Assessing the performance of counterfactual predictions

January 19, 2023

Abstract

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.

Keywords: causal inference, prediction model, treatment drop-in, transportability, model performance, machine learning

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 baseline may vary, particularly for models with a longer time horizon [1]. 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" [2, 3]. 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 [4].

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 [5–7]. 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 [7, 8].

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.

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

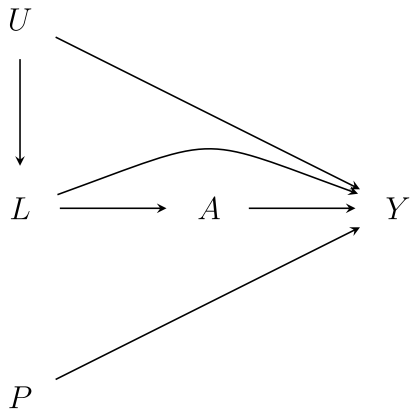


Figure 1: Example directed acyclic graph for prediction in a setting with a single time fixed treatment A over follow up.

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 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 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 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 [9–11]. 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.

4 Identifiability conditions

We will assume the following identifiability assumptions which have been described in more detail elsewhere [12–14].

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.

5 Tailoring a model for counterfactual predictions

As we show in section A.1 of the appendix, under the conditions above the expected potential outcome conditional on covariates X^* is identified by the expression

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

or, equivalently

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

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 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 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 [15]. 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.

6 Assessing model performance

Using the same conditions, in section A.2 of the appendix we show the model performance metric ψ is identifiable using data from the test set through the expression

$$\psi = E\left[E[(Y - \mu_{\widehat{\beta}}(X^*))^2 \mid X, A = a, D_{test} = 1] \mid D_{test} = 1\right] \tag{3}$$

or, equivalently using an inverse probability weighted expression

$$\psi = E\left[\frac{I(A = a)}{\Pr(A = a \mid X, D_{test} = 1)}(Y - \mu_{\widehat{\beta}}(X^*))^2 \mid D_{test} = 1\right] \tag{4}$$

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), an estimator of the target MSE is

$$\widehat{\psi}_{CL} = \frac{1}{n_{test}} \sum_{i=1}^{n} I(D_{test,i} = 1)\widehat{h}_a(X_i) \tag{5}$$

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 (4), an alternative weight-based estimator of the target MSE is

$$\widehat{\psi}_{IPW} = \frac{1}{n_{test}} \sum_{i=1}^{n} \frac{I(A_i = a, D_{test,i} = 1)}{\widehat{e}_a(X_i)} (Y_i - \mu_{\widehat{\beta}}(X_i^*))^2 \tag{6}$$

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 (5) relies on correctly specifying the model for the conditional loss and the weighting estimator (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 [16]. An alternative is to use a doubly-robust estimator which combines models for $\widehat{h}_a(X)$ and $\widehat{e}_a(X)$, such as

$$\widehat{\psi}_{DR} = \frac{1}{n_{test}} \sum_{i=1}^{n} I(D_{test,i} = 1) \left[\widehat{h}_a(X_i) + \frac{I(A_i = a)}{\widehat{e}_a(X_i)} \left\{ (Y - \mu_{\widehat{\beta}}(X_i^*))^2 - \widehat{h}_a(X_i) \right\} \right] \tag{7}$$

As we show in the Appendix, under mild regularity conditions [17], 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 $\sqrt{n}$.

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 [10]. 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 6 above. Failure to do so, can lead to sub-optimal selection with respect to the counterfactual prediction of interest.

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 [18, 19].

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 $X \sim$ 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 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. 19.5). In contrast, when using $\widehat{\psi}_{IPW}$, the inverse-probability weighted estimate of the MSE,

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.

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 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.

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 [20]. 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,

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.

i.e. $e_a(X) = \text{expit}(\alpha_0 + \sum_{i=1}^{10} \alpha_{1,i} X_{(i)})$.

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.

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.

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 [21]. 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 [22] 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.

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

Table 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.

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.

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 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.

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 [23]. Several authors have proposed model performance metrics which are similar to our own [24–28].

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 [10]. 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 [29] were available or under a more general proximal inference framework [30]. It's also possible to develop sensitivity analyses for exploring how violations of this assumption might affect model performance estimates [31].

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 [32, 33]. 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 [18, 19, 34, 35]. 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. 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).

2. 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).

3. 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).

4. 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).

5. 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).

6. 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).

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

8. 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).

9. 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).

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

11. 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).

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

13. 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).

14. 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).

15. 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).

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

17. 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).

18. Steingrimsson, J. A., Gatsonis, C. & Dahabreh, I. J. *Transporting a Prediction Model for Use in a New Target Population* 2021. arXiv: 2101.11182 [stat].

19. 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. arXiv: 2210.01980 [stat].

20. 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).

21. 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).

22. 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).

23. 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).

24. Schuler, A., Baiocchi, M., Tibshirani, R. & Shah, N. *A Comparison of Methods for Model Selection When Estimating Individual Treatment Effects* 2018. arXiv: 1804.05146 [cs, stat].

25. 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).

26. Xu, Y. & Yadlowsky, S. *Calibration Error for Heterogeneous Treatment Effects* 2022. arXiv: 2203.13364 [stat].

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

28. 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.

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

30. Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X. & Miao, W. *An Introduction to Proximal Causal Learning* 2020. arXiv: `2009.10982 [stat]`.

31. 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`.

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

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

34. 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).

35. 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).

A Time-fixed treatment initiation

A.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 4, namely

$$E[Y^a \mid X^*] = E[E[Y \mid X, A = a, D_{train} = 1] \mid X^*, D_{train} = 1] \tag{A1}$$

and

$$E[Y^a \mid X^*] = E\left[\frac{I(A = a)}{\Pr(A = a \mid X, D_{train} = 1)} Y \mid X^*, D_{train} = 1\right] \tag{A2}$$

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. ■

A.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

$$\psi_{\widehat{\beta}} = E\left(E[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid X, A = a, D_{test} = 1] \mid D_{test} = 1\right) \tag{A3}$$

and

$$\psi_{\widehat{\beta}} = E\left[\frac{I(A = a)}{\Pr(A = a \mid X, D_{test} = 1)} L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid D_{test} = 1\right] \tag{A4}$$

in the test set under the time-fixed setup described in section 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. ■

A.3 Plug-in estimation

Using sample analogs for the identified expressions A3 and A4, 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 $\widehat{\psi}_{IPW}$.

A.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

$$\psi_{\widehat{\beta}} = E\left\{E_{f^{int}}\left(E[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid X, A = a, D_{test} = 1] \mid D_{test} = 1\right)\right\} \tag{A5}$$

and

$$\psi_{\widehat{\beta}} = E\left[\frac{I(A=a)}{\Pr(A=a \mid X, D_{test}=1)} L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid D_{test}=1\right] \tag{A6}$$

in the test set under the time-fixed setup described in section 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.

B Time-varying treatment initiation

B.1 Set up

Here we extend the set up of section 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 A1

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

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

B.3 Identification of general loss functions

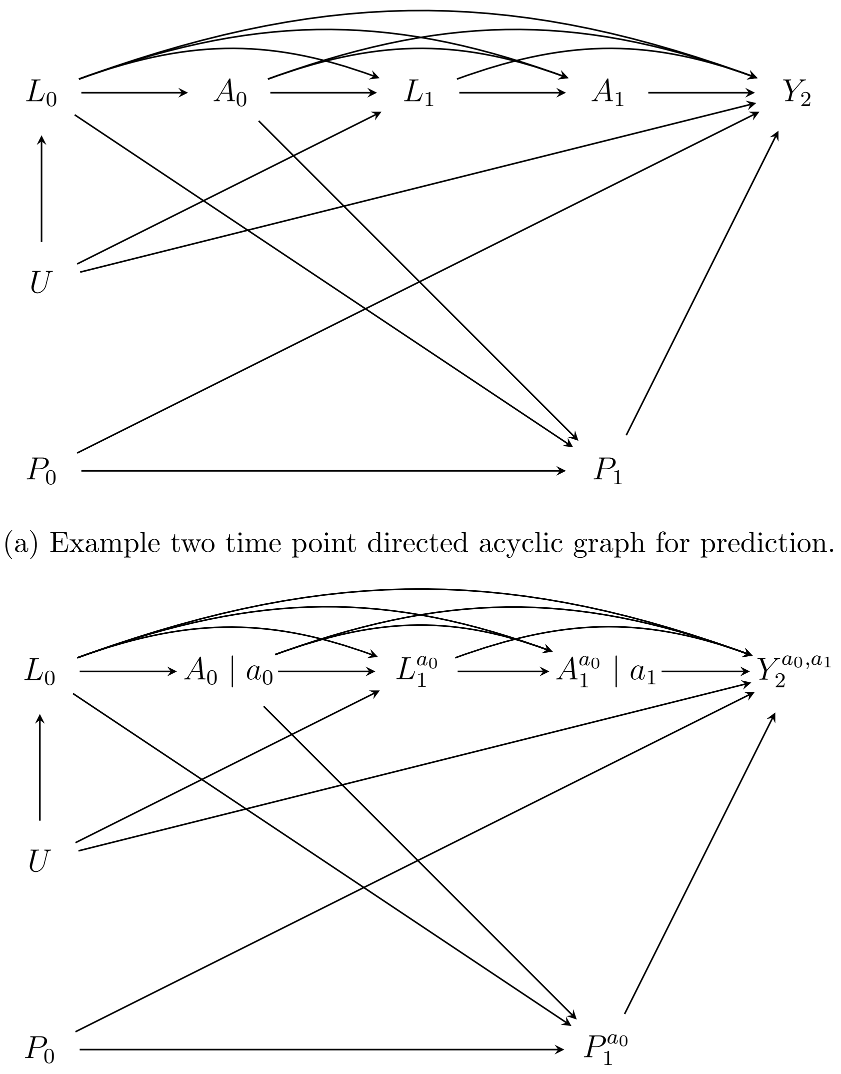
Under time-varying treatment initiation, the expected counterfactual loss for general loss function $L\{Y^g, \mu_{\widehat{\beta}}\}$ is identified by the functionals

$$\begin{aligned}\psi_{\widehat{\beta}} = E_{X_0}\Bigg[E_{X_1}\Bigg\{ \dots E_{X_{K-1}}\Bigg(E_{X_K}[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid \overline{X}_K, \overline{A}_K = \overline{a}^g_K, D_{test} = 1] \\ \mid \overline{X}_{K-1}, \overline{A}_{K-1} = \overline{a}^g_{K-1}, D_{test} = 1\Bigg) \dots \mid X_0, A_0 = a^g_0, D_{test} = 1\Bigg\} \mid D_{test} = 1\Bigg]\end{aligned} \tag{A7}$$

and

$$\psi_{\widehat{\beta}} = E\left[\frac{I(\overline{A}_K = \overline{a}^g_K, D_{test} = 1)}{\prod_{k=0}^K \Pr(A_k = a^g_k \mid \overline{X}_k, \overline{A}_{k-1} = \overline{a}^g_{k-1}, D_{test} = 1)} L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid D_{test} = 1\right] \tag{A8}$$

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



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

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

Figure A1: Example directed acyclic graph (DAG) and single world intervention graph (SWIG) for a two time point process.

Proof. For the first representation we have

$$\begin{aligned}
\psi_{\widehat{\beta}} &= E[L\{Y^g, \mu_{\widehat{\beta}}(X^*)\}] \\
&= E[L\{Y^g, \mu_{\widehat{\beta}}(X^*)\} \mid D_{test} = 1] \\
&= E(E[L\{Y^g, \mu_{\widehat{\beta}}(X^*)\} \mid X_0, D_{test} = 1] \mid D_{test} = 1) \\
&= E(E[L\{Y^g, \mu_{\widehat{\beta}}(X^*)\} \mid X_0, A_0 = a_0^g, D_{test} = 1] \mid D_{test} = 1)
\end{aligned}$$

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, 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 $\overline{A}_k = \overline{a}_k^g$, such that

$$\begin{aligned}
\psi_{\widehat{\beta}} = E_{X_0}\Bigg[E_{X_1}\Bigg\{ \ldots E_{X_{K-1}}\Bigg(E_{X_K}[L\{Y^g, \mu_{\widehat{\beta}}(X^*)\} \mid \overline{X}_K, \overline{A}_K = \overline{a}_K^g, D_{test} = 1] \\
\mid \overline{X}_{K-1}, \overline{A}_{K-1} = \overline{a}_{K-1}^g, D_{test} = 1\Bigg) \ldots \mid X_0, A_0 = a_0^g, D_{test} = 1\Bigg\} \mid D_{test} = 1\Bigg] \\
= E_{X_0}\Bigg[E_{X_1}\Bigg\{ \ldots E_{X_{K-1}}\Bigg(E_{X_K}[L\{Y, \mu_{\widehat{\beta}}(X^*)\} \mid \overline{X}_K, \overline{A}_K = \overline{a}_K^g, D_{test} = 1] \\
\mid \overline{X}_{K-1}, \overline{A}_{K-1} = \overline{a}_{K-1}^g, D_{test} = 1\Bigg) \ldots \mid X_0, A_0 = a_0^g, D_{test} = 1\Bigg\} \mid D_{test} = 1\Bigg]
\end{aligned}$$

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

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

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

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

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

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

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

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

. ■

B.4 Plug-in estimation

Using sample analogs for the identified expressions A7 and A8, 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

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

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

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

$\widehat{h}_a(X)$ 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 $\Pr(A_k = a_k^g \mid \overline{X}_k, \overline{A}_{k-1} = \overline{a}_{k-1}^g, D_{test} = 1)$. 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 $I(\overline{A}_K = \overline{a}_K^g, D_{test,i} = 1) = 1$ 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 Doubly robust estimators

C.1 Efficient influence function

As we've shown previously, under the identifiability conditions of section 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

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

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.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 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 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

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

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

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

C.3.1 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 $\psi = E\left[\frac{I(A=a)}{\Pr(A=a|X)} L(Y, \mu_{\widehat{\beta}}(X^*))\right]$

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

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

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

C.3.2 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.

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

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

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

By the central limit theorem,

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

completing the proof.

D 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

$$\psi_{\widehat{\beta}} = E[I(Y^a = 1) \mid \mu_{\widehat{\beta}}(X^*)]. \tag{A9}$$

D.1 Identification

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

$$\psi_{\widehat{\beta}} = E[E\{I(Y = 1) \mid X, A = a, \mu_{\widehat{\beta}}(X^*), D_{test} = 1\} \mid \mu_{\widehat{\beta}}(X^*), D_{test} = 1] \tag{A10}$$

and

$$\psi_{\widehat{\beta}} = E\left[\frac{I(A = a)}{\Pr(A = a \mid X, \mu_{\widehat{\beta}}(X^*), D_{test} = 1)} I(Y = 1) \mid \mu_{\widehat{\beta}}(X^*), D_{test} = 1\right] \tag{A11}$$

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. ■

D.2 Estimation

Unlike previous sections, estimation of the full risk calibration curve using sample analogs of the identified expressions A10 and A11 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.

E 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

$$\psi_{\widehat{\beta}} = E[I\left(\mu_\beta(X_i^*) > \mu_\beta(X_j^*)\right) \mid Y_i^a = 1, Y_j^a = 0]. \tag{A12}$$

E.1 Identification

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

$$\psi_{\widehat{\beta}} = \frac{\mathrm{E}\left[I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*)\right) h_a(X_i, X_j)\right]}{\mathrm{E}\left[h_a(X_i, X_j)\right]} \tag{A13}$$

and

$$\psi_{\widehat{\beta}} = \frac{\mathrm{E}\left[\frac{I\left(\mu_{\widehat{\beta}}(X_i^*) > \mu_{\widehat{\beta}}(X_j^*), Y_i=1, Y_j=0, A_i=a, A_j=a\right)}{e_a(X_i, X_j)}\right]}{\mathrm{E}\left[\frac{I(Y_i=1, Y_j=0, A_i=a, A_j=a)}{e_a(X_i, X_j)}\right]} \tag{A14}$$

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

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

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 $e_a(X_i, X_j)$. ■

E.2 Plug-in estimation

Using sample analogs for the identified expressions A13 and A14, 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

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

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 $\widehat{\psi}_{IPW}$.

F 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 [21]. 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.

F.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.