# Estimation and evaluation of counterfactual prediction models

Christopher B. Boyer, Ph.D.
Department of Quantitative Health Sciences

April 7, 2025

Cleveland Clinic

# A common clinical science task

Predicting a patient's *future* health status based on their current state and medical history. E.g.,

$\Longrightarrow$ What is a patient's 10-year risk of cardiovascular disease ($Y_i$) based on their demographics and current blood pressure and lipid levels ($X^*$)?

# A common clinical science task

Predicting a patient's *future* health status based on their current state and medical history. E.g.,

$\implies$ What is a patient's 10-year risk of cardiovascular disease ($Y_i$) based on their demographics and current blood pressure and lipid levels ($X^*$)?

We assume this is well approximated by the probability of the outcome among some reference class composed of those with same/similar $X^*$, i.e.

$$Y_i \sim f(Y|X^*)$$

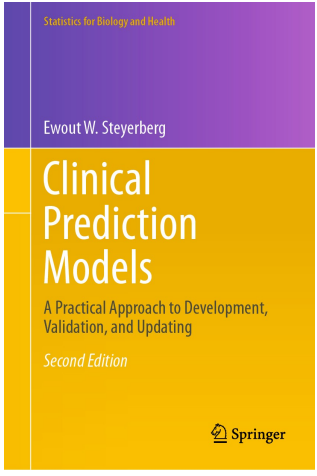
and therefore amenable to statistical modeling of the form:

$$\mathrm{E}[Y|X^*] = g(X^*; \beta)$$

# Clinical prediction modeling

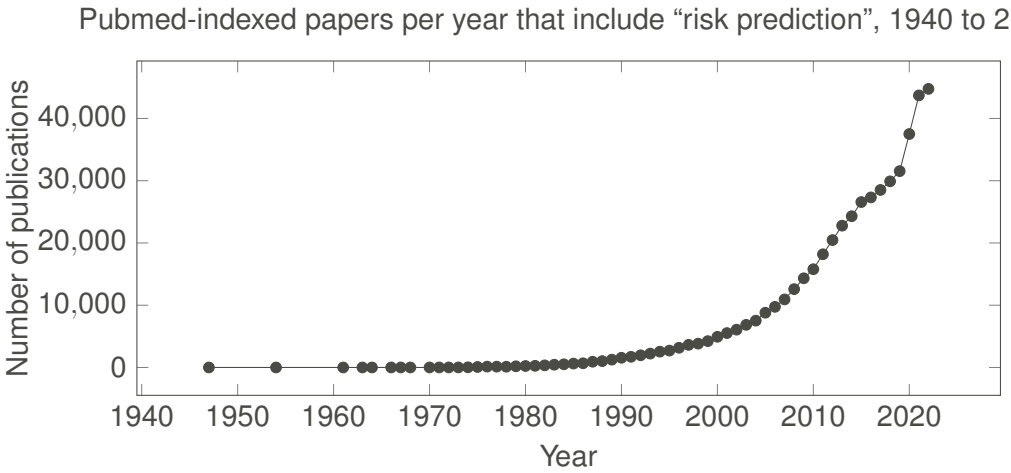
Paradigm:

1. Collect data from a sample of patients.
2. Train a model for the prediction estimand based on available clinical inputs.
3. Evaluate model performance in an independent sample.
4. Apply model prospectively to new patients.
5. Monitor model performance over time and in new settings.




Cleveland Clinic

# There's a model for that...




Cleveland Clinic

# What is wrong with the conventional approach?

Conventional (hereafter **“factual”**) prediction methods have a lot to offer:

- A large and mature set of methods, based on traditional regression or more flexible (but data hungry) machine learning approaches.

- Framework that focuses on agnostic evaluation of model performance rather than whether the model is “correct”.

# What is wrong with the conventional approach?

Conventional (hereafter **“factual”**) prediction methods have a lot to offer:

- A large and mature set of methods, based on traditional regression or more flexible (but data hungry) machine learning approaches.
- Framework that focuses on agnostic evaluation of model performance rather than whether the model is “correct”.

Bold claim: often these models are applied as sophisticated answers to the wrong question.

# A more useful input to clinical decision making

Many actual use cases involve “what if” questions about predicted states under *hypothetical interventions* (hereafter **“counterfactual”** prediction).

# A more useful input to clinical decision making

Many actual use cases involve “what if” questions about predicted states under *hypothetical interventions* (hereafter **“counterfactual”** prediction).

Case 1: Decision support

What would the patient’s 10-year risk of cardiovascular disease be if...

# A more useful input to clinical decision making

Many actual use cases involve “what if” questions about predicted states under *hypothetical interventions* (hereafter **“counterfactual”** prediction).

Case 1: Decision support

What would the patient’s 10-year risk of cardiovascular disease be if...

$\Longrightarrow$ they start statins today?

$\Longrightarrow$ they start statins once they have two successive visits with LDL $>$ 190 mg/dl?

$\Longrightarrow$ they start lifestyle interventions first and then statins if LDL $>$ 160 mg/dl?

$\Longrightarrow$ they never start statins?

# A more useful input to clinical decision making

Case 2: Changes in treatment patterns

E.g., a model trained when 5% of patients receive treatment after baseline is now being used in a setting where 60% recieve treatment due to changes in treatment guidelines.

# A more useful input to clinical decision making

Case 2: Changes in treatment patterns

E.g., a model trained when 5% of patients receive treatment after baseline is now being used in a setting where 60% recieve treatment due to changes in treatment guidelines.

Case 3: Removing pernicious influences

E.g., a model that targets the risk of death in the absence of surgery is trained in an observational setting where some of the patients ultimately received surgery.

# Our contribution

- How to estimate (build models for) these counterfactual outcomes using data from an observational study or randomized trial (or both)?
- How to evaluate these models given that the outcome may not be fully observed?
- How to allow for effective separation of prediction and causal inference tasks?
- What are the conditions under which all of this works and what can we do if they do not hold?

# A framework for counterfactual prediction

1. What is the clinical decision to be made?
2. Study design and sampling
3. Causal model and estimand
4. Identifiability
5. Estimation and inference
6. Performance evaluation
7. Sensitivity analysis
8. Communication and dissemination

# (1) Clinical decision

Example:

For a patient admitted to the hospital with characteristics $X^*$, what is the 14-day risk of venous thromboembolism if they receive heparin prophylaxis?

# (1) Clinical decision

Example:

For a patient admitted to the hospital with characteristics $X^*$, what is the 14-day risk of venous thromboembolism if they receive heparin prophylaxis?

Questions:

- How is prophylaxis defined (dosage, length, route)?
- How to handle death (competing risk, composite)?
- Should we include risk of bleeding?

# (2) Study design and sampling

We have data $O$ from $n$ participants for model development where

$$O = \{(X_i, A_i, Y_i, D_i) : i = 1, \ldots, n\}$$

and we define

- $X$ : vector of covariates measured at baseline.
- $A$ : a binary indicator of treatment initiation post baseline.
- $Y$ : a clinical outcome (here assumed to be binary).
- $D$ : a split indicator, where $D = 1$ is test and $D = 0$ is training.

Note: $X$ here includes possible confounders $L$ as well as predictors of outcome $P$ that are not confounders, i.e. $X_i = (L_i, P_i)$.

# (3) Causal model and estimand

Let $Y^a$ be the potential outcome under an intervention that sets $A$ to $a \in \mathcal{A}$.

# (3) Causal model and estimand

Let $Y^a$ be the potential outcome under an intervention that sets $A$ to $a \in \mathcal{A}$.

Goal: build a model $g_a(X^*)$ for the conditional expectation of the potential outcome in the target population, i.e.

$$g_a(X^*) = \mathrm{E}[Y^a | X^* = x^*],$$

where $X^*$ is a subset of all covariates $X$ to allow for separation between covariates for prediction and those necessary for causal identification.

# (3) Causal model and estimand

Let $Y^a$ be the potential outcome under an intervention that sets $A$ to $a \in \mathcal{A}$.

Goal: build a model $g_a(X^*)$ for the conditional expectation of the potential outcome in the target population, i.e.

$$g_a(X^*) = \mathrm{E}[Y^a | X^* = x^*],$$

where $X^*$ is a subset of all covariates $X$ to allow for separation between covariates for prediction and those necessary for causal identification.

Note: for now, we assume that our data $O$ are sampled directly from the target population.

# (4) Identifiability conditions

To identify our counterfactual prediction estimand, we require the following identifiability assumptions:

A1. *Consistency.* If $A = a$, then $Y^a = Y$

A2. *Conditional exchangeability.* $Y^a \perp\!\!\!\perp A \mid X$

A3. *Positivity.* For all $x$ with positive density, i.e. $f_X(x) > 0$, $\Pr[A = a \mid X = x] > 0$

# (4) Identifiability conditions

To identify our counterfactual prediction estimand, we require the following identifiability assumptions:

A1. *Consistency.* If $A = a$, then $Y^a = Y$

A2. *Conditional exchangeability.* $Y^a \perp\!\!\!\perp A \mid X$

A3. *Positivity.* For all $x$ with positive density, i.e. $f_X(x) > 0$, $\Pr[A = a \mid X = x] > 0$

Note: we also assume that, by design, the train/test split is random such that

$$(Y, A, X) \perp\!\!\!\perp D$$

# (4) Identifiability conditions

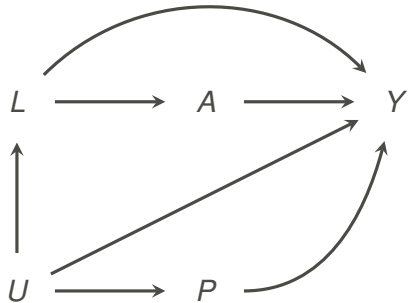



Figure: Causal directed acyclic graph showing law of observed data where identifiability conditions hold.

# (5) Estimation and inference

Under the conditions above $\mathrm{E}[Y^a|X^*]$ is identified by

$$g_a(X^*) \equiv \mathrm{E}[\mathrm{E}[Y \mid X, A = a, D = 0] \mid X^*, D = 0] \tag{1}$$

# (5) Estimation and inference

Under the conditions above $\mathrm{E}[Y^a|X^*]$ is identified by

$$g_a(X^*) \equiv \mathrm{E}[\mathrm{E}[Y \mid X, A = a, D = 0] \mid X^*, D = 0] \tag{1}$$

or, equivalently, using an inverse probability weighted expression

$$g_a(X^*) = \mathrm{E}\left[\frac{I(A = a)}{\Pr[A = a \mid X, D = 0]} Y \middle| X^*, D = 0\right] \tag{2}$$

The two expressions for $g_a(X^*)$ suggest two possible approaches for developing a model for counterfactual prediction from the training data.

# (5) Estimation and inference

Approach 1: Outcome modeling

1. Subset to participants with $A = a$ in the training data and fit a model $g_a(X)$ for the observed $Y$ conditional $X$, i.e. $\mathrm{E}[Y|X, A = a, D = 0] = g_a(X)$.

# (5) Estimation and inference

Approach 1: Outcome modeling

1. Subset to participants with $A = a$ in the training data and fit a model $g_a(X)$ for the observed $Y$ conditional $X$, i.e. $\mathrm{E}[Y|X, A = a, D = 0] = g_a(X)$.

2. Marginalize (standardize) over the covariates in $X$ that are not in $X^*$. When the dimension of $X^*$ is small, this can be done nonparametrically, otherwise, a second step is needed, either:

# (5) Estimation and inference

Approach 1: Outcome modeling

1. Subset to participants with $A = a$ in the training data and fit a model $g_a(X)$ for the observed $Y$ conditional $X$, i.e. $\mathrm{E}[Y|X, A = a, D = 0] = g_a(X)$.

2. Marginalize (standardize) over the covariates in $X$ that are not in $X^*$. When the dimension of $X^*$ is small, this can be done nonparametrically, otherwise, a second step is needed, either:

Option 1. Model the estimated $\widehat{g}_a(X)$ as a function of $X^*$, i.e. $\mathrm{E}[\widehat{g}_a(X)|X^*, D = 0] = g_a(X^*)$.

# (5) Estimation and inference

Approach 1: Outcome modeling

1. Subset to participants with $A = a$ in the training data and fit a model $g_a(X)$ for the observed $Y$ conditional $X$, i.e. $\mathrm{E}[Y|X, A = a, D = 0] = g_a(X)$.

2. Marginalize (standardize) over the covariates in $X$ that are not in $X^*$. When the dimension of $X^*$ is small, this can be done nonparametrically, otherwise, a second step is needed, either:

Option 1. Model the estimated $\widehat{g}_a(X)$ as a function of $X^*$, i.e. $\mathrm{E}[\widehat{g}_a(X)|X^*, D = 0] = g_a(X^*)$.

Option 2. Model the conditional density of $X$ given $X^*$, i.e. $f(X|X^*, D = 0) = h(X^*)$, and generate predictions from $\int g_a(x)h(x^*)dx$.

# (5) Estimation and inference

Approach 2: Inverse probability weighting

1. For each individual, create weights $W(a)$ equal to the probability of receiving treatment level $A = a$ conditional on covariates $X$ necessary to ensure exchangeability, i.e., sample analogs of

$$W(a) = \frac{I(A = a)}{\Pr[A = a \mid X, D = 0]}$$

# (5) Estimation and inference

Approach 2: Inverse probability weighting

1. For each individual, create weights $W(a)$ equal to the probability of receiving treatment level $A = a$ conditional on covariates $X$ necessary to ensure exchangeability, i.e., sample analogs of

$$W(a) = \frac{I(A = a)}{\Pr[A = a \mid X, D = 0]}$$

2. Fit model $g(X^*)$ using weighted optimization based on $W(a)$, for instance via weighted maximum likelihood.

# (6) Performance evaluation

To evaluate performance, one typically chooses a **performance statistic**, $\psi$, that compares model predictions with the observed outcome in a **hold out** dataset:

- E.g., mean squared error, c-statistic, calibration curve, L1 loss, etc.

# (6) Performance evaluation

To evaluate performance, one typically chooses a **performance statistic**, $\psi$, that compares model predictions with the observed outcome in a **hold out** dataset:

- E.g., mean squared error, c-statistic, calibration curve, L1 loss, etc.

The problem: the potential outcome $Y^a$ is not fully "observed", e.g. we could define a counterfactual MSE

$$\psi(a) \equiv \mathrm{E}[(Y^a - \widehat{g}(X^*))^2]$$

but we don't have $Y^a$ for everyone

# (6) Performance evaluation

To evaluate performance, one typically chooses a **performance statistic**, $\psi$, that compares model predictions with the observed outcome in a **hold out** dataset:

- E.g., mean squared error, c-statistic, calibration curve, L1 loss, etc.

The problem: the potential outcome $Y^a$ is not fully "observed", e.g. we could define a counterfactual MSE

$$\psi(a) \equiv \mathrm{E}[(Y^a - \widehat{g}(X^*))^2]$$

but we don't have $Y^a$ for everyone

$\implies$ Nonetheless, under certain assumptions, $\psi(a)$ may still be identified from the observed data!

# (6) Performance evaluation

The counterfactual MSE $\psi(a)$ is identifiable under A1-A3 using data from the test set through the expression

$$\psi(a) \equiv \mathrm{E}\left[\mathrm{E}[(Y - \widehat{g}(X^*))^2 \mid X, A = a, D = 1] \mid D = 1\right] \tag{3}$$

# (6) Performance evaluation

The counterfactual MSE $\psi(a)$ is identifiable under A1-A3 using data from the test set through the expression

$$\psi(a) \equiv \mathrm{E}\left[\mathrm{E}[(Y - \widehat{g}(X^*))^2 \mid X, A = a, D = 1] \mid D = 1\right] \tag{3}$$

or, equivalently, using an inverse probability weighted expression,

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

regardless of whether the model $\widehat{g}(X^*)$ is correctly specified.

# (6) Performance evaluation

Using sample analogs for expression 3, we obtain the **conditional loss estimator**

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

where $\widehat{h}_a(X)$ is an estimator for $E[(Y - \widehat{g}(X^*))^2 \mid X, A = a, D = 1]$.

Steps:

# (6) Performance evaluation

Using sample analogs for expression 3, we obtain the **conditional loss estimator**

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

where $\widehat{h}_a(X)$ is an estimator for $E[(Y - \widehat{g}(X^*))^2 \mid X, A = a, D = 1]$.

Steps:

1. Subset to those with $A = a$ in the test set.

# (6) Performance evaluation

Using sample analogs for expression 3, we obtain the **conditional loss estimator**

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

where $\widehat{h}_a(X)$ is an estimator for $E[(Y - \widehat{g}(X^*))^2 \mid X, A = a, D = 1]$.

Steps:

1. Subset to those with $A = a$ in the test set.
2. Estimate a model for the outcome $(Y - \widehat{g}(X^*))^2$ conditional on full $X$ using either statistical or machine learning model.

# (6) Performance evaluation

Using sample analogs for expression 3, we obtain the **conditional loss estimator**

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

where $\widehat{h}_a(X)$ is an estimator for $E[(Y - \widehat{g}(X^*))^2 \mid X, A = a, D = 1]$.

Steps:

1. Subset to those with $A = a$ in the test set.
2. Estimate a model for the outcome $(Y - \widehat{g}(X^*))^2$ conditional on full $X$ using either statistical or machine learning model.
3. Use the model to predict outcome under $A = a$ for everyone in the test set, take the sample average.

# (6) Performance evaluation

Using sample analogs for expression 4, we obtain the **inverse probability weighting estimator**

$$\widehat{\psi}_{IPW} = \frac{1}{n_{test}} \sum_{i=1}^{n} \frac{I(A_i = a, D_i = 1)}{\widehat{e}_a(X_i)} (Y - \widehat{g}(X^*))^2$$

where $\widehat{e}_a(X)$ is an estimator for $\Pr[A = a \mid X, D = 1]$.

Steps:

# (6) Performance evaluation

Using sample analogs for expression 4, we obtain the **inverse probability weighting estimator**

$$\widehat{\psi}_{IPW} = \frac{1}{n_{test}} \sum_{i=1}^{n} \frac{I(A_i = a, D_i = 1)}{\widehat{e}_a(X_i)} (Y - \widehat{g}(X^*))^2$$

where $\widehat{e}_a(X)$ is an estimator for $\Pr[A = a \mid X, D = 1]$.

Steps:

1. Estimate a model for the probability of treatment in the test set, i.e. $e_a(X) = \Pr[A = a \mid X, D = 1]$.

# (6) Performance evaluation

Using sample analogs for expression 4, we obtain the **inverse probability weighting estimator**

$$\widehat{\psi}_{IPW} = \frac{1}{n_{test}} \sum_{i=1}^{n} \frac{I(A_i = a, D_i = 1)}{\widehat{e}_a(X_i)} (Y - \widehat{g}(X^*))^2$$

where $\widehat{e}_a(X)$ is an estimator for $\Pr[A = a \mid X, D = 1]$.

Steps:

1. Estimate a model for the probability of treatment in the test set, i.e. $e_a(X) = \Pr[A = a \mid X, D = 1]$.
2. Form inverse probability weights $W(a) = \dfrac{I(A = a)}{e_a(X)}$

# (6) Performance evaluation

Using sample analogs for expression 4, we obtain the **inverse probability weighting estimator**

$$\widehat{\psi}_{IPW} = \frac{1}{n_{test}} \sum_{i=1}^{n} \frac{I(A_i = a, D_i = 1)}{\widehat{e}_a(X_i)} (Y - \widehat{g}(X^*))^2$$

where $\widehat{e}_a(X)$ is an estimator for $\Pr[A = a \mid X, D = 1]$.

Steps:

1. Estimate a model for the probability of treatment in the test set, i.e. $e_a(X) = \Pr[A = a \mid X, D = 1]$.
2. Form inverse probability weights $W(a) = \dfrac{I(A = a)}{e_a(X)}$
3. Calculate the weighted mean of $(Y - \widehat{g}(X^*))^2$ in the test set using weights from previous step.

# (6) Performance evaluation

We can also construct the **doubly-robust estimator**

$$\widehat{\psi}_{DR} = \frac{1}{n_{test}} \sum_{i=1}^{n} I(D_i = 1) \left[ \widehat{h}_a(X_i) + \frac{I(A_i = a)}{\widehat{e}_a(X_i)} \left\{ (Y - \widehat{g}(X^*))^2 - \widehat{h}_a(X) \right\} \right]$$

which combines models for the conditional loss and the probability of treatment.

# (6) Performance evaluation

We can also construct the **doubly-robust estimator**

$$\widehat{\psi}_{DR} = \frac{1}{n_{test}} \sum_{i=1}^{n} I(D_i = 1) \left[ \widehat{h}_a(X_i) + \frac{I(A_i = a)}{\widehat{e}_a(X_i)} \left\{ (Y - \widehat{g}(X^*))^2 - \widehat{h}_a(X) \right\} \right]$$

which combines models for the conditional loss and the probability of treatment.

$\Longrightarrow$ The doubly-robust estimator will be consistent if either one of $h_a(X)$ or $e_a(X)$ (or both) is correctly specified!

# (6) Performance evaluation

We can also construct the **doubly-robust estimator**

$$\widehat{\psi}_{DR} = \frac{1}{n_{test}} \sum_{i=1}^{n} I(D_i = 1) \left[ \widehat{h}_a(X_i) + \frac{I(A_i = a)}{\widehat{e}_a(X_i)} \left\{ (Y - \widehat{g}(X^*))^2 - \widehat{h}_a(X) \right\} \right]$$

which combines models for the conditional loss and the probability of treatment.

$\implies$ The doubly-robust estimator will be consistent if either one of $h_a(X)$ or $e_a(X)$ (or both) is correctly specified!

It also permits the use of nonparameteric or machine learning estimators that converge at rate less than $\sqrt{n}$ due to product of the errors in empirical process term.

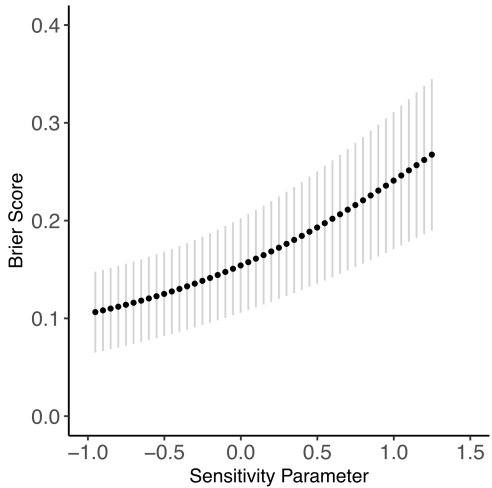
# (7) Sensitivity analysis

Exponential tilt model:

$$f(Y^a = y \mid X) \propto e^{\eta q(y)} f(Y = y \mid X)$$

Steps:

1. Specify grid of $\eta$ and function $q(\cdot)$ based on subject matter knowledge.
2. Re-estimate model or performance statistic under each $\eta$.




Cleveland Clinic

# (8) Communication and dissemination

How do we actually communicate/deploy these models?

# (8) Communication and dissemination

How do we actually communicate/deploy these models?

Approach 1: Outcome modeling

$\Longrightarrow$ when using two stage estimation (option 1) report the coefficients from second-stage pseudo-outcome regression, i.e. $\mathrm{E}[\widehat{g}_a(X)|X^*, D = 0] = g_a(X^*)$.

$\Longrightarrow$ when modeling the density (option 2) evaluate the integral for a grid of $X^*$ values.

# (8) Communication and dissemination

How do we actually communicate/deploy these models?

Approach 1: Outcome modeling

$\Longrightarrow$ when using two stage estimation (option 1) report the coefficients from second-stage pseudo-outcome regression, i.e. $\mathrm{E}[\widehat{g}_a(X)|X^*, D = 0] = g_a(X^*)$.

$\Longrightarrow$ when modeling the density (option 2) evaluate the integral for a grid of $X^*$ values.

Approach 2: Inverse probability weighting

$\Longrightarrow$ report the coefficients from the weighted regression.

If using black-box approaches (e.g. neural nets, random forests, etc) you can still deploy as you would previously.

# Empirical example

MESA = Multi-Ethnic Study of Atherosclerosis

Longitudinal cohort of 6,814 participants aged 45 to 84 from six communities and one of the validation datasets for the development of the pooled cohort equations.

Outcome: 10-year risk of ASCVD

Intervention: withhold statins over the follow up period ($A = 0$).

# Empirical example

Prediction models, $g(X^*)$:

- logit = main effects logistic regression (factual)
- IPW = logistic regression for treatment initiation + weighted main effects logistic regression (counterfactual)

Predictors ($X^*$): age, sex, smoking status, diabetes history, systolic blood pressure, anti-hypertensive medication use and total and HDL serum cholesterol levels.

Covariates ($X$): baseline demographics (12), risk factors (19), and medication use (6)

# Empirical example: do we think identification is credible?

Table: Intention-to-treat and per protocol effects of statin therapy in emulation compared to benchmark trial.

| | 5-year ASCVD | | 10-year ASCVD | |
|---|---|---|---|---|
| | HR | 95% CI | HR | 95% CI |
| Target Trial Emulation: MESA | | | | |
| ITT | **0.79** | **(0.65, 0.93)** | 0.70 | (0.56, 0.88) |
| Per protocol | 0.68 | (0.48, 0.94) | 0.60 | (0.39, 0.92) |
| Benchmark Trial: HPS | | | | |
| ITT | **0.76** | **(0.72, 0.81)** | | |

HR = Hazard Ratio, CI = Confidence Interval

# Empirical example: model fit

| | Factual (Logit) | | | Counterfactual (IPW) | | |
|---|---|---|---|---|---|---|
| Characteristic ($X^*$) | OR | 95% CI | p-value | OR | 95% CI | p-value |
| age | 1.27 | (1.18, 1.37) | <0.001 | 1.20 | (1.11, 1.30) | <0.001 |
| sex | 1.64 | (1.27, 2.13) | <0.001 | 1.59 | (1.21, 2.11) | 0.001 |
| smoker | 1.86 | (1.41, 2.46) | <0.001 | 1.62 | (1.19, 2.16) | 0.002 |
| diabetes | 1.28 | (1.00, 1.63) | 0.051 | 1.52 | (1.17, 1.98) | 0.002 |
| sbp | 1.25 | (1.15, 1.36) | <0.001 | 1.27 | (1.16, 1.39) | <0.001 |
| hdl | 0.81 | (0.73, 0.89) | <0.001 | 0.75 | (0.67, 0.84) | <0.001 |
| chol | 1.03 | (1.00, 1.06) | 0.034 | 1.09 | (1.06, 1.13) | <0.001 |
| hyp meds. | 1.35 | (1.04, 1.74) | 0.025 | 1.57 | (1.16, 2.11) | 0.003 |
| sbp $\times$ hyp meds | 0.83 | (0.75, 0.93) | 0.002 | 0.88 | (0.78, 0.99) | 0.039 |

OR = Odds Ratio, CI = Confidence Interval

# Empirical example: model performance

Performance statistics ($\psi$): MSE, AUC

Estimators ($\widehat{\psi}$):

- CL = main effects logistic regression for $h_a(X)$
- IPW = main effects logistic regression for $e_a(X)$
- DR = combine $h_a(X)$ + $e_a(X)$

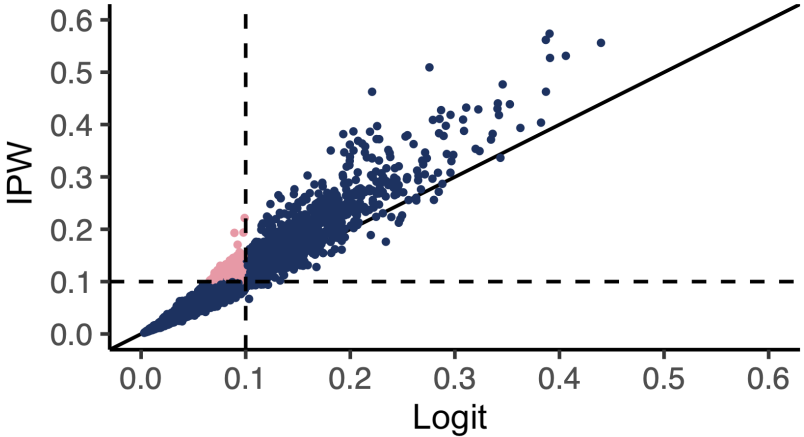
Covariates ($X$): same as before

# Empirical example: model performance

| | MSE | | AUC | |
|---|---|---|---|---|
| Estimator | Logit | IPW | Logit | IPW |
| Naïve | **0.069** | **0.072** | **0.710** | **0.708** |
| | (0.003) | (0.003) | (0.013) | (0.014) |
| CL | 0.086 | 0.085 | 0.719 | 0.727 |
| | (0.005) | (0.004) | (0.015) | (0.015) |
| IPW | 0.109 | 0.099 | 0.753 | 0.778 |
| | (0.013) | (0.009) | (0.025) | (0.029) |
| DR | 0.090 | 0.087 | 0.740 | 0.751 |
| | (0.006) | (0.005) | (0.023) | (0.023) |

The columns refer to the posited prediction model. Standard error estimates are shown in parentheses obtained via 1000 bootstrap replicates.

# What is the clinical significance?




Cleveland Clinic

# An alternative approach: transporting from a trial

In many cases, we may suspect that the conditional exchangeability assumption is violated in an observational study (i.e. **there's likely unmeasured confounding**).

# An alternative approach: transporting from a trial

In many cases, we may suspect that the conditional exchangeability assumption is violated in an observational study (i.e. **there's likely unmeasured confounding**).

Alternative: use data from a trial in which *A* is randomized.

# An alternative approach: transporting from a trial

In many cases, we may suspect that the conditional exchangeability assumption is violated in an observational study (i.e. **there's likely unmeasured confounding**).

Alternative: use data from a trial in which $A$ is randomized.

However, the trial setting will generally differ from the target setting where the model is applied.

$\Longrightarrow$ E.g., due to eligibility, location, participation, blinding, treatment delivery, etc.

# An alternative approach: transporting from a trial

In many cases, we may suspect that the conditional exchangeability assumption is violated in an observational study (i.e. **there's likely unmeasured confounding**).

Alternative: use data from a trial in which $A$ is randomized.

However, the trial setting will generally differ from the target setting where the model is applied.

$\Longrightarrow$ E.g., due to eligibility, location, participation, blinding, treatment delivery, etc.

Our contribution: methods and identifiability criteria for **"transporting"** a counterfactual prediction model from a trial to the target population and evaluating its performance.

# (2) Study design and sampling

We have data $O_1$ from $n_1$ participants **in a trial** where

$$O_1 = \{(S_i = 1, X_i, A_i, Y_i, D_i) : i = 1, \ldots, n_1\}$$

We also have covariate data $O_0$ from $n_0$ participants **in the target population** where

$$O_0 = \{(S_i = 0, X_i, D_i) : i = 1, \ldots, n_0\}$$

We define $X$, $A$, $Y$, and $D$ as previously and we have

- $S$ : an indicator of data source, where $S = 1$ is trial and $S = 0$ is target population.

# (2) Study design and sampling

Target population sample ($O_0$) :

| ID | X | S | A | Y |
|---|---|---|---|---|
| 1 | 15.2 | 0 | – | – |
| 2 | 0.5 | 0 | – | – |
| 3 | 4.7 | 0 | – | – |

+

Trial sample ($O_1$) :

| ID | X | S | A | Y |
|---|---|---|---|---|
| 4 | 2.3 | 1 | 1 | 10 |
| 5 | 14.2 | 1 | 0 | 20 |
| 6 | 8.9 | 1 | 1 | 30 |

# (3) Causal model and estimand

Our goal is now to **"transport"** a model $g_a(X^*)$, that is, to target the conditional expectation of the counterfactual outcome in the target population

$$\mathrm{E}[Y^a | X^* = x^*, S = 0]$$

under the hypothetical intervention $A = a$.

As before, $X^*$ is a subset of all covariates $X$ to allow for separation between covariates for prediction and those necessary for causal identification.

# (4) Identifiability conditions

A1* *Consistency*. If $A_i = a$, then $Y_i^a = Y_i$.

A2* *Conditional exchangeability in the trial*. $(Y^a \perp\!\!\!\perp A|X, S = 1)$.

A3* *Positivity in the trial*. For all $x$ with positive density, i.e. $f_{X|S=1}(x|S = 1) > 0$, we have $\Pr[A = a|X = x, S = 1] > 0$.

A4* *Conditional exchangeability of trial participation*. $Y^a \perp\!\!\!\perp S|X$.

A5* *Overlap of participation*. For all $x$ with positive density, i.e. $f_{X|S=0}(x|S = 0) > 0$, we have $\Pr[S = 1|X = x] > 0$.

# (4) Identifiability conditions

A1* *Consistency*. If $A_i = a$, then $Y_i^a = Y_i$.

A2* *Conditional exchangeability in the trial*. $(Y^a \perp\!\!\!\perp A|X, S = 1)$.

A3* *Positivity in the trial*. For all $x$ with positive density, i.e. $f_{X|S=1}(x|S = 1) > 0$, we have $\Pr[A = a|X = x, S = 1] > 0$.

A4* *Conditional exchangeability of trial participation*. $Y^a \perp\!\!\!\perp S|X$.

A5* *Overlap of participation*. For all $x$ with positive density, i.e. $f_{X|S=0}(x|S = 0) > 0$, we have $\Pr[S = 1|X = x] > 0$.

A1*-A3* are same as in observational setting, however A2* and A3* are assured by design in the trial. In transportability, in essence we exchange them for A4* and A5*.

# (4) Identifiability conditions

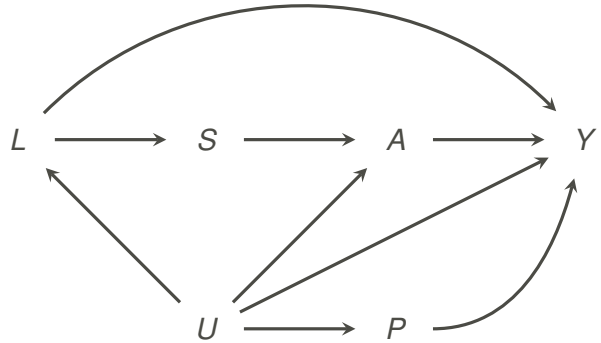



Figure: Causal directed acyclic graph showing law of observed data where identifiability conditions hold.

# (5) Estimation and inference

If assumptions A1* through A5* hold, then $g_a(X^*)$ tailored to the target population is identified by

$$g_a(X^*) = \mathrm{E}[\mathrm{E}[Y|X, S = 1, A = a]|X^*, S = 0] \tag{5}$$

# (5) Estimation and inference

If assumptions A1* through A5* hold, then $g_a(X^*)$ tailored to the target population is identified by

$$g_a(X^*) = \mathrm{E}[\mathrm{E}[Y|X, S = 1, A = a]|X^*, S = 0] \tag{5}$$

or the equivalent inverse probability weighting representation

$$g_a(X^*) = \frac{1}{Pr[S = 0]} \mathrm{E}\left[ \frac{Pr[S = 0|X] I(S = 1, A = a)}{Pr[S = 1|X] Pr[A = a|X, S = 1]} Y \middle| X^* \right]. \tag{6}$$

Similar estimation procedures as described above for the observational analysis can be used.

# (6) Performance evaluation

If assumptions A1* through A5* hold, then the counterfactual MSE in the target population can be written as

$$\psi_{tr}(a) = \mathrm{E}[\mathrm{E}[(Y - g_a(X^*))^2 | X, S = 1, A = a] | S = 0]. \quad (7)$$

# (6) Performance evaluation

If assumptions A1* through A5* hold, then the counterfactual MSE in the target population can be written as

$$\psi_{tr}(a) = \mathrm{E}[\mathrm{E}[(Y - g_a(X^*))^2 | X, S = 1, A = a] | S = 0]. \quad (7)$$

or the equivalent inverse probability weighting representation

$$\psi_{tr}(a) = \frac{1}{Pr[S = 0]} \mathrm{E}\left[ \frac{Pr[S = 0|X] I(S = 1, A = a)}{Pr[S = 1|X] Pr[A = a|X, S = 1]} (Y - g_a(X^*))^2 \middle| X^* \right]. \quad (8)$$

Similar estimation procedures as described above for the observational analysis can be used. A doubly-robust estimator is also obtainable.

# What if we have data from a trial and an observational study?

Observational study sample ($O_0$) :

| ID | X | S | A | Y |
|---|---|---|---|---|
| 1 | 15.2 | 0 | 0 | 15 |
| 2 | 0.5 | 0 | 0 | 10 |
| 3 | 4.7 | 0 | 1 | 25 |

+

Trial sample ($O_1$) :

| ID | X | S | A | Y |
|---|---|---|---|---|
| 4 | 2.3 | 1 | 1 | 10 |
| 5 | 14.2 | 1 | 0 | 20 |
| 6 | 8.9 | 1 | 1 | 30 |

# Benchmarking

We define **benchmarking** as comparing causal estimates from an observational analysis with those from a (prior) randomized trial. E.g., we could compare

$$g_{trial}(X^*) = \mathrm{E}[\mathrm{E}[Y|X, S = 1, A = a]|X^*, S = 0]$$

to

$$g_{obs}(X^*) = \mathrm{E}[\mathrm{E}[Y|X, S = 0, A = a]|X^*, S = 0].$$

In practice, benchmarking the result for each covariate pattern in $X^*$ may be infeasible in which case we may prefer $\mathrm{E}[\widehat{g}_{trial}(X^*)] \approx \mathrm{E}[\widehat{g}_{trial}(X^*)]$

# Benchmarking

We define **benchmarking** as comparing causal estimates from an observational analysis with those from a (prior) randomized trial. E.g., we could compare

$$g_{trial}(X^*) = \mathrm{E}[\mathrm{E}[Y|X, S = 1, A = a]|X^*, S = 0]$$

to

$$g_{obs}(X^*) = \mathrm{E}[\mathrm{E}[Y|X, S = 0, A = a]|X^*, S = 0].$$

Successful benchmarking, e.g. $\widehat{g}_{trial}(X^*) \approx \widehat{g}_{trial}(X^*)$, can increase trust that the assumptions underpinning observational analysis hold, but it does not guarantee validity[1].

---

In practice, benchmarking the result for each covariate pattern in $X^*$ may be infeasible in which case we may prefer $\mathrm{E}[\widehat{g}_{trial}(X^*)] \approx \mathrm{E}[\widehat{g}_{trial}(X^*)]$

# Joint analysis

If benchmarking is successful, then a natural question is: **can we simply combine them?**

# Joint analysis

If benchmarking is successful, then a natural question is: **can we simply combine them?**

We define **joint analysis** of the trial and observational data as

$$g_{joint}(X^*) = \mathrm{E}[\mathrm{E}[Y|X, A = a]|X^*, S = 0]$$

that is, fitting a model in the *combined data* from the randomized trial and observational study then standardizing to the covariate distribution in the target population.

# Joint analysis

If benchmarking is successful, then a natural question is: **can we simply combine them?**

We define **joint analysis** of the trial and observational data as

$$g_{joint}(X^*) = \mathrm{E}[\mathrm{E}[Y|X, A = a]|X^*, S = 0]$$

that is, fitting a model in the *combined data* from the randomized trial and observational study then standardizing to the covariate distribution in the target population.

If assumptions hold, joint analysis will always be more efficient!

# Empirical example

CASS = Coronary Artery Surgery Study

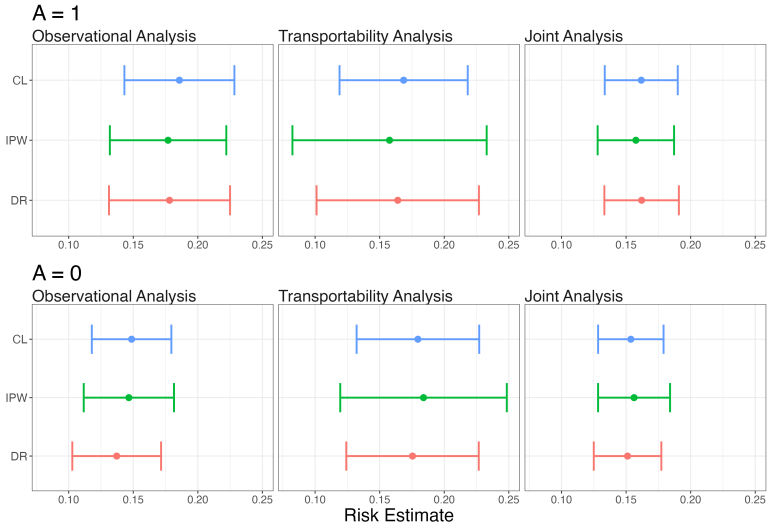
Participants ($N = 1,686$) could select to be a part of a randomized trial ($S = 1$) and if they declined they were offered to participate in an observational study ($S = 0$).

Outcome: 10-year cumulative risk of mortality

Intervention(s): Coronary artery bypass grafting surgery plus medical therapy (hereafter surgery, $A = 1$) versus only medical therapy ($A = 0$).

Prediction model: random forest fit using 50% of data (training) and evaluted in hold out (test).

# Empirical example




Cleveland Clinic

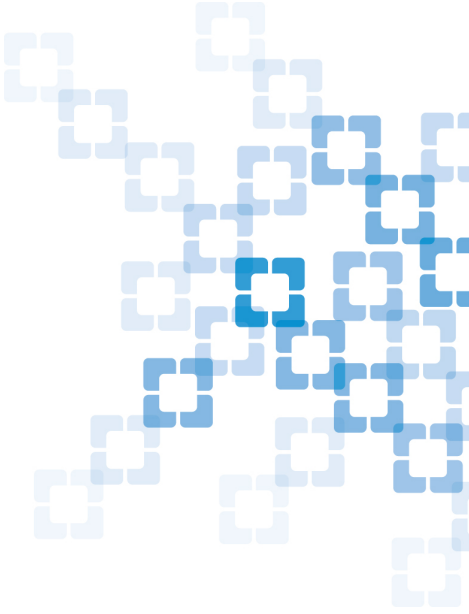
# References


1. Boyer, C., Dahabreh, I. J., & Steingrimsson, J. A. (2025). “Estimating and evaluating counterfactual prediction models”. *Statistics in Medicine* (In Press). Preprint: https://doi.org/10.48550/arXiv.2308.13026

2. Voter, S., Dahabreh, I. J., Boyer, C., Rahbar, H., Kontos, D., & Steingrimsson, J. A. (2025). “Counterfactual prediction from machine learning models: transportability and joint analysis for model development and evaluation using multi-source data”. *BMC Diagnostic and Prognostic Research* (In Press).

# Thank you! Questions?

Contact me:
boyerc5@ccf.org
@boyercb



Cleveland Clinic