &= E(E[Y \mid X, A = a, D_{train} = 1] \mid X^*, D_{train} = 1) \end{aligned}$$

where the first line follows from the random sampling of the training set, the second from the law of iterated expectations, the third from the exchangeability condition, and the fourth from the consistency condition. Recall that X^* is a subset of X. For the second representation, we show that it is equivalent to the first

$$\begin{aligned} E[Y^a \mid X^*] &= E(E[Y \mid X, A = a, D_{train} = 1] \mid X^*, D_{train} = 1) \\ &= E\left(E\left[\frac{I(A = a)}{\Pr(A = a \mid X, D_{train} = 1)} Y \mid X, D_{train} = 1\right] \mid X^*, D_{train} = 1\right) \\ &= E\left(\frac{I(A = a)}{\Pr(A = a \mid X, D_{train} = 1)} E\left[Y \mid X, D_{train} = 1\right] \mid X^*, D_{train} = 1\right) \\ &= E\left[\frac{I(A = a)}{\Pr(A = a \mid X, D_{train} = 1)} Y \mid X^*, D_{train} = 1\right] \end{aligned}$$

where the second line follows from the definition of conditional expectation, the third removes the constant fraction outside expectation, and the last reverses the law of iterated expectations. ■

C.1.2 Identification of general loss functions

Here we show, for general counterfactual loss function $L\{Y^a, \mu_{\widehat{\beta}}\}$, the expected loss is identified by the functionals

$$\psi_{\widehat{\beta}} = E\left(E[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid X, A = a, D_{test} = 1] \mid D_{test} = 1\right) \tag{C.3}$$

and

$$\psi_{\widehat{\beta}} = E\left[\frac{I(A = a)}{\Pr(A = a \mid X, D_{test} = 1)} L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid D_{test} = 1\right] \tag{C.4}$$

in the test set under the time-fixed setup described in section 3.2. Many common performance measures, such as the mean squared error, Brier score, and absolute error, are special cases of the general loss function.

Proof. For the first representation we have

$$
\begin{aligned}
\psi_{\widehat{\beta}} &= E[L\{Y^a, \mu_{\widehat{\beta}}(X^*)\}] \\
&= E[L\{Y^a, \mu_{\widehat{\beta}}(X^*)\} \mid D_{test} = 1] \\
&= E(E[L\{Y^a, \mu_{\widehat{\beta}}(X^*)\} \mid X, D_{test} = 1] \mid D_{test} = 1) \\
&= E(E[L\{Y^a, \mu_{\widehat{\beta}}(X^*)\} \mid X, A = a, D_{test} = 1] \mid D_{test} = 1) \\
&= E(E[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid X, A = a, D_{test} = 1] \mid D_{test} = 1)
\end{aligned}
$$

where the first line follows from the definition of $\psi_{\widehat{\beta}}$, the second from random sampling of the test set, the third from the law of iterated expectations, the fourth from the exchangeability condition, and the fifth from the consistency condition. Recall that X^* is a subset of X. For the second representation, we show that it is equivalent to the first

$$
\begin{aligned}
\psi_{\widehat{\beta}} &= E(E[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid X, A = a, D_{test} = 1] \mid D_{test} = 1) \\
&= E\left(E\left[\frac{I(A = a)}{\Pr(A = a \mid X, D_{test} = 1)} L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid X, D_{test} = 1\right] \mid D_{test} = 1\right) \\
&= E\left(\frac{I(A = a)}{\Pr(A = a \mid X, D_{test} = 1)} E\left[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid X, D_{test} = 1\right] \mid D_{test} = 1\right) \\
&= E\left[\frac{I(A = a)}{\Pr(A = a \mid X, D_{test} = 1)} L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid D_{test} = 1\right]
\end{aligned}
$$

where the second line follows from the definition of conditional expectation, the third removes the constant fraction outside expectation, and the last reverses the law of iterated expectations. ■

C.1.3 Plug-in estimation

Using sample analogs for the identified expressions C.3 and C.4, we obtain two plug-in estimators for the expected loss for a generalized loss function

$$\widehat{\psi}_{CL} = \frac{1}{n_{test}} \sum_{i=1}^{n} I(D_{test,i} = 1)\widehat{h}_a(X_i)$$

and

$$\widehat{\psi}_{IPW} = \frac{1}{n_{test}} \sum_{i=1}^{n} \frac{I(A_i = a, D_{test,i} = 1)}{\widehat{e}_a(X_i)} L\{Y, \mu_{\widehat{\beta}}(X_i^*)\}$$

where $\widehat{h}_a(X)$ is an estimator for $E[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid X, A = a, D_{test} = 1]$ and $\widehat{e}_a(X)$ is an estimator for $\Pr(A = a \mid X, D_{test} = 1)$. Using the terminology in Morrison et al., we call the first plug-in estimator the conditional loss estimator $\widehat{\psi}_{CL}$ and the second the inverse probability weighted estimator $\widehat{\psi}_{IPW}$.

C.1.4 Random and dynamic regimes

Above we consider static interventions which set treatment A to a particular value a. We might also consider interventions which probabilistically set A based on a known density, possibly conditional on pre-treatment covariates, e.g. $f^{int}(A \mid X)$. For instance, instead of a counterfactual prediction if everyone or no one had been treated, we may be interested in the prediction if 20% or 50% were treated. We term such an intervention a *random* intervention to contrast it with *static* interventions considered previously. Random interventions are closer to the counterfactual interventions of interest under dataset shift which may be approximated as probabilistic changes in the natural course of treatment due to changes in guidelines or prescribing patterns or the wider-availability. For general counterfactual loss function $L\{Y^g, \mu_{\widehat{\beta}}\}$, the expected loss under a random intervention is identified by the functionals

$$\psi_{\widehat{\beta}} = E\left\{E_{f^{int}}\left(E[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid X, A = a, D_{test} = 1] \mid D_{test} = 1\right)\right\} \tag{C.5}$$

and

$$\psi_{\widehat{\beta}} = E\left[\frac{I(A = a)}{\Pr(A = a \mid X, D_{test} = 1)} L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid D_{test} = 1\right] \tag{C.6}$$

in the test set under the time-fixed setup described in section 3.2. The primary difference between these expressions and the ones in section A.1. is that the expectation is taken with respect to the intervention density.

C.2 Time-varying treatment initiation

C.2.1 Set up

Here we extend the set up of section 3.2 in the case that treatment initiation is time-varying over the follow up period. We now observe n i.i.d. longitudinal samples $\{O_i\}_{i=1}^n$ from a source population. For each observation, let

$$O_i = (\overline{X}_K, \overline{A}_K, Y_{K+1})$$

where overbars denote the full history of a variable, such that $\overline{X}_k = (X_0, \ldots, X_k)$, and variables X_k, A_k, and Y_{K+1} are defined as previously. We still assume interest lies in building a prediction model for the outcome Y_{K+1} conditional on baseline covariates X^* which are now a subset of X_0, i.e. $X^* \subset X_0$. An example DAG for a two time point process is shown in Figure C.1

We would like to assess the performance of the model in a counterfactual version of the source population in which a new treatment policy is implemented. As previously, Y^a is the potential outcome under an intervention which sets treatment A to a. For a sequence of time-varying treatments $\overline{A}_k$, we further define a *treatment regime* as a collection of functions $\{g_k(\overline{a}_{k-1}, \overline{x}_k) : k = 0, \ldots, K\}$ for determining treatment assignment at each time k, possibly based on past treatment and covariate history. For a hypothetical treatment regime g, we would like to determine the performance of fitted model $\mu_{\widehat{\beta}}(X^*)$ under the new regime by estimating the expected loss

$$\psi_{\widehat{\beta}} = E[L\{Y^g, \mu_{\widehat{\beta}}(X^*)\}]$$

for generalized loss function $L\{Y^g, \mu_{\widehat{\beta}}(X^*)\}$.

C.2.2 Identifiability conditions

We now consider modified identifiability conditions under time-varying treatment initiation. For all k from 0 to K, we require

1. *Exchangeability:* $Y^g_{K+1} \perp\!\!\!\perp A_k \mid \overline{X}_k, \overline{A}_{k-1}$
2. *Consistency:* $Y_{K+1} = Y^g_{K+1}$ and $\overline{X}_k = \overline{X}^g_k$ if $\overline{A}_k = \overline{a}^g_k$
3. *Positivity:* $1 > \Pr(A_k = a_k \mid \overline{X}_k = \overline{X}_k, \overline{A}_{k-1} = \overline{a}_{k-1}) > 0$

C.2.3 Identification of general loss functions

Under time-varying treatment initiation, the expected counterfactual loss for general loss function $L\{Y^g, \mu_{\widehat{\beta}}\}$ is identified by the functionals

$$\begin{aligned}\psi_{\widehat{\beta}} = E_{X_0}\Bigg[E_{X_1}\Bigg\{ \dots E_{X_{K-1}}\Bigg(E_{X_K}[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid \overline{X}_K, \overline{A}_K = \overline{a}^g_K, D_{test} = 1] \\ \mid \overline{X}_{K-1}, \overline{A}_{K-1} = \overline{a}^g_{K-1}, D_{test} = 1\Bigg) \dots \mid X_0, A_0 = a^g_0, D_{test} = 1\Bigg\} \mid D_{test} = 1\Bigg]\end{aligned} \tag{C.7}$$

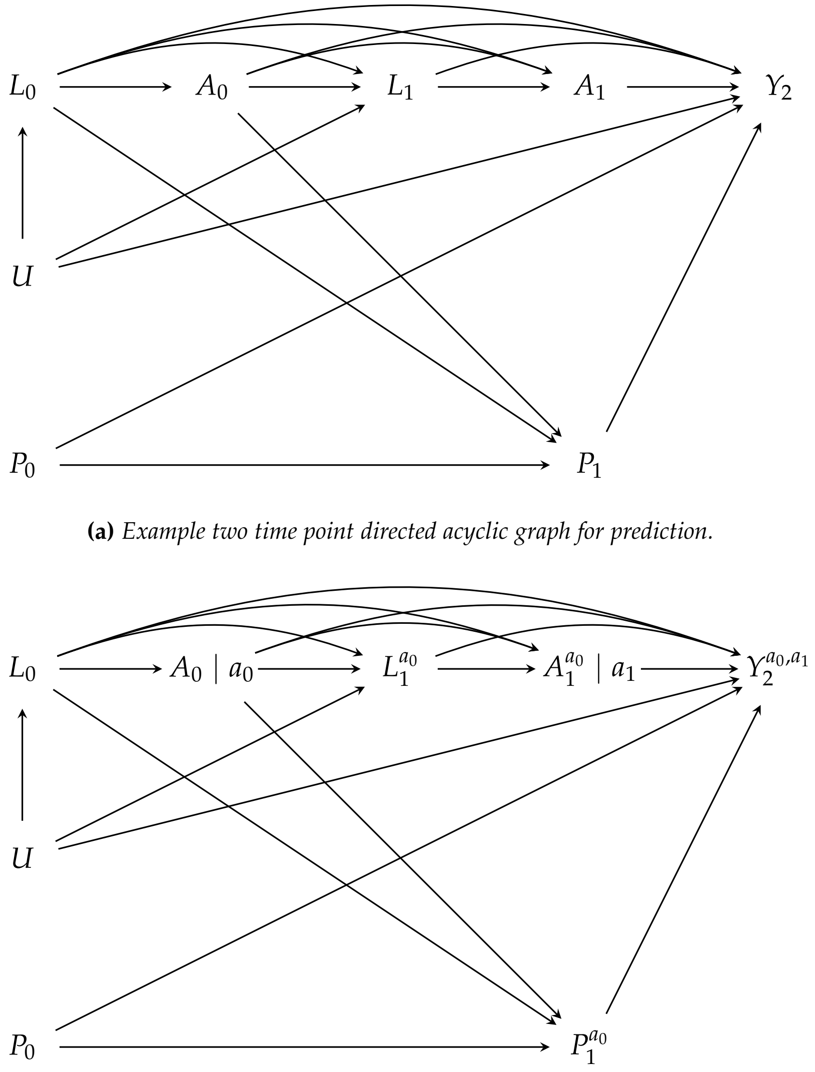
and

$$\psi_{\widehat{\beta}} = E\left[\frac{I(\overline{A}_K = \overline{a}^g_K, D_{test} = 1)}{\prod_{k=0}^K \Pr(A_k = a^g_k \mid \overline{X}_k, \overline{A}_{k-1} = \overline{a}^g_{k-1}, D_{test} = 1)} L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid D_{test} = 1\right] \tag{C.8}$$

in the test set, where the first is a sequence of iterated expectations and the second is an inverse-probability weighted expectation.

Proof. For the first representation we have

$$\begin{aligned}\psi_{\widehat{\beta}} &= E[L\{Y^g, \mu_{\widehat{\beta}}(X^*)\}] \\ &= E[L\{Y^g, \mu_{\widehat{\beta}}(X^*)\} \mid D_{test} = 1] \\ &= E(E[L\{Y^g, \mu_{\widehat{\beta}}(X^*)\} \mid X_0, D_{test} = 1] \mid D_{test} = 1) \\ &= E(E[L\{Y^g, \mu_{\widehat{\beta}}(X^*)\} \mid X_0, A_0 = a^g_0, D_{test} = 1] \mid D_{test} = 1)\end{aligned}$$



(a) *Example two time point directed acyclic graph for prediction.*

(b) *Single world intervention graph of intervention on A_0 and A_1.*

Figure C.1: *Example directed acyclic graph (DAG) and single world intervention graph (SWIG) for a two time point process.*

where the first line follows from the definition of $\psi_{\widehat{\beta}}$, the second from random sampling of the test set, the third from the law of iterated expectations, and the fourth from the exchangeability condition. Arguing recursively from $k = 0$ to K, we can repeatedly invoke iterated expectations and exchanageability to insert $\overline{X}_k$ and $\overline{A}_k = \overline{a}^g_k$, such that

$$\begin{aligned}
\psi_{\widehat{\beta}} &= E_{X_0}\Bigg[E_{X_1}\Bigg\{\dots E_{X_{K-1}}\Bigg(E_{X_K}[L\{Y^g, \mu_{\widehat{\beta}}(X^*)\} \mid \overline{X}_K, \overline{A}_K = \overline{a}^g_K, D_{test} = 1] \\
&\qquad \mid \overline{X}_{K-1}, \overline{A}_{K-1} = \overline{a}^g_{K-1}, D_{test} = 1\Bigg) \dots \mid X_0, A_0 = a^g_0, D_{test} = 1\Bigg\} \mid D_{test} = 1\Bigg] \\
&= E_{X_0}\Bigg[E_{X_1}\Bigg\{\dots E_{X_{K-1}}\Bigg(E_{X_K}[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid \overline{X}_K, \overline{A}_K = \overline{a}^g_K, D_{test} = 1] \\
&\qquad \mid \overline{X}_{K-1}, \overline{A}_{K-1} = \overline{a}^g_{K-1}, D_{test} = 1\Bigg) \dots \mid X_0, A_0 = a^g_0, D_{test} = 1\Bigg\} \mid D_{test} = 1\Bigg]
\end{aligned}$$

where the last line follows by consistency. For the second representation, note that for the inner most expectations we can proceed as previously

$$\begin{aligned}
&E(E[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid \overline{X}_K, \overline{A}_k = \overline{a}^g_K, D_{test} = 1] \mid \overline{X}_{K-1}, \overline{A}_{k-1} = \overline{a}^g_{K-1}, D_{test} = 1) \\
&= E\left(E\left[W_K L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid \overline{X}_K, \overline{A}_{K-1}, D_{test} = 1\right] \mid \overline{X}_{K-1}, \overline{A}_{K-1}, D_{test} = 1\right) \\
&= E\left(W_K E\left[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid \overline{X}_K, \overline{A}_{K-1}, D_{test} = 1\right] \mid \overline{X}_{K-1}, \overline{A}_{K-1}, D_{test} = 1\right) \\
&= E\left[W_K L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid \overline{X}_{K-1}, \overline{A}_{K-1}, D_{test} = 1\right]
\end{aligned}$$

where the second line follows from the definition of conditional expectation, the third removes the constant fraction outside expectation, and the last reverses the law of iterated expectations and where

$$W_K = \frac{I(A_K = a^g_K, D_{test} = 1)}{\Pr(A_K = a^g_K \mid \overline{X}_K, \overline{A}_{K-1}, D_{test} = 1)}$$

Arguing recursively from $k = 0$ to K, we get

$$\psi_{\widehat{\beta}} = E\left[\frac{I(\overline{A}_K = \overline{a}^g_K, D_{test} = 1)}{\prod_{k=0}^{K} \Pr(A_k = a^g_k \mid \overline{X}_k, \overline{A}_{k-1} = \overline{a}^g_{k-1}, D_{test} = 1)} L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid D_{test} = 1\right]$$

which is the inverse-probability weighted representation with weights equal to

$$W_k = \frac{I(\overline{A}_K = \overline{a}^g_K, D_{test} = 1)}{\prod_{k=0}^{K} \Pr(A_k = a^g_k \mid \overline{X}_k, \overline{A}_{k-1} = \overline{a}^g_{k-1}, D_{test} = 1)}$$

. ■

C.2.4 Plug-in estimation

Using sample analogs for the identified expressions C.7 and C.8, we obtain two plug-in estimators for the expected counterfactual loss under a generalized loss function

$$\widehat{\psi}_{CL} = \sum_{i=1}^{n} I(D_{test,i} = 1)\widehat{h}_{a_0}(X_i)$$

and

$$\widehat{\psi}_{IPW} = \sum_{i=1}^{n} \frac{I(\overline{A}_K = \overline{a}^g_K, D_{test,i} = 1)}{\prod_{k=0}^{K} \widehat{e}_{a_k}(X_i)} L\{Y, \mu_{\widehat{\beta}}(X_i^*)\}$$

where $h_{t+1} = L\{Y, \mu_{\widehat{\beta}}(X_i^*)$ and $h_{a_0}(X)$ is recursively defined for t = $K, \ldots, 0$

$$h_{a_t} : (x_t, a_t) E[h_{a_{t+1}}(X_{t+1}) \mid \overline{X}_t, \overline{A}_t = \overline{a}^g_t]$$

$\widehat{h}_a(X)$ is an estimator for $E[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid X, A = a, D_{test} = 1]$ and $\widehat{e}_{a_k}(X)$ is an estimator for $\Pr(A_k = a^g_k \mid \overline{X}_k, \overline{A}_{k-1} = \overline{a}^g_{k-1}, D_{test} = 1)$. Note that as the number of time points (i.e. K) increases, the proportion in the test set who actually follow the regime of interest, i.e. those for whom $I(\overline{A}_K = \overline{a}^g_K, D_{test,i} = 1) = 1$ may be prohibitively small, in which case plug-in estimation may not be feasible. In this case, additional modeling assumptions will be necessary to borrow information from other regimes.

C.3 Doubly robust estimators

C.3.1 Efficient influence function

As we've shown previously, under the identifiability conditions of section 3.4, the expected counterfactual loss of a generalized loss function $L\{Y^a, \mu(X^*)\}$ is identified by the observed data functional

$$\psi = E\left(E[L\{Y, \mu(X^*)\} \mid X, A = a]\right).$$

The influence function for ψ under a nonparametric model for the observable data $O = (X, A, Y)$ is

$$\begin{aligned}\chi^1_{P_0} = \frac{I(A=a)}{\Pr(A = a \mid X)}&(L\{Y, \mu(X^*)\} - E[L\{Y, \mu(X^*)\} \mid X, A = a]) + \\ &(E[L\{Y, \mu(X^*)\} \mid X, A = a] - \psi).\end{aligned}$$

As the influence function under a nonparametric model is always unique, it is also the efficient influence function.

Proof. To show that $\chi^1_{P_0}$ is the efficient influence function, we will use the well-known fact that the influence function is a solution to

$$\left.\frac{d}{dt}\psi_{P_t}\right|_{t=0} = E_{P_0}(\chi^1_{P_0} g_{P_0})$$

where g_{P_0} is the score of the obeservable data under the true law P_0 and P_t is a parametric submodel indexed by $t \in [0, 1]$ and the pathwise derivative of the submodel is evaluated at $t = 0$ corresponding to the true law P_0. Let $h_a(X) = E_{P_0}[L\{Y, \mu(X^*)\} \mid X, A = a]$. Beginning with the left hand side

$$
\begin{aligned}
\frac{d}{dt}\psi_{P_t}\bigg|_{t=0} &= \frac{d}{dt}E_{P_t}\left(E_{P_t}[L\{Y,\mu(X^*)\} \mid X, A=a]\right)\bigg|_{t=0} \\
&= \frac{\partial}{\partial t}E_{P_t}\left(E_{P_0}[L\{Y,\mu(X^*)\} \mid X, A=a]\right)\bigg|_{t=0} + \\
&\qquad E_{P_0}\left(\frac{\partial}{\partial t}E_{P_t}[L\{Y,\mu(X^*)\} \mid X, A=a]\bigg|_{t=0}\right) \\
&= E_{P_0}\left[\{h_a(X)-\psi\}\, g_{X,A,Y}(O)\right] + \\
&\qquad E_{P_0}\left\{\left(\frac{I(A=a)}{\Pr(A=a \mid X)}\Big[L\{Y,\mu(X^*)\} - h_a(X)\Big]\right) g_{X,A,Y}(O)\right\} \\
&= E_{P_0}\left\{\left(h_a(X) - \psi + \frac{I(A=a)}{\Pr(A=a \mid X)}\Big[L\{Y,\mu(X^*)\} - h_a(X)\Big]\right) g_{X,A,Y}(O)\right\}
\end{aligned}
$$

where the first line is the definition, the second line applies the chain rule, the third applies definition of the score, and the last uses linearity of expectations. Returning to original supposition, it follows that the influence function is

$$
\begin{aligned}
\chi^1_{P_0} = &\frac{I(A=a)}{\Pr(A=a \mid X)}(L\{Y,\mu(X^*)\} - E[L\{Y,\mu(X^*)\} \mid X, A=a]) + \\
&(E[L\{Y,\mu(X^*)\} \mid X, A=a] - \psi).
\end{aligned}
$$

■

C.3.2 One-step estimator

Given the efficient influence function above and random sampling in the test set, the one-step estimator for ψ is given by

$$
\widehat{\psi}_{DR} = \frac{1}{n_{test}}\sum_{i=1}^{n} I(D_{test,i}=1)\widehat{h}_a(X_i) + \frac{I(A_i=a, D_{test,i}=1)}{\widehat{e}_a(X_i)}\Big[L\{Y,\mu(X_i^*)\} - \widehat{h}_a(X_i)\Big]
$$

C.3.3 Asymptotic properties

In previous sections, the asymptotic properties of $\widehat{\psi}_{CL}$ and $\widehat{\psi}_{IPW}$ follow from standard parametric theory[1]. However, here the asymptotic properties of $\widehat{\psi}_{DR}$ are complicated by

[1] after separating estimation of $\mu_\beta(X^*)$ from the evaluation of performance by random partition of test set.

the estimation of two nuisance functions, $\widehat{h}_a(X)$ and $\widehat{e}_a(X)$, and the fact that, we do not immediately assume a parametric model for either. To simplify the derivation of the large sample properties of $\widehat{\psi}_{DR}$ we begin by defining

$$H\left(e'_a(X), h'_a(X)\right) = h'_a(X) + \frac{I(A=a)}{e'_a(X)}\left[L\left(Y, \mu\left(X^*\right)\right) - h'_a(X)\right]$$

for arbitrary functions $e'_a(X)$, and $h'_a(X)$. Here we suppress the dependence on being in the test set for ease of exposition, but note that the rest procedes the same if we were to limit our focus to the test set. Note, the doubly robust estimator can be written as $\widehat{\psi}_{DR} = \frac{1}{n}\sum_{i=1}^{n} H\left(\widehat{e}_a\left(X_i\right), \widehat{h}_a\left(X_i\right)\right)$. We define the probability limits of $\widehat{e}_a(X)$ and $\widehat{h}_a(X)$ as $e^*_a(X)$ and $h^*_a(X)$, respectively. By definition, when $\widehat{e}_a(X)$ and $\widehat{h}_a(X)$ are correctly specified, the limits are $e^*_a(X) = \Pr[A = a \mid X]$ and $h^*_a(X) = \mathrm{E}\left[L\left(Y, \mu\left(X^*\right)\right) \mid X, A = a\right]$.

To derive the asymptotic properties of $\widehat{\psi}_{DR}$, we make the following assumptions:

D1. $H(\widehat{e}_a(X), \widehat{h}_a(X))$ and its limit $H\left(e^*_a(X), h^*_a(X)\right)$ fall in a Donsker class.

D2. $\left\|H(\widehat{e}_a(X), \widehat{h}_a(X)) - H\left(e^*_a(X), h^*_a(X)\right)\right\| \xrightarrow{P} 0.$

D3. (Finite second moment). $\mathrm{E}\left[H\left(e^*_a(X), h^*_a(X)\right)^2\right] < \infty$.

D4. (Model double robustness). At least one of the models $\widehat{e}_a(X)$ or $\widehat{h}_a(X)$ is correctly specified. That is, at least one of $e^*_a(X) = \Pr[A = a \mid X]$ or $h^*_a(X) = \mathrm{E}\left[L\left(Y, \mu\left(X^*\right)\right) \mid X, A = a\right]$ holds, but not necessarily both.

Assumption D1 is a well-known restriction on the complexity of the functionals $\widehat{e}_a(X)$ and $\widehat{h}_a(X)$. As long as $\widehat{e}_a(X), \widehat{h}_a(X), e^*_a(X)$, and $h^*_a(X)$ are Donsker and all are uniformly bounded then Assumption D1 holds by the Donsker preservation theorem. Many commonly used models such as generalized linear models fall within the Donsker class. This requirement can be further relaxed through sample-splitting, in which case more flexible machine learning algorithms such as random forests, gradient boosting, or neural networks may be used to estimate $\widehat{e}_a(X)$ and $\widehat{h}_a(X)$.

Using Assumptions D1 through D4, below we prove:

1. (Consistency) $\widehat{\psi}_{DR} \xrightarrow{P} \psi$.

2. (Asymptotic distribution) $\widehat{\psi}_{DR}$ has the asymptotic representation

$$\sqrt{n}\left(\widehat{\psi}_{DR} - \psi\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} H\left(e_a^*(X_i), h_a^*(X_i)\right) - \mathrm{E}\left[H\left(e_a^*(X), h_a^*(X)\right)\right]\right) + Re + o_P(1),$$

where

$$Re \leq \sqrt{n} O_P\left(\left\|\widehat{h}_a(X) - \mathrm{E}\left[L\left(Y, \mu(X^*)\right) \mid X, A = a\right]\right\|_2^2 \times \left\|\widehat{e}_a(X) - \Pr[A = a \mid X]\right\|_2^2\right)$$

and thus if $\widehat{h}_a(X)$ and $\widehat{e}_a(X)$ converge at combined rate of at least $\sqrt{n}$ then

$$\sqrt{n}\left(\widehat{\psi}_{DR} - \psi\right) \xrightarrow{d} N\left(0, \mathrm{Var}\left[H(e_a^*(X), h_a^*(X))\right]\right)$$

Consistency

Using the probability limits $e_a^*(X)$ and $h_a^*(X)$ defined previously, the double robust estimator $\widehat{\psi}_{DR}$ converges in probability to

$$\widehat{\psi}_{DR} \xrightarrow{P} E\left[h_a^*(X) + \frac{I(A=a)}{e_a^*(X)}\left(L\left(Y, \mu\left(X^*\right)\right) - h_a^*(X)\right)\right]$$

Here we show that the right-hand side is equal to ψ under assumptions D1- D4 when either:

1. $\widehat{e}_a(X)$ is correctly specified

2. $\widehat{h}_a(X)$ is correctly specified

First consider the case where $\widehat{e}_a(X)$ is correctly specified, that is $e_a^*(X) = \Pr[A = a \mid X]$, but we do not assume that the limit $h_a^*(X)$ is equal to $\mathrm{E}\left[L\left(Y, g\left(X^*\right)\right) \mid X, A = a\right])$. Recall, as shown previously $\psi = E\left[\frac{I(A=a)}{\Pr(A=a|X)} L(Y, \mu_{\widehat{\beta}}(X^*))\right]$

$$\begin{aligned}
\widehat{\psi}_{DR} &\xrightarrow{P} E\left[h_a^*(X) + \frac{I(A=a)}{e_a^*(X)}\left(L\left(Y, \mu\left(X^*\right)\right) - h_a^*(X)\right)\right] \\
&= E\left[h_a^*(X) - \frac{I(A=a)}{e_a^*(X)} h_a^*(X)\right] + \psi \\
&= E\left[E\left[h_a^*(X) - \frac{I(A=a)}{e_a^*(X)} h_a^*(X) \mid X\right]\right] + \psi \\
&= E\left[h_a^*(X) - \frac{1}{e_a^*(X)} h_a^*(X) E\left[I(A=a) \mid X\right]\right] + \psi \\
&= E\left[h_a^*(X) - \frac{1}{e_a^*(X)} h_a^*(X) \Pr\left[A=a \mid X\right]\right] + \psi \\
&= E\left[h_a^*(X) - h_a^*(X)\right] + \psi \\
&= \psi.
\end{aligned}$$

Next consider the case when $\widehat{h}_a(X)$ is correctly specified, that is

$$h_a^*(X) = \mathrm{E}\left[L\left(Y, g\left(X^*\right)\right) \mid X, A=a\right]$$

and this time we do not make the assumptions that the limit $e_a^*(X)$ is equal to $\Pr[A=a \mid X]$. Recall, as shown previously $\psi = E\left[E\left[L(Y, \mu_{\widehat{\beta}}(X^*)) \mid X, A=a\right]\right]$.

$$\begin{aligned}
\widehat{\psi}_{DR} &\xrightarrow{P} E\left[h_a^*(X) + \frac{I(A=a)}{e_a^*(X)}\left(L\left(Y, \mu\left(X^*\right)\right) - h_a^*(X)\right)\right] \\
&= E\left[h_a^*(X)\right] + E\left[\frac{I(A=a)}{e_a^*(X)}\left(L\left(Y, \mu\left(X^*\right)\right) - h_a^*(X)\right)\right] \\
&= \psi + E\left[\frac{I(A=a)}{e_a^*(X)}\left(L\left(Y, \mu\left(X^*\right)\right) - h_a^*(X)\right)\right] \\
&= \psi + E\left[E\left[\frac{I(A=a)}{e_a^*(X)}\left(L\left(Y, \mu\left(X^*\right)\right) - h_a^*(X)\right) \mid X\right]\right] \\
&= \psi + E\left[\frac{I(A=a)}{e_a^*(X)} E\left[\left(L\left(Y, \mu\left(X^*\right)\right) - h_a^*(X)\right) \mid X\right]\right] \\
&= \psi + E\left[E\left[\left(L\left(Y, \mu\left(X^*\right)\right) - h_a^*(X)\right) \mid X, A=a\right]\right] \\
&= \psi + E\left[E\left[L\left(Y, \mu\left(X^*\right)\right) \mid X, A=a\right] - h_a^*(X)\right] \\
&= \psi + E\left[h_a^*(X) - h_a^*(X)\right] \\
&= \psi.
\end{aligned}$$

Asymptotic distribution

For a random variable W we define notation

$$\mathbb{G}_n(W) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} W_i - \mathrm{E}[W]\right).$$

and thus the asymptotic representation of $\widehat{\psi}_{DR}$ can be written

$$\begin{aligned}\sqrt{n}\left(\widehat{\psi}_{DR} - \psi\right) = & \, \mathbb{G}_n(H(\widehat{e}_a(X), \widehat{h}_a(X))) - \mathbb{G}_n\left(H\left(e_a^*(X), h_a^*(X)\right)\right) \\ & + \mathbb{G}_n\left(H\left(e_a^*(X), h_a^*(X)\right)\right) \\ & + \sqrt{n}(\mathrm{E}[H(\widehat{e}_a(X), \widehat{h}_a(X))] - \psi)\end{aligned}$$

where we add and subtract the term $\mathbb{G}_n\left(H\left(e_a^*(X), h_a^*(X)\right)\right)$ and add another zero term in $+\sqrt{n}(\mathrm{E}[H(\widehat{e}_a(X), \widehat{h}_a(X))] - \psi)$. For the first term, Assumption D1 implies

$$\mathbb{G}_n(H(\widehat{e}_a(X), \widehat{h}_a(X))) - \mathbb{G}_n\left(H\left(e_a^*(X), h_a^*(X)\right)\right) = o_P(1)$$

Let

$$Re = \sqrt{n}(\mathrm{E}[H(\widehat{e}_a(X), \widehat{h}_a(X))] - \psi)$$

now we have

$$\sqrt{n}\left(\widehat{\psi}_{DR} - \psi\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(H\left(e_a^*(X_i), h_a^*(X_i)\right) - \mathrm{E}\left[H\left(e_a^*(X), h_a^*(X)\right)\right]\right)\right) + Re + o_P(1)$$

Let's try to calculate the upper bound of Re. First, note

$$n^{-1/2}Re = \underbrace{\mathrm{E}\left[\widehat{h}_a(X)\right]}_{R_1} + \underbrace{\mathrm{E}\left[\frac{I(A=a)}{\widehat{e}_a(X)}\left[L\left(Y, \mu\left(X^*\right)\right) - \widehat{h}_a(X)\right]\right]}_{R_2} - \psi.$$

We rewrite term R_2 as:

$$
\begin{aligned}
R_2 &= \mathrm{E}\left[\frac{I(A=a)}{\widehat{e}_a(X)}\left\{L\left(Y,\mu\left(X^*\right)\right)-\widehat{h}_a(X)\right\}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\frac{I(A=a)}{\widehat{e}_a(X)}\left\{L\left(Y,\mu\left(X^*\right)\right)-\widehat{h}_a(X)\right\} \mid X\right]\right] \\
&= \mathrm{E}\left[\frac{1}{\widehat{e}_a(X)}\mathrm{E}\left[\frac{I(A=a)}{\Pr[A=a\mid X]}\Pr[A=a\mid X]\left\{L\left(Y,\mu\left(X^*\right)\right)-\widehat{h}_a(X)\right\} \mid X\right]\right] \\
&= \mathrm{E}\left[\frac{1}{\widehat{e}_a(X)}\mathrm{E}\left[\Pr[A=a\mid X]\left\{L\left(Y,\mu\left(X^*\right)\right)-\widehat{h}_a(X)\right\} \mid X, A=a\right]\right] \\
&= \mathrm{E}\left[\frac{1}{\widehat{e}_a(X)}\Pr[A=a\mid X]\left\{\mathrm{E}\left[L\left(Y,\mu\left(X^*\right)\right) \mid X, A=a\right]-\widehat{h}_a(X)\right\}\right]
\end{aligned}
$$

Combining the above gives

$$
\begin{aligned}
n^{-1/2}Re &= \mathrm{E}\left[\widehat{h}_a(X)\right]+\mathrm{E}\left[\frac{I(A=a)}{e'_a(X)}\left[L\left(Y,\mu\left(X^*\right)\right)-h'_a(X)\right]\right]-\psi \\
&= \mathrm{E}\left[\widehat{h}_a(X)\right]+\mathrm{E}\left[\frac{1}{\widehat{e}_a(X)}\Pr[A=a\mid X]\left\{\mathrm{E}\left[L\left(Y,\mu\left(X^*\right)\right) \mid X, A=a\right]-\widehat{h}_a(X)\right\}\right] \\
&- E\left[E\left[L(Y,\mu_{\widehat{\beta}}(X^*)) \mid X, A=a\right]\right] \\
&= E\left[\left\{\mathrm{E}\left[L\left(Y,\mu\left(X^*\right)\right) \mid X, A=a\right]-\widehat{h}_a(X)\right\}\times\left\{\frac{1}{\widehat{e}_a(X)}\Pr[A=a\mid X]-1\right\}\right]
\end{aligned}
$$

Using the Cauchy-Schwartz inequality we get.

$$
\begin{aligned}
Re &\leq \sqrt{n}\left(\mathrm{E}\left[\left\{\mathrm{E}\left[L\left(Y,\mu\left(X^*\right)\right) \mid X, A=a\right]-\widehat{h}_a(X)\right\}^2\right]\right)^{1/2} \\
&\times\left(\mathrm{E}\left[\left\{\frac{1}{\widehat{e}_a(X)}\Pr[A=a\mid X]-1\right\}^2\right]\right)^{1/2} \\
&\leq \sqrt{n}O_P\left(\left\|\mathrm{E}\left[L\left(Y,\mu\left(X^*\right)\right) \mid X, A=a\right]-\widehat{h}_a(X)\right\|_2^2\times\left\|\widehat{e}_a(X)-\Pr[A=a\mid X]\right\|_2^2\right)
\end{aligned}
$$

If both models $\widehat{e}_a(X)$ and $\widehat{h}_a(X)$ are correctly specified and converge at a combined rate faster than $\sqrt{n}$, then $Re = o_P(1)$ and

$$
\begin{aligned}
\sqrt{n}\left(\widehat{\psi}_{DR}-\psi\right) = \sqrt{n}&\left(\frac{1}{n}\sum_{i=1}^{n} H\left(\Pr\left[A=a\mid X_i\right],\mathrm{E}\left[L\left(Y,g\left(X^*\right)\right) \mid A=a, X_i\right]\right)\right. \\
&\left.-\mathrm{E}\left[H\left(\Pr[A=a\mid X],\mathrm{E}\left[L\left(Y,g\left(X^*\right)\right) \mid A=a, X\right]\right)\right]\right)+o_P(1)
\end{aligned}
$$

By the central limit theorem,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} H\left(e_a^*(X_i), h_a^*(X_i)\right) - \mathrm{E}\left[H\left(e_a^*(X_i), h_a^*(X_i)\right)\right]\right) \xrightarrow{d} N\left(0, \mathrm{Var}\left[H\left(e_a^*(X), h_a^*(X)\right)\right]\right)$$

completing the proof.

C.4 Risk calibration curve

Another common metric of the performance of risk prediction models is model calibration, that is are the risk estimates produced by the model reliable in the sense that for 100 patients who receive a risk prediction of 17% does the outcome really occur for roughly 17 of them over the follow up period. This can be nonparametrically evalutated by estimating the so-called "calibration" curve, i.e. the observed risk as a function of the predicted risk. For counterfactual predictions the relevant calibration curve though is the counterfactual risk that would be observed under intervetion $A = a$ as a function of the predicted risk, or

$$\psi_{\widehat{\beta}} = E[I(Y^a = 1) \mid \mu_{\widehat{\beta}}(X^*)]. \tag{C.9}$$

C.4.1 Identification

Here we show that the counterfactual calibration curve is identified by the observed data functionals

$$\psi_{\widehat{\beta}} = E[E\{I(Y = 1) \mid X, A = a, \mu_{\widehat{\beta}}(X^*), D_{test} = 1\} \mid \mu_{\widehat{\beta}}(X^*), D_{test} = 1] \tag{C.10}$$

and

$$\psi_{\widehat{\beta}} = E\left[\frac{I(A = a)}{\Pr(A = a \mid X, \mu_{\widehat{\beta}}(X^*), D_{test} = 1)} I(Y = 1) \mid \mu_{\widehat{\beta}}(X^*), D_{test} = 1\right] \tag{C.11}$$

in the test set.

Proof. For the first representation we have

$$\begin{aligned}
\psi_{\widehat{\beta}} &= E[I(Y^a = 1) \mid \mu_{\widehat{\beta}}(X^*)] \\
&= E[I(Y^a = 1) \mid \mu_{\widehat{\beta}}(X^*), D_{test} = 1] \\
&= E[E\{I(Y^a = 1) \mid X, \mu_{\widehat{\beta}}(X^*), D_{test} = 1\} \mid \mu_{\widehat{\beta}}(X^*), D_{test} = 1] \\
&= E[E\{I(Y^a = 1) \mid X, A = a, \mu_{\widehat{\beta}}(X^*), D_{test} = 1\} \mid \mu_{\widehat{\beta}}(X^*), D_{test} = 1] \\
&= E[E\{I(Y = 1) \mid X, A = a, \mu_{\widehat{\beta}}(X^*), D_{test} = 1\} \mid \mu_{\widehat{\beta}}(X^*), D_{test} = 1]
\end{aligned}$$

where the first line follows from the definition of $\psi_{\widehat{\beta}}$, the second from random sampling of the test set, the third from the law of iterated expectations, the fourth from the exchangeability condition, and the fifth from the consistency condition. Recall that X^* is a subset of X. For the second representation, we show that it is equivalent to the first

$$\begin{aligned}
\psi_{\widehat{\beta}} &= E[E\{I(Y = 1) \mid X, A = a, \mu_{\widehat{\beta}}(X^*), D_{test} = 1\} \mid \mu_{\widehat{\beta}}(X^*), D_{test} = 1] \\
&= E\left[E\left\{\frac{I(A = a)}{\Pr(A = a \mid X, \mu_{\widehat{\beta}}(X^*), D_{test} = 1)} I(Y = 1) \mid X, \mu_{\widehat{\beta}}(X^*), D_{test} = 1\right\} \mid \mu_{\widehat{\beta}}(X^*), D_{test} = 1\right] \\
&= E\left[\frac{I(A = a)}{\Pr(A = a \mid X, \mu_{\widehat{\beta}}(X^*), D_{test} = 1)} E\left\{I(Y = 1) \mid X, \mu_{\widehat{\beta}}(X^*), D_{test} = 1\right\} \mid \mu_{\widehat{\beta}}(X^*), D_{test} = 1\right] \\
&= E\left[\frac{I(A = a)}{\Pr(A = a \mid X, \mu_{\widehat{\beta}}(X^*), D_{test} = 1)} I(Y = 1) \mid \mu_{\widehat{\beta}}(X^*), D_{test} = 1\right]
\end{aligned}$$

where the second line follows from the definition of conditional expectation, the third removes the constant fraction outside expectation, and the last reverses the law of iterated expectations. ■

C.4.2 Estimation

Unlike previous sections, estimation of the full risk calibration curve using sample analogs of the identified expressions C.10 and C.11 is generally infeasible because they are conditional on a continuous risk score. Instead analysts typically perform either kernel or binned estimation of the calibration curve functional. In the case of the counterfactual risk calibration curve under a hypothetical intervention, the expression above suggest modifying these approaches either through the use of inverse probability weights or an outcome model.

C.5 Area under ROC curve

A final common metric for the performance of a risk prediction model $\mu_\beta(X^*)$ is the area under the receiver operating characteristic (ROC) curve, often referred to as simply the area under the curve (AUC). The AUC can be interpreted as the probability that a randomly sampled observation with the outcome has a higher predicted value than a randomly sampled observation without the outcome. In that sense, it is a measure of the discriminative ability of the model, i.e. the ability to distinguish between cases and noncases. For counterfactual predictions the relevant AUC though is the counterfactual AUC that would be observed under intervetion $A = a$, or

$$\psi_{\widehat{\beta}} = E[I\left(\mu_\beta(X_i^*) > \mu_\beta(X_j^*)\right) \mid Y_i^a = 1, Y_j^a = 0]. \tag{C.12}$$

C.5.1 Identification

Here we show that the counterfactual AUC is identified by the observed data functionals in the test set

$$\psi_{\widehat{\beta}} = \frac{\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right) h_a(X_i, X_j)\right]}{\mathrm{E}\left[h_a(X_i, X_j)\right]} \tag{C.13}$$

and

$$\psi_{\widehat{\beta}} = \frac{\mathrm{E}\left[\frac{I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*), Y_i=1, Y_j=0, A_i=a, A_j=a\right)}{e_a(X_i,X_j)}\right]}{\mathrm{E}\left[\frac{I\left(Y_i=1, Y_j=0, A_i=a, A_j=a\right)}{e_a(X_i,X_j)}\right]} \tag{C.14}$$

where the subscripts i and j denote a random pair of observations from the test set. We also define

$$h_a(X_i, X_j) = \Pr\left[Y_i = 1 \mid X_i, A_i = a, D_{test,i} = 1\right] \Pr\left[Y_j = 0 \mid X_j, A_j = a, D_{test,j} = 1\right]$$

and

$$e_a(X_i, X_j) = \Pr\left[A_i = a \mid X_i, D_{test,i} = 1\right] \Pr\left[A_j = a \mid X_j, D_{test,j} = 1\right]$$

for a pair of covariate vectors X_i and X_j.

To identify the AUC, we require a modified set of identification conditions, namely:

1. *Exchangeability.* $Y^a \perp\!\!\!\perp A \mid X$

2. *Consistency.* $Y^a = Y$ if $A = a$

3. *Positivity.* (i) $\Pr(A = a|X = x) > 0$ for all x that have positive density in $f(X, A = a)$, (ii) $\mathrm{E}\left[\Pr[Y = 1|X_i, A = a] \Pr[Y = 0|X_j, A = a]\right] > 0$, where i is a random observation that has the outcome and j is random observation without the outcome.

Proof. For the first representation we have

$$
\begin{aligned}
\psi_{\widehat{\beta}} &= \mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right) \mid Y_i^a = 1, Y_j^a = 0\right] \\
&= \frac{\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*), Y_i^a = 1, Y_j^a = 0\right)\right]}{\Pr\left[Y_i^a = 1, Y_j^a = 0\right]} \\
&= \frac{\mathrm{E}\left[\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*), Y_i^a = 1, Y_j^a = 0\right) \mid X_i, X_j\right]\right]}{\mathrm{E}\left[\Pr\left[Y_i^a = 1, Y_j^a = 0 \mid X_i, X_j\right]\right]} \\
&= \frac{\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right) \Pr\left[Y_i^a = 1, Y_j^a = 0 \mid X_i, X_j\right]\right]}{\mathrm{E}\left[\Pr\left[Y_i^a = 1, Y_j^a = 0 \mid X_i, X_j\right]\right]} \\
&= \frac{\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right) \Pr\left[Y_i^a = 1, Y_j^a = 0 \mid X_i, X_j, A_i = a, A_j = a\right]\right]}{\mathrm{E}\left[\Pr\left[Y_i^a = 1, Y_j^a = 0 \mid A_i = a, A_j = a, X_i, X_j\right]\right]} \\
&= \frac{\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right) \Pr\left[Y_i^a = 1 \mid X_i, A_i = a\right] \Pr\left[Y_j^a = 0 \mid X_j, A_j = a\right]\right]}{\mathrm{E}\left[\Pr\left[Y_i^a = 1 \mid X_i, A_i = a\right] \Pr\left[Y_j^a = 0 \mid X_j, A_j = a\right]\right]} \\
&= \frac{\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right) \Pr\left[Y_i = 1 \mid X_i, A_i = a\right] \Pr\left[Y_j = 0 \mid X_j, A_j = a\right]\right]}{\mathrm{E}\left[\Pr\left[Y_i = 1 \mid X_i, A_i = a\right] \Pr\left[Y_j = 0 \mid X_j, A_j = a\right]\right]} \\
&= \frac{\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right) \Pr\left[Y_i = 1 \mid X_i, A_i = a\right] \Pr\left[Y_j = 0 \mid X_j, A_j = a\right]\right]}{\mathrm{E}\left[\Pr\left[Y_i = 1 \mid X_i, A_i = a\right] \Pr\left[Y_j = 0 \mid X_j, A_j = a\right]\right]} \\
&= \frac{\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right) h_a(X_i, X_j)\right]}{\mathrm{E}\left[h_a(X_i, X_j)\right]}
\end{aligned}
$$

where the first line follows from the definition of $\psi_{\widehat{\beta}}$, the second from the definition of conditional probability, the third from the law of iterated expectations, the fourth from the definition of conditional expectation, the fifth from the exchangeability condition, the sixth from independence of potential outcomes, the seventh from the consistency condition, the eighth from random sampling of the test set, and the ninth applies the definition of $h_a(X_i, X_j)$. Recall that X^* is a subset of X. For the second representation, we will show that it is equivalent to the first. Starting from line five above

$$
\begin{aligned}
\psi_{\widehat{\beta}} &= \frac{\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right)\Pr\left[Y_i^a = 1, Y_j^a = 0 \mid X_i, X_j, A_i = a, A_j = a\right]\right]}{\mathrm{E}\left[\Pr\left[Y_i^a = 1, Y_j^a = 0 \mid A_i = a, A_j = a, X_i, X_j\right]\right]} \\
&= \frac{\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right)\Pr\left[Y_i = 1, Y_j = 0 \mid X_i, X_j, A_i = a, A_j = a\right]\right]}{\mathrm{E}\left[\Pr\left[Y_i = 1, Y_j = 0 \mid A_i = a, A_j = a, X_i, X_j\right]\right]} \\
&= \frac{\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right)\frac{\Pr\left[Y_i=1,Y_j=0,A_i=a,A_j=a|X_i,X_j\right]}{\Pr\left[A_i=a,A_j=a|X_i,X_j\right]}\right]}{\mathrm{E}\left[\frac{\Pr\left[Y_i=1,Y_j=0,A_i=a,A_j=a|X_i,X_j\right]}{\Pr\left[A_i=a,A_j=a|X_i,X_j\right]}\right]} \\
&= \frac{\mathrm{E}\left[\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right)\frac{\Pr\left[Y_i=1,Y_j=0,A_i=a,A_j=a|X_i,X_j\right]}{\Pr\left[A_i=a,A_j=a|X_i,X_j\right]} \mid X_i, X_j\right]\right]}{\mathrm{E}\left[\mathrm{E}\left[\frac{\Pr\left[Y_i=1,Y_j=0,A_i=a,A_j=a|X_i,X_j\right]}{\Pr\left[A_i=a,A_j=a|X_i,X_j\right]} \mid X_i, X_j\right]\right]} \\
&= \frac{\mathrm{E}\left[\frac{I\left(\mu_{\widehat{\beta}}(X_i^*)>\mu_{\widehat{\beta}}(X_j^*)\right)}{\Pr[A_i=a|X_i]\Pr\left[A_j=a|X_j\right]}\Pr\left[Y_i = 1, Y_j = 0, A_i = a, A_j = a \mid X_i, X_j\right]\right]}{\mathrm{E}\left[\frac{\Pr\left[Y_i=1,Y_j=0,A_i=a,A_j=a|X_i,X_j\right]}{\Pr[A_i=a|X_i]\Pr[A_j=a|X_j]}\right]} \\
&= \frac{\mathrm{E}\left[\frac{I\left(\mu_{\widehat{\beta}}(X_i^*)>\mu_{\widehat{\beta}}(X_j^*),Y_i=1,Y_j=0,A_i=a,A_j=a\right)}{\Pr[A_i=a|X_i]\Pr\left[A_j=a|X_j\right]}\right]}{\mathrm{E}\left[\frac{I\left(Y_i=1,Y_j=0,A_i=a,A_j=a\right)}{\Pr[A_i=a|X_i]\Pr[A_j=a|X_j]}\right]} \\
&= \frac{\mathrm{E}\left[\frac{I\left(\mu_{\widehat{\beta}}(X_i^*)>\mu_{\widehat{\beta}}(X_j^*),Y_i=1,Y_j=0,A_i=a,A_j=a\right)}{e_a(X_i,X_j)}\right]}{\mathrm{E}\left[\frac{I\left(Y_i=1,Y_j=0,A_i=a,A_j=a\right)}{e_a(X_i,X_j)}\right]}
\end{aligned}
$$

where the second line follows from consistency, the third from the definition of conditional probability, the fourth from iterated expectations, the fifth removes the constant fraction outside expectation, the sixth reverses the law of iterated expectations and the last applies random sampling of the test set and the definition of $e_a(X_i, X_j)$. ■

C.5.2 Plug-in estimation

Using sample analogs for the identified expressions C.13 and C.14, we obtain two plug-in estimators for the counterfactual AUC

$$\widehat{\psi}_{OM} = \frac{\sum_{i \neq j}^{n} \widehat{h}_a(X_i, X_j) I(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*), D_{test,i} = 1, D_{test,j} = 1)}{\sum_{i \neq j}^{n} \widehat{h}_a(X_i, X_j) I(D_{test,i} = 1, D_{test,j} = 1)}$$

and

$$\widehat{\psi}_{IPW} = \frac{\sum\limits_{i \neq j}^{n} \dfrac{I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*), Y_i = 1, Y_j = 0, A_i = a, A_j = a, D_{test,i} = 1, D_{test,j} = 1\right)}{\widehat{e}_a(X_i, X_j)}}{\sum\limits_{i \neq j}^{n} \dfrac{I\left(Y_i = 1, Y_j = 0, A_i = a, A_j = a, D_{test,i} = 1, D_{test,j} = 1\right)}{\widehat{e}_a(X_i, X_j)}}$$

where $\widehat{h}_a(X_i, X_j)$ is an estimator for $\Pr[Y_i = 1 | X_i, A_i = a, D_{test,i} = 1] \Pr[Y_j = 0 | X_j, A_j = a, D_{test,j} = 1]$ and $\widehat{e}_a(X_i, X_j)$ is an estimator for $\Pr[A_i = a | X_i, D_{test,i} = 1] \Pr[A_j = a | X_j, D_{test,j} = 1]$. Here, we call the first plug-in estimator the outcome model estimator $\widehat{\psi}_{OM}$ and the second the inverse probability weighted estimator $\widehat{\psi}_{IPW}$.

C.6 Additional application details

The Multi-Ethnic Study on Atherosclerosis (MESA) study is a population-based sample of 6,814 men and women aged 45 to 84 drawn from six communities (Baltimore; Chicago; Forsyth County, North Carolina; Los Angeles; New York; and St. Paul, Minnesota) in the United States between 2000 and 2002. The sampling procedure, design, and methods of the study have been described previously [68]. Study teams conducted five examination visits between 2000 and 2011 in 18 to 24 month intervals focused on the prevalence, correlates, and progression of subclinical cardiovascular disease. These examinations included assessments of lipid-lowering (primarily statins) and other medication use as well as cardiovascular risk factors such as systolic blood pressure, serum cholesterol, cigarette smoking, height, weight, and diabetes.

Our goal was to emulate a single-arm trial corresponding to the AHA guidelines on initiation of statin therapy for primary prevention of cardiovascular disease in the MESA cohort and use the emulated trial to develop a prediction model for the treatment-naive risk. The AHA guidelines stipulate that patients aged 40 to 75 with serum LDL cholesterol levels between 70 mg/dL and 190 mg/dL and no history of cardiovascular disease should initiate statins if their risk exceeds 7.5%. Therefore, we considered MESA participants who completed the baseline examination, had no recent history of statin use, no history of cardiovascular disease, and who met the criteria described in the guidelines (excluding the risk threshold) as eligible to participate in the trial. The primary endpoint was time to atherosclerotic cardiovascular disease (ASCVD), defined as nonfatal myocardial infarction, coronary heart disease death, or ischemic stroke.

Follow up began at the second examination cycle to enable a "wash out" period for statin use and to ensure adequate pre-treatment covariates to control confouding. We constructed a sequence of nested trials starting at each examination cycle from exam 2 through exam 5 and pooled the results from all 4 trials into a single analysis and used a robust variance estimator to account for correlation among duplicated participants. In each nested trial, we used the corresponding questionnaire to determine eligibility as well as statin initiators versus non-initiators. Because the exact timing of statin initiation was not known with precision, in each trial, we estimated the start of follow up for initiators and non-initators by drawing a random month between their current and previous examinations. We explored alternative definitions of the start of follow up in sensitivity analyses in the appendix. To mimic the targeted single-arm trial we limited to non-initiators for development of the prediction models.

C.6.1 Propensity score models

In the emulated single arm trial, statin initiation can be viewed as "non-adherence" which can be adjusted for by inverse probability weighting, therefore we censored participants when they initiated statins. To calculate the weights, we estimated two logistic regression models: one for the probability of remaining untreated given past covariate history (denominator model) and one for probability of remaining untreated given the selected baseline predictors (numerator model). In the denominator model we included the following covariates:

- *Demographic factors* - Age, gender, marital status, education, race/ethnicity, employment, health insurance status, depression, perceived discrimination, emotional support, anger and anxiety scales, and neighborhood score.
- *Risk factors* - Systolic and diastolic blood pressure, serum cholesterol levels (LDL, HDL, Triglycerides), hypertension, diabetes, waist circumference, smoking, alcohol consumption, exercise, family history of CVD, calcium score, hypertrophy on ECG, CRP, IL-6, number of pregnancies, oral contraceptive use, age of menopause.
- *Medication use* - Anti-hypertensive use, insulin use, daily aspirin use, anti-depressant use, vasodilator use, anti-arryhtmic use.

Time-varying demographic factors and risk factors were lagged such that values from the previous examination cycle were used.