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 **??**, 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 **??**. 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

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

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 **??**. 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 **??** 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

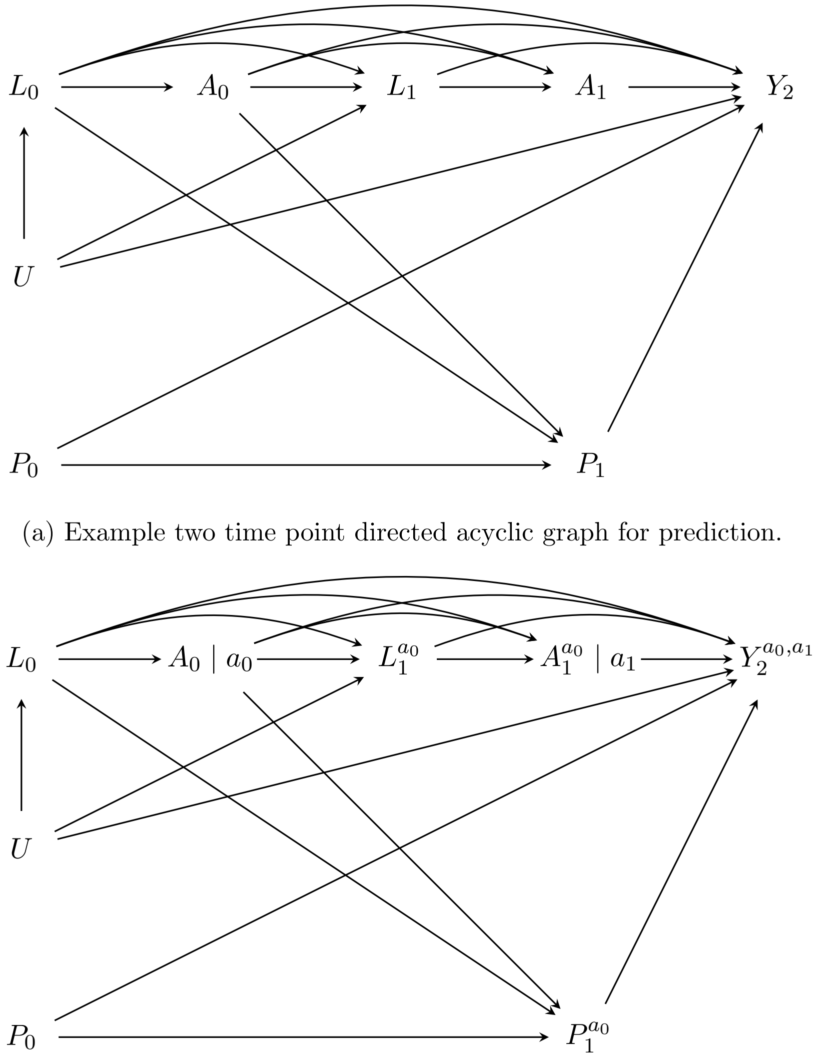
We now consider modified identifiability conditions under time-varying treatment initiation. For all k from 0 to K, we require

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

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



(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.

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.

Proof. For the first representation we have

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

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

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

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

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

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

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

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

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

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

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

. ■

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

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

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

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

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

C Doubly robust estimators

C.1 Efficient influence function

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

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

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

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

■

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

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

C.3 Asymptotic properties

The asymptotic properties of $\widehat{\psi}_{DR}$ are complicated by the presence of multiple nuisance functions and the fact that, in this section, we do not immediately assume a parametric model for the data. To derive the large sample properties we first define

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

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. The doubly robust estimator can be written as $\widehat{\psi}_{DR} = \frac{1}{n}\sum_{i=1}^{n} H\left(\widehat{e}_a\left(X_i\right), \widehat{h}_a\left(X_i\right)\right)$, where $\widehat{e}_a\left(X_i\right)$, and $\widehat{h}_a\left(X_i\right)$ are as defined before. Denote the limits of $\widehat{e}_a(X)$ and $\widehat{h}_a(X)$ as $e_a^*(X)$ and $h_a^*(X)$, respectively. Under correct model specification, the limits are equal to $e_a^*(X) = \Pr[A = a \mid X]$ and $h_a^*(X) = \mathrm{E}\left[L\left(Y, \mu\left(X^*\right)\right) \mid X, A=a\right]$.

We make the following four 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 [18].

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 restriction on the complexity of the functionals $\widehat{e}_a(X)$ and $\widehat{h}_a(X)$. As long as $\widehat{e}_a(X), \widehat{h}_a(X), e_a^*(X)$, and $h_a^*(X)$ are Donsker and all are uniformly bounded then Assumption

D1 holds by the Donsker preservation theorem. Many commonly used models such as generalized linear models fall within the Donsker class. This requirement can be further relaxed through sample-splitting, in which case more flexible machine learning algorithms 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 $\psi = E\left[\frac{I(A=a)}{\Pr(A=a|X)} L(Y, \mu_{\widehat{\beta}}(X^*))\right]$

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

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

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

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

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

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.

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

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

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

By the central limit theorem,

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

completing the proof.

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 \mid X_i, X_j]}{\Pr[A_i=a, A_j=a \mid X_i, X_j]}\right]}{\mathrm{E}\left[\frac{\Pr[Y_i=1, Y_j=0, A_i=a, A_j=a \mid X_i, X_j]}{\Pr[A_i=a, A_j=a \mid 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 \mid X_i, X_j]}{\Pr[A_i=a, A_j=a \mid 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 \mid X_i, X_j]}{\Pr[A_i=a, A_j=a \mid 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 \mid X_i]\Pr[A_j=a \mid 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 \mid X_i, X_j]}{\Pr[A_i=a \mid X_i]\Pr[A_j=a \mid 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 \mid X_i]\Pr[A_j=a \mid X_j]}\right]}{\mathrm{E}\left[\frac{I(Y_i=1, Y_j=0, A_i=a, A_j=a)}{\Pr[A_i=a \mid X_i]\Pr[A_j=a \mid 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 $\widehat{\psi}_{IPW}$.