New approaches to factual and counterfactual risk prediction for cardiovascular disease

Dissertation Defense

Christopher Boyer Advisor: Goodarz Danaei

Department of Epidemiology

Harvard School of Public Health

January 6, 2023

Outline

- Introduction
- Paper 1: *"Factual and counterfactual prediction using the parametric g–formula"*
- Paper 2: *"Target trials for prediction: emulating a trial to estimate the treatment–naive risk"*
- Paper 3: *"Assessing performance of counterfactual predictions"*
- Future directions

Introduction

Background

The goal of a clinical prediction model is to quantitatively assess an individual's prognosis for a certain health outcome given their history, often with the purpose of aiding clinical decision–making.

Prognosis may be operationalized as the conditional probability of disease (Y) given a set of "risk factors" or "predictors" (X), i.e.

$$\underbrace{Pr(Y = 1 \mid X = x)}_{\text{conditional risk of event Y given that X = x}}$$

We generally further assume this probability can be approximated by a simple function of X, e.g. a model $\mu_\beta(X)$, which maps predictors to risks and can be learned from historic data about clinical outcomes.

$$Pr(Y = 1 \mid X = x) = \underbrace{\mu_\beta(x)}_{\text{model for x}}$$

Background

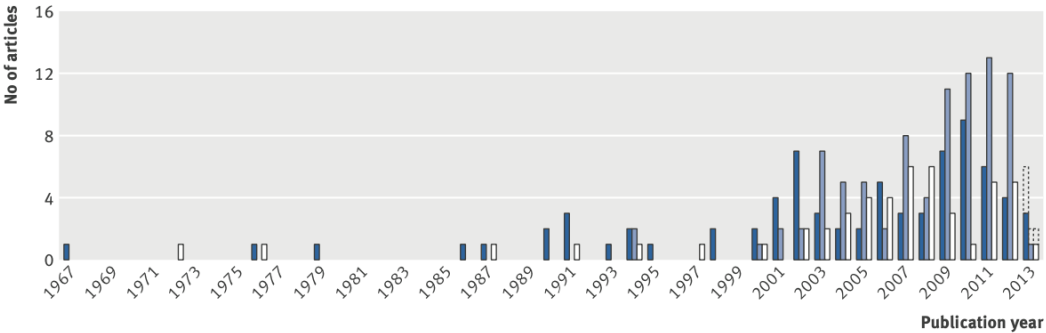


Fig 2 | Numbers of articles in which only one or more models were developed (dark blue), only one or more models were externally validated (light blue), or one or more models were developed and externally validated (white), ordered by publication year (up to June 2013). Predictions of the total numbers in 2013 are displayed with dotted lines

Source: Damen et al. 2016

To explain or predict?

Several famous papers have posited a divide between "two cultures" of statistical modeling.

Prediction

- Forward–looking
- Data–driven
- X chosen for task and predictability
- Focused on variance and agnostic evaluation of model performance

Causal inference

- Retrospective
- Theory–driven
- X chosen for identifiability
- Focused on unbiasedness

For a time, the two camps largely existed in isolation from one another. However, it's increasingly appreciated that there are also many fruitful areas of overlap.

In particular, there is increasing recognition that many common prediction tasks have counterfactual elements that may be more appropriately tackled using causal methods (**counterfactual prediction**).

Set up

We observe i.i.d. longitudinal samples $\{O_i\}_{i=1}^n$ from n participants across K time points,

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

where

- X_k : vector of time–varying covariates.
- A_k : a binary indicator of treatment.
- C_k : a censoring indicator.
- D_k : a competing event indicator.
- Y_k : a survival outcome indicator.
- T : failure time.

and overbars denote past history such that $\overline{X}_k = (X_0, \ldots, X_k)$.

Note: X_k here includes possible time–varying confounders L_k as well as predictors of outcome P_k that are not confounders, i.e. $X_k = (L_k, P_k)$.

Factual risk predictions

predictions that target estimands involving strictly *observable* quantities, e.g.

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

mostly answer question: *"what is...?"*

Example:

- What is the 10–year risk of developing cardiovascular disease (Y_{K+1}) for a 60–year old man with type–II diabetes and a history of smoking ($X_0 = x_0$)?

Assumptions:

- Prediction observations $\{X_i^{obs}\}_{i=1}^n$ are independent samples from same "population" process $\mathcal{P}$ as the training data.
- No loss to follow up and no competing events such that outcomes are fully observed.

Counterfactual risk predictions

predictions that target estimands involving *counterfactuals*, e.g.

$$Pr(Y^{g}_{K+1} = 1 \mid X_0 = x_0)$$

mostly answer question: *"what if/what would...?"*

Example:

- What would my predicted 10–year risk of cardiovascular disease (Y^{g}_{K+1}) be if I started anti–hypertensive medication now ($g : \underline{A}_k = 1$) based on my current "risk factor" levels (X_0)?

Assumptions*:

- Exchangeability $Y^{g}_{K+1} \perp\!\!\!\perp A_k \mid X_0, \overline{A}_{k-1}, \ \forall k$
- Consistency $Y^{g}_{K+1} = Y_K$ for $\underline{A}_k = \underline{a}^{g}_{k}$
- Positivity $1 > f(a_k \mid x_0, \overline{a}_{k-1}) > 0, \ \forall k$

Note: many prediction targets of interest are counterfactual!

* Still require predictions to be drawn from same population and no loss to follow up and no competing events.

Objective of this dissertation

A humble attempt to dive into this gap between prediction and causal inference.

- **Paper 1** takes a method traditionally used for causal inference, the g–formula, and re–purposes as a means of flexibly targeting both factual and counterfactual predictions.
- **Paper 2** focuses on a specific real–world counterfactual prediction problem (predicting the treatment naive risk), clarifies the target trial it corresponds with, and proposes two approaches that preserve some advantages of separation.
- **Paper 3** asks, in spirit of prediction literature, how can we agnostically evaluate the counterfactual predictive performance of any model irrespective of whether it is correctly specified.

Paper 1: *"Factual and counterfactual prediction using the parametric g-formula"*

Background

- Clinical risk prediction models have proliferated and are particularly influential in cardiology.
  - More than 700 CVD risk models proposed.
  - Incorporated into treatment guidelines (such as those of AHA, NHS, and ECC).
  - Used to determine trial eligibility.
  - Used for population health planning.
- Despite the large number of models, most use a relatively narrow set of methods.
  - Data drawn from a (large) observational cohort or cohorts.
  - Predictors taken from a single baseline examination cycle.
  - Use standard parametric regression techniques to model the 10–year risk.

Background

Our idea:

What if we used the g–formula as a risk prediction model?

Potential Advantages:

1. Leverage multiple repeated measurements over time to generate *dynamic predictions* that "update" as more information is collected.
2. As a model of the joint distribution (i.e. time–varying risk factors and the outcome), generates predictions about the *natural course* that may be relevant.
3. May identify longitudinal *"risk trajectories"* that may improve the ability of the model to discriminate between cases and non–cases.
4. Generates both *factual* and *counterfactual predictions* of the sort that may be useful to patients, clinicians, and researchers.
5. Can also generate *population–level* risk predictions which may be useful for policy planning, counterfactual exploration, cost–effectiveness analysis etc.

Study aims

Our aims:

- Derive a modified version of the parametric g–formula that can be used as a risk prediction model.
- Evaluate it's ability to generate both factual and counterfactual predictions via simulation.
- Develop a real–world implementation using data from the Framingham Offspring study.

Theory: What is the g–formula?

The g–formula typically estimates the marginal distribution of outcome Y under a hypothetical time–varying intervention g, e.g.

$$Pr(Y^g_{K+1} = 1)$$

Under the assumptions of exchangeability, consistency, and positivity, the g–formula is identified[1] by:

$$Pr(Y^g_{K+1} = 1) = \sum_{\bar{l}_k} Pr(Y_{K+1} = 1 \mid \bar{A}_k = \bar{a}^g_k, \bar{L}_k = \bar{l}_k) \prod_{i=1}^{k} f(l_i \mid \bar{l}_{i-1}, \bar{a}^g_{i-1})$$

The g–formula can be thought of as an **extension of standardization** to the time–varying setting.

[1] For deterministic interventions

- The g–formula has mostly been used to estimate *population–level estimands*, e.g. Taubman et al. (2009).

- However, we can modify it to target *conditional estimands*[2] which are the goal in counterfactual prediction, i.e.

$$Pr(Y^g_{K+1} = 1 \mid \overline{X}_k = \overline{x}_k) = \sum_{\underline{l}_k} Pr(Y_{K+1} = 1 \mid \overline{X}_k = \overline{x}_k, \underline{A}_k = \underline{a}^g_k, \underline{L}_k = \underline{l}_k) \prod_{i=k}^{K} f(l_i \mid \bar{l}_{i-1}, \overline{a}^g_{i-1})$$

  where we're now summing over just the "future" from k to K.

- Equivalent to standardizing just over possible "futures".

[2] Here I'm assuming that covariates can be split into modifiable treatment A_k and other risk–factors sufficient to control for counfouding L_k, i.e. $X_k = (L_k, A_k)$

Theory: How do we get factual predictions?

- When we *set* treatment variables to specific values or distributions, the g–formula generates *counterfactual predictions*.

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

$$\sum_{\underline{l}_k} Pr(Y_{K+1} = 1 \mid \overline{X}_k = \overline{x}_k, \underline{A}_k = \underline{a}^{g}_k, \underline{L}_k = \underline{l}_k) \times \prod_{i=k}^{K} f(l_i \mid \overline{l}_{i-1}, \overline{a}^{g}_{i-1})$$

- When we allow all variables to follow their *natural course*, i.e. the values they would take on absent our intervention, the g–formula generates *factual predictions*.

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

$$\sum_{\underline{l}_k} \sum_{\underline{a}_k} Pr(Y_{K+1} = 1 \mid \overline{X}_k = \overline{x}_k, \underline{A}_k = \underline{a}_k, \underline{L}_k = \underline{l}_k) \prod_{i=k}^{K} f(l_i \mid \overline{l}_{i-1}, \overline{a}_{i-1}) f(a_i \mid \overline{l}_i, \overline{a}_{i-1})$$

- Conceptually, we can think of factual predictions as being equivalent to a specific counterfactual prediction, i.e. one under a *random dynamic regime* that is equivalent to the *natural course*.

$$f^{g}(a_k \mid \overline{l}_k, \overline{a}_{k-1}) = f^{obs}(a_k \mid \overline{l}_k, \overline{a}_{k-1})$$

- In most cases, we can't nonparametrically estimate the components of the g–formula due to the *curse of dimensionality.*
- Therefore, in practice we often use parametric models to estimate the components and monte carlo simulation to approximate the sum/integral.
- This is the *parametric g–formula*.

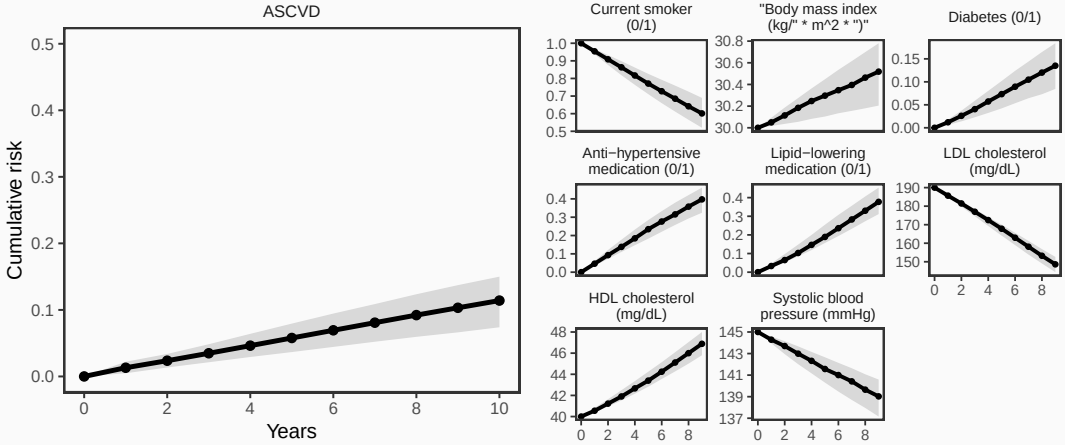
$$Pr(Y_K = 1 \mid \overline{X}_k = \overline{x}_k) =$$

$$\sum_{\underline{l}_k} \underbrace{Pr(Y_K = 1 \mid \overline{X}_k = \overline{x}_k, \underline{A}_k = \underline{a}_k, \underline{L}_k = \underline{l}_k)}_{\text{outcome model}} \times \prod_{i=k}^{K} \underbrace{f(l_i \mid \overline{l}_{i-1}, \overline{a}_{i-1})}_{\text{covariate models}} \times \underbrace{f(a_i \mid \overline{l}_i, \overline{a}_{i-1})}_{\text{treatment model}}$$

- For factual predictions, this involves models for the outcome as well as models for the natural course of the "risk factors" including treatments and covariates.

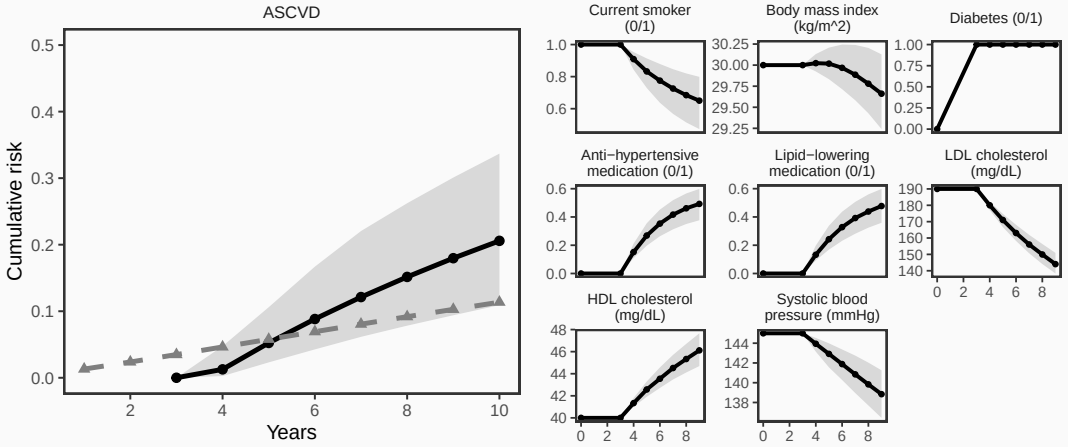
Example: Natural course

Profile: a 60 year–old male smoker with a BMI of 30 kg/m^2, no history of diabetes, no history of treatment and elevated risk factors levels.



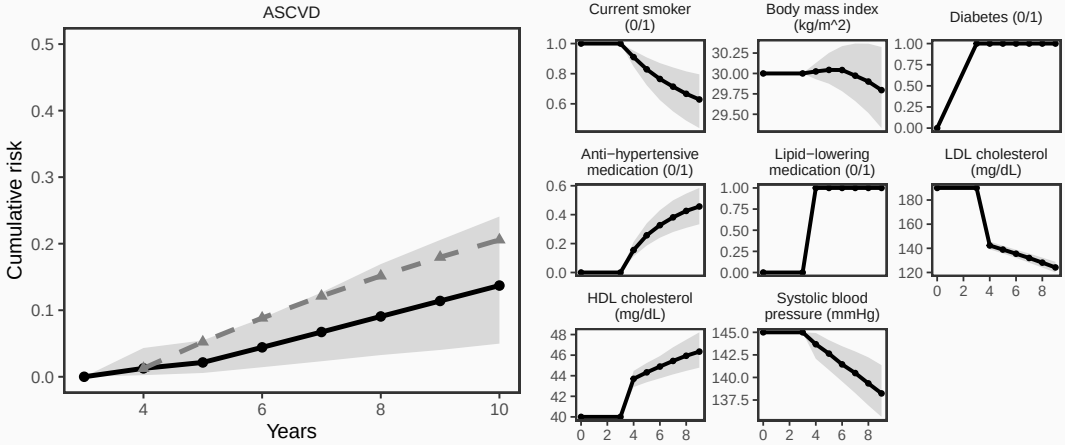
Example: Updated prediction 3 years later

Profile Updated after 3 years and a diabetes diagnosis. *Dotted line = predictions from previous panel.*



Example: Counterfactual prediction

Profile: What would his risk be if he started statins? *Dotted line = updated predictions from previous panel.*



Data generating process:

$K = 10$ time point process, five continous covariates ($L_0, L_{1_k}, L_{2_k}, P_0, P_{1_k}$) three of which ($L_{1_k}$, L_{2_k}, P_{1_k}) vary over time and two of which (L_0, P_0) are fixed at baseline, A_k is time–varying treatment, C_{k+1} is censoring indicator, D_{k+1} is competing event and Y_{k+1} is outcome. Feedback between L_{1_k}, L_{2_k}, and A_k.

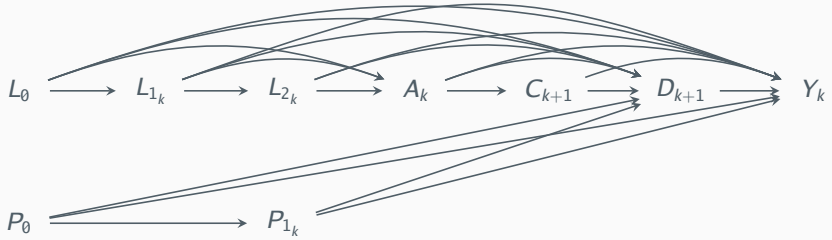


Figure: Template directed acyclic graph for simulation data generation process at the kth time point.

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

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

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

3. *Counterfactual prediction.* Training data are generated from process above but test data are generated from a population in which treatment is unnavailable, i.e. A_k are deterministically set to zero. Target is counterfactual cumulative risk $\Pr(Y_{10}^{\underline{a}_k=0} = 1 \mid \overline{X}_{k^*} = \overline{x}_{k^*})$ for $k^* \in \{0, 3, 6\}$.

Simulation

Comparators:

1. *G–formula.*
2. *Landmark regression without lags.*
3. *Landmark regression with lags.*

Model specification scenarios:

1. All models correctly specified.
2. Covariate models are misspecified.
3. Outcome models are misspecified.

Monte carlo simulation:

- Draw 500 samples of size 6000. Split into 3000 train and 3000 test.
- Estimate models in training.
- Predict target risk in test conditional on baseline ($k^* = 0$) only, first four values ($k^* = 3$) and first values ($k^* = 6$).
- Evaluate using dynamic MSE(Δk, k^*) and AUC(Δk, k^*)

Table: Monte carlo simulation results comparing g–formula and landmark approaches.

	MSE(Δk, k^*)			AUC(Δk, k^*)		
k^*	g–formula	landmark	landmark (lags)	g–formula	landmark	landmark (lags)
Scenario 1: Factual prediction						
0	**0.112**	0.116	0.116	**0.881**	0.880	0.880
	(0.006)	(0.007)	(0.007)	(0.011)	(0.011)	(0.011)
3	**0.099**	0.104	0.104	**0.903**	0.901	0.900
	(0.006)	(0.006)	(0.006)	(0.010)	(0.010)	(0.010)
6	**0.087**	0.091	0.091	**0.918**	0.916	0.916
	(0.006)	(0.006)	(0.006)	(0.010)	(0.010)	(0.010)

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

Table: Monte carlo simulation results comparing g–formula and landmark approaches.

	MSE(Δk, k^*)			AUC(Δk, k^*)		
k^*	g–formula	landmark	landmark (lags)	g–formula	landmark	landmark (lags)
Scenario 3: Counterfactual prediction						
0	**0.164**	0.388	0.388	**0.960**	0.943	0.943
	(0.009)	(0.018)	(0.018)	(0.006)	(0.007)	(0.007)
3	**0.116**	0.290	0.285	**0.970**	0.965	0.966
	(0.007)	(0.014)	(0.014)	(0.004)	(0.005)	(0.005)
6	**0.086**	0.158	0.158	**0.972**	0.971	0.971
	(0.006)	(0.012)	(0.015)	(0.004)	(0.004)	(0.004)

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

Application

Data source:

- Framingham Offspring Study: 5,135 offspring of the original Framingham Heart Study participants.
- We include 2,828 cohort members who were under 75 years of age at the start of the fifth examination cycle, had complete baseline data, were not on lipid–lowering treatment and had no prior cardiovascular disease event.

Outcome:

- Atherosclerotic cardiovascular disease defined as nonfatal MI, coronary heart disease death, or stroke.

Covariates:

- Age, sex, smoking status, BMI, diabetes status, systolic blood pressure, total cholesterol, HDL–cholesterol, statins and anti–hypertensive medication use.
- Chosen to be in line with those used in original Framingham risk equations/PCE.

Application

Missing data:

- Last observation carried forward for one cycle and then censor if the subject misses more than two consecutive exams.

Estimation strategy:

- Fit g–formula component models using:
  - pooled logistic regression models for ASCVD.
  - pooled logistic regression models for death.
  - covariate–specific pooled generalized linear regression models (see apendix).
- As a comparators fit landmark Cox models with and without lagged terms. Use Breslow estimator to get risks.

Application

Factual prediction:

- Target: 10–year risk of cardiovascular disease, $Pr(Y_{10} = 1 \mid \overline{X}_k^*)$, conditional on
  - baseline covariates only ($k^* = 0$)
  - the first 4 years of covariates ($k^* = 3$)
  - the first 7 years of covariates ($k^* = 6$)

- Performance: Calculate optimism corrected MSE(Δk, k^*) and AUC(Δk, k^*) using 500 bootstrap replicates.

Counterfactual prediction:

- Target: 10–year risk of cardiovascular disease **if never treated with statins**, $Pr(Y_{10}^{a=0} = 1 \mid \overline{X}_k^*)$, conditional on same set as above.

- Performance: Validate by calculating MSE(Δk, k^*) and AUC(Δk, k^*) **among ancestors in Framingham Heart Study**.

Application: Factual prediction

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

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

Application: Counterfactual prediction

Table: Validation of model performance for counterfactual prediction of 'treatment–naive' risk in Framingham Heart Study.

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

Summary

Conclusions

- Parametric g–formula can be used for time–dependent factual and counterfactual predictions, although the latter require untestable assumptions.
- When models are correctly specified (or approximately so), g–formula outperforms traditional approaches by efficiently using longitudinal data.
- Unlike traditional approaches, can flexibly target a range of counterfactual estimands using a single modeling framework.

Limitations

- Causal assumptions may be implausible in many prediction settings. Better data needed.
- G–null paradox, it may be impossible to correctly specify g–formula under the null for even a moderate number of time points. Supports using more flexible models.
- Unlike traditional approaches, can flexibly target a range of counterfactual estimands using a single modeling framework.

Paper 2: *"Target trials for prediction: emulating a trial to estimate the treatment–naive risk"*

Motivation

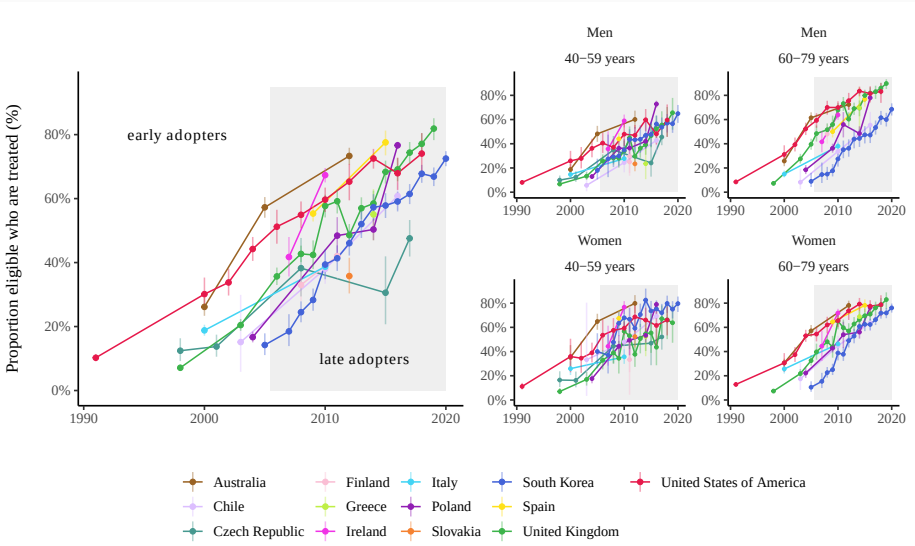
- A common use of prediction models in clinical care is to guide decisions about initiating treatment.
- This generally involves a model–based estimate of a patient's risk and then a decision rule to determine whether they should start treatment.
  - E.g. the AHA guidelines suggest initiating statins when 10–year risk of ASCVD exceeds 7.5% as determined by the pooled cohort equations.
- Ideally, treatment initiation would be based on the **treatment–naive risk**.
- However, in practice models are trained using observational data where treatment is initiated over follow up and therefore estimate the **natural course risk**.
- Could lead to underallocation of treatment (depending on rates of initiation and magnitude of effects).
- Existing approaches mostly either (a) ignore this issue or (b) censor at time of treatment initiation. Some authors are interested in using causal methods to estimate counterfactual risk of interest but have not approached it rigorously.

Aims

Aims

- Clarify the target trial corresponding to treatment–naive risk.
- Propose methods for estimating treatment–naive risk from observational data.
  - Focus on methods that fit within constraints of prediction task.
  - Focus on methods that allow algorithms traditionally preferred by modelers.
- Apply to real world example of statins in MESA.

Why statins?


Proportion eligible who are treated (%)
early adopters
late adopters
Men
40−59 years
60−79 years
Women
40−59 years
60−79 years
Australia
Chile
Czech Republic
Finland
Greece
Ireland
Italy
Poland
Slovakia
South Korea
Spain
United Kingdom
United States of America

Set up

We observe i.i.d. longitudinal samples $\{O_i\}_{i=1}^n$ from n participants across K time points,

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

where

- X_k : vector of time–varying covariates.
- A_k : a binary indicator of treatment.
- C_k : a censoring indicator.
- Y_k : a survival outcome indicator.
- T : failure time.

and overbars denote past history such that $\overline{X}_k = (X_0, \ldots, X_k)$.

Note: X_k here includes possible time–varying confounders L_k as well as predictors of outcome P_k that are not confounders, i.e. $X_k = (L_k, P_k)$.

Set up

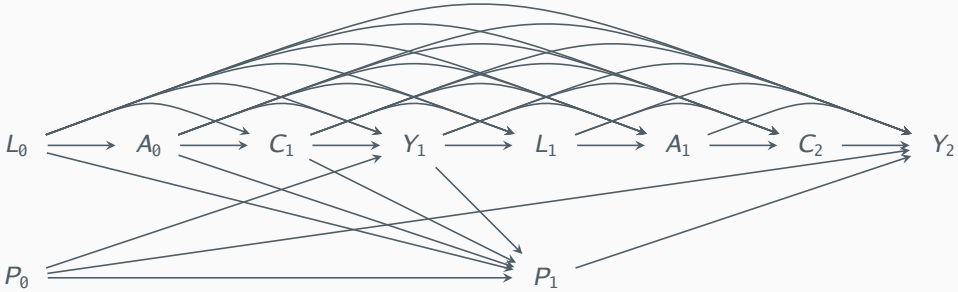


Figure: Example two time point directed acyclic graph for prediction.

Treatment–naive risk

The **treatment–naive risk** is defined as

$$\Pr[T^{\bar{a}=0} < t \mid X^* = x]$$

where predictors X^* are a subset of baseline covariates X_0, i.e. $X^* \subset X_0$, generally chosen for ease of collection and prognostic value.

Imagine we had access to $T^{\bar{a}=0}$, then this is just a standard prediction task for which there's a vast toolbox to draw from. For instance, in cardiovascular risk prediction modeling we might fit Cox regression

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

and estimate cumulative incidence using an estimator of $\Lambda_0(t)$.

However, when trained on data with treatment initiation we instead we observe T under **natural course** of treatment.

Target trial

Ideally, we would estimate $\Pr[T^{\overline{a}=0} < t \mid X^* = x]$ in a single arm trial in which treatment was withheld from all participants over follow up.

However, this trial is not ethical nor feasible. Instead, we can emulate it using observational data. Doing so can also help us clarify several key decisions about how data should be set up and how our ultimate model should be used.

- Restrict to those actually eligible for screening using prediction model (e.g. treatment free at baseline, meet guideline criteria for screening).
- Have a well defined time zero and good definition of the treatment to be withheld.

From this perspective, treatment initiation is a form of time–varying non–adherence that must be "adjusted" for using g–methods.

Our approach

To estimate the appropriate risk we need causal methods. However, we'd like something fit–to–purpose and which allows us ultimately to still use standard prediction tools. We focus on survival methods as these are common in epidemiology.

Proposal:

1. *SNAFTM:* Assume structural model for treatment removal, g–estimate structural parameters, construct pseudo–outcomes and use them in place of standard outcome.

2. *IPCW:* Treat initiation of treatment as non–adherence, censor participants when they initiate, and then use inverse probability weighting to adjust for informative censoring in prediction algorithm.

We require the following identifiability conditions:

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

and either of

3a. *Positivity:* $\Pr[A_k = 0 \mid \overline{X}_k, \overline{A}_{k-1} = 0, T > k] > 0$
3b. *Known semi–parametric model:* $T^{\bar{a}=0}$ follows a SNAFTM.

for all $k = 0, \ldots, K$.

Summary of SNAFTM approach

1. Assume observed time (T) relates to time under no treatment ($T^{\overline{a}=0}$) via SNAFTM.

$$T^{\overline{a}=0} = \int_0^T \exp\{\psi A_t\} dt \quad \text{or} \quad T^{\overline{a}=0} = \int_0^T \exp\{\gamma(t, \overline{A}_t, \overline{X}_t; \psi)\} dt$$

2. Estimate ψ using g–estimation under identifiability conditions.

3. Form pseudo–outcomes $H(\widehat{\psi})$ using g–estimates of ψ.

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

4. Use pseudo–outcomes to estimate $\Pr_n[H(\widehat{\psi}) < t \mid X^*]$ using any standard prediction algorithm.

Summary of IPCW approach

1. Censor participants when they deviate from the "never treat" regime.

2. Calculate stabilized weights W_c based on probability of remaining untreated, include baseline predictors X^* in numerator so they can be conditioned on in resulting model later.

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

3. Use any algorithm that permits time–varying weighted optimization to estimate treatment–naive risk, e.g. pooled logistic regression

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

Application

Multi–Ethnic Study of Atherosclerosis (MESA):

- A population–based sample of 6,814 residents 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.
- Conducted five examination visits between 2000 and 2011 in 18 to 24 month intervals
- 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.
- Importantly, the prevalence of statin treatment increased rapidly during the study period, both nationally and in this particular cohort.
- MESA was also one of the validation cohorts during the development of the pooled cohort equations.

Application

Our aims:

- Emulate a nested target trial for statin therapy in MESA and benchmark against published statin trials.
- If feasible, then emulate the single arm trial for withholding statins corresponding to the decision–point in the AHA guidelines.
- Account for time–varying non–adherence using the methods discussed and use emulation to build a model for the treatment–naive risk.
- Compare performance of treatment–naive model with models that do not account for treatment initiation.

Benchmarking against statin trials

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

Protocol component	Target trial specification	Emulation
Eligibility	Age 40 to 79 years No prior statin use No history of ASCVD LDL–C $\geq$ 70 mg/dL	same
Treatment strategies	(1) initiation of statins within 3 months of baseline randomization (2) no initiation of statins over follow up	same
Treatment assignment	non–blinded random assignment to either (1) or (2) at baseline	same but randomization is emulated conditional on covariates necessary to control confounding
Outcomes	5 and 10–year cumulative incidence of ASCVD defined as nonfatal myocardial infarction, coronary heart disease death, or ischemic stroke	same
Follow up	Start at baseline and follow until ASCVD event, non–ASCVD death, or until 10 years have elapsed, whichever happens first	same but exact starting time was estimated from time of questionnaire return
Statistical analysis	*ITT* – compare cumulative incidence of ASCVD under each strategy, adjusting for prognostic factors to increase efficiency *Per protocol* – Use IPW/g–estimation to account for time–varying non–adherence.	same but additionally emulating baseline randomization conditional on covariates

Benchmarking against statin trials

Table: Intention to treat and per protocol effects of statin therapy in nested trial emulation, Multi–Ethnic Study of Atherosclerosis, 2000 to 2010

	5–year[a]		10–year	
	HR	**95% CI**	**HR**	**95% CI**
ITT				
Pooled logit	0.79	(0.65, 0.93)	0.70	(0.56, 0.88)
g–estimation	0.77	(0.56, 0.98)	0.69	(0.47, 1.06)
Weibull κ	1.7		1.7	
Adherence–adjusted				
IPCW	0.68	(0.48, 0.94)	0.60	(0.45, 1.93)
g–estimation	0.66	(0.48, 0.94)	0.59	(0.45, 1.93)
Weibull κ	1.7		1.7	

HR = Hazard Ratio, CI = Confidence Interval

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

Benchmarking against statin trials

Summary:

- Estimated ITT effects in two–arm trial emulation in MESA compare favorably with those from randomized trials (e.g. HR = 0.79 vs HR = 0.75).
- Non–adherence common in both real trial and emulation.
- Adherence–adjusted analyses suggest that effects of always vs. never treat comparison may be stronger.
- Cumulative incidence curves suggest effects accumulate over time and hazards are nonconstant.
- While this is not dispositive, there was no evidence of gross deviations from trial results.

Conclusion: **We can proceed with building models for the treatment–naive risk in MESA.**

Single–arm trial emulation

- Follow up began at the 2nd examination cycle to enable a “wash out” period for statin use and to ensure adequate pre–treatment covariates to control confouding.
- Eligibility based on the implied decision point in the AHA guidelines.
  - Age 40 to 79 years.
  - No prior statin use.
  - No history of ASCVD.
  - LDL–C $\geq$ 70 mg/dL and LDL–C $\leq$ 190 mg/dL
- Participants followed for 10 years until first CVD event, death, or end of follow up.
- At each examination cycle, we used the corresponding questionnaire to determine non–adherence due to statin initiation.
- Because the exact timing of statin initiation was not known with precision, at each exam, we estimated the start of follow up for initiators and non–initators by drawing a random month between their current and previous examinations[3].

[3] We explored alternative definitions of the start of follow up in sensitivity analyses in the appendix.

Single–arm trial emulation

- Of the 6,814 MESA participants who completed the baseline examination, 4,149 met the eligibility criteria for our trial emulation. Over the follow up period there were
  - 288 ASCVD events
  - 190 non–ASCVD deaths
- Outcome is 10–year cumulative incidence of ASCVD defined as nonfatal myocardial infarction, coronary heart disease death, or ischemic stroke.
- After ten years approximately 40% of MESA participants had initiated statins.
- We impute missing values of time–varying covariates using multiple imputation based on previous exams.

Statin initiation and adherence

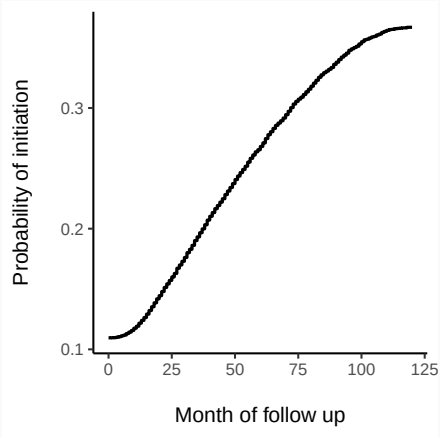


Figure: Probability of statin initiation and probability of adherence among initiators and non–initiators in nested target trial emulation, Multi–Ethnic Study of Atherosclerosis, 2000 to 2010.

Predicting treatment naive risk

- Predictors: age, sex, smoking status, diabetes history, systolic blood pressure, anti–hypertensive medication use and total and HDL serum cholesterol levels.

Models:

1. **Factual**: Cox PH model which ignores treatment initiation.
2. **Counterfactual (SNAFTM)**: Cox PH model using g–estimated pseudo–outcomes from SNAFTM.
3. **Counterfactual (IPCW)**: Pooled logistic regression with time–varying inverse probability weights.

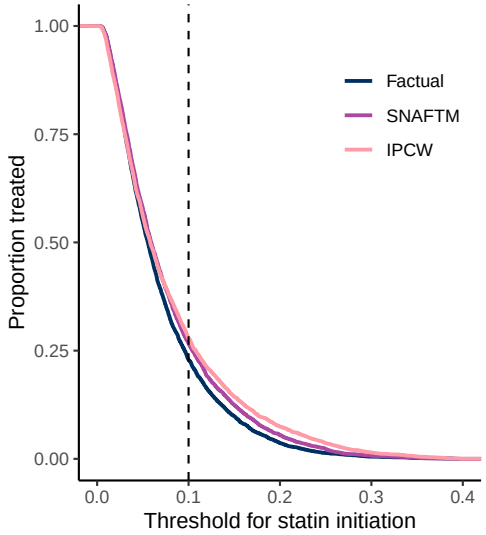
Predicting treatment naive risk

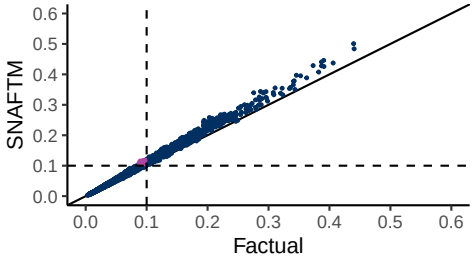
Table: Comparison of Cox proportional hazards models for predicting the factual and counterfactual statin–naive risk, Multi–Ethnic Study of Atherosclerosis, 2000 to 2010.

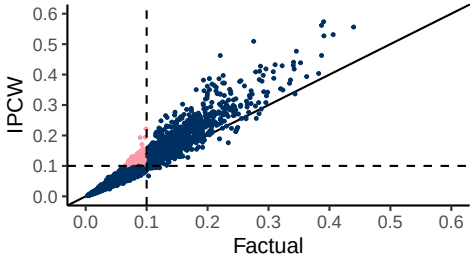
	Factual			Counterfactual (SNAFTM)			Counterfactual (IPCW)		
Characteristic	**HR[1]**	**95% CI[1]**	**p–value**	**HR[1]**	**95% CI[1]**	**p–value**	**HR[1]**	**95% CI[1]**	**p–value**
age	1.27	(1.18, 1.37)	<0.001	1.28	(1.19, 1.38)	<0.001	1.20	(1.11, 1.30)	<0.001
gender	1.64	(1.27, 2.13)	<0.001	1.66	(1.28,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

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

Predicting treatment naive risk







Summary

- Benchmarking suggests estimating effect of statins may at least be plausible in this population.
- Risk factor associations are stronger in treatment–naive models (makes sense as many are also indicators for treatment).
- Risks from treatment naive model are generally larger (makes sense as treatment is protective).
- In this population, using a factual model leads to an underallocation of statin therapy of between 5 to 9 percentage points at common decision thresholds of 7.5% and 10%.

Limitations

- Both methods require correctly specified models for treatment[4], g–estimation may be more robust to poor overlap, however also requires correct specification of effect modification by time–varying confounders.
- Under administrative censoring, g–estimation of SNAFTM must solve estimating equations that are nonsmooth and can sometimes fail to find a solution or produce multiple solutions.
  - Alternatives such as SNCFTM and SNCSTM, which under some paramterizations of blip function still estimate parameters of SNAFTM.
- We've taken for granted decisions about treatment should be informed by risk, rather than say estimated treatment effect.
  - Is it better to be using CATE here to begin with?
- Positivity assumption
- Not clear how to transparently assess counterfactual prediction performance.

[4] doubly robust estimating equations for g–estimation of SNAFTM is possible

Paper 3: *"Assessing the performance of counterfactual predictions"*

Motivation

In the spirit of the prediction literature, we'd like to be able to compare performance of counterfactual predictions agnostic to whether our model is "correct", e.g.

$$E[(Y^g - \mu_{\hat{\beta}}(X^*))^2]$$

Problem: We don't observe Y^g for all individuals

Why is this important?

- Prediction models are often deployed in settings that are different from those in which they are trained.
- For models with a significant prediction horizon, one of the ways settings may differ is that the natural course of treatment after baseline may vary.
- Even when models are deployed in the same setting, treatment policies may change over time leading to problems of "domain adaption" or "dataset shift".
- Beyond differences in training and deployment, there are instances in which the target prediction estimand is explicitly counterfactual.

Motivation

- We need methods for tailoring models to target counterfactual queries, even when outcome data may not be available.
- We also need performance metrics that agnostically evaluate model performance in these new environments independent of whether they are correctly specified.
- Literature on what to do is jumbled (e.g.subset to those who are untreated) and not causal.

Aims:

1. Clarify estimation of common counterfactual performance metrics when only observed training/test data are available.
2. Propose estimators and evaluate them in simulation.
3. Apply to real world example from paper 2.

Set up

We observe a simple random sample from a source population $\{(X_i, A_i, Y_i)\}_{i=1}^{n}$ in which the initiation of treatment follows it's natural course.

where

- X : vector of time–varying covariates.
- A : a binary indicator of treatment.
- Y : the outcome of interest.

Furthermore:

- To fix concepts, we assume for now A is a point treatment (weakend in appendix).
- The dataset is randomly split into a training set and a test set. Let D_{train} and D_{test} be indicators of the split.
- Let $\mu_\beta(X^*)$ denote a parametric model indexed by parameter β and $\mu_{\widehat{\beta}}(X^*)$ be the "fitted" model using parameter estimates $\widehat{\beta}$.

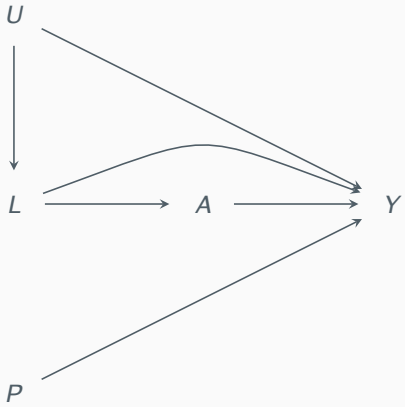


Figure: Example directed acyclic graph.

Performance target

- To determine the performance of the model, we generally relate its predictions $\mu_{\widehat{\beta}}(X^*)$ to the observed outcomes Y^a using any of a number of common metrics.
- However, for counterfactual predictions, this is not as simple as the potential outcome Y^a is not observed for all individuals.
- Yet, as we show, under certain conditions the metric may still be identified using only the observed data in the test set.
- An example target performance metric of interest is the mean–squared error (MSE)

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

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.

Under the following identification conditions

1. *Exchangeability* $Y^a \perp\!\!\!\perp A \mid X$
2. *Consistency* $Y^a = Y$ if $A = a$
3. *Positivity* $1 > \Pr(A = a \mid X) > 0$

The MSE under a counterfactual intervention which sets A to a is identified by

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

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

Using sample analogs for the identified expressions, we obtain two plug–in estimators for the counterfactual MSE

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

and

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

and a doubly–robust estimator

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

where $\widehat{h}_a(X)$ is an estimator for $E[(Y - \mu_{\widehat{\beta}}(X^*))^2 \mid X, A = a, D_{test} = 1]$ and $\widehat{e}_a(X)$ is an estimator for $\Pr(A = a \mid X, D_{test} = 1)$.

Simulation study 1

Goal: 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.

Data generation:

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

Model target: Want to know how it would perform in same population if no one were treated.

$$E[Y^{a=0} \mid X^*] = \mu_\beta(X^*)$$

Simulation study 1

Models (μ_β):

1. OLS correct – $Y \sim 1 + X + X^2$
2. OLS misspecified – $Y \sim 1 + X$
3. WLS correct – $Y \sim 1 + X + X^2$, weights from logistic $A \sim 1 + X$
4. WLS misspecified – $Y \sim 1 + X$, weights from logistic $A \sim 1 + X$

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

1. naive – naively estimate MSE without adjustment for treatment.

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

2. IPW – use $\widehat{\psi}_{IPW}$ to target treatment–naive MSE.

Simulation study 1

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.

Goal: illustrate

(iii) the importance of correct specification of the nuisance models when estimating counterfactual performance.

(iv) the properties of the doubly–robust estimator under misspecification of the nuisance models.

Data generation:

$$X \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma})$$

$$A \sim \text{Bernoulli}\{\text{expit}(-0.3 + 0.2 \sum_{i=1}^{3} X_{(i)} + 0.3 \sum_{i=1}^{3} X_{(i)}^2)\}$$

$$Y \sim \text{Bernoulli}\{\text{expit}(-0.3 + 0.2 \sum_{i=1}^{3} X_{(i)} + 0.3 \sum_{i=1}^{3} X_{(i)}^2 - 0.5A)\}$$

Simulation study 2

Model $\mu_\beta(X^*)$: logistic regression $Y \sim 1 + X_{(1)} + X_{(2)} + X_{(3)}$

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

1. naive – naively estimate MSE without adjustment for treatment (same as befor).
2. CL – use $\widehat{\psi}_{CL}$ to target treatment–naive MSE.
3. IPW – use $\widehat{\psi}_{IPW}$ to target treatment–naive MSE.
4. DR – use $\widehat{\psi}_{DR}$ to target treatment–naive MSE.

We consider both correctly and missspecified nuisance functions $e_a(X)$ and $h_a(X)$.

- $e_a(X)$ correct – logistic $A \sim 1 + \sum_{i=1}^{3} X_{(i)} + \sum_{i=1}^{3} X_{(i)}^2$
- $e_a(X)$ misspecified – logistic $A \sim 1 + X_{(1)} + X_{(2)} + X_{(3)}$
- $h_a(X)$ correct – logistic $Y \sim 1 + \sum_{i=1}^{3} X_{(i)} + \sum_{i=1}^{3} X_{(i)}^2$
- $h_a(X)$ misspecified – logistic $Y \sim 1 + X_{(1)} + X_{(2)} + X_{(3)}$
- $e_a(X)$ gam – generalized additive model $A \sim 1 + X_{(1)} + X_{(2)} + X_{(3)}$
- $h_a(X)$ gam – generalized additive model $Y \sim 1 + X_{(1)} + X_{(2)} + X_{(3)}$

Simulation study 2

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.

Application

- Use emulated statin–naive trial from MESA.
- Evaluate performance of a model for statin–naive risk by estimating counterfactual MSE.
- Changes:
  - For simplicity assume initiation is time–fixed instead of time–varying.
  - Compare simple logistic and IPCW logistic models instead of Cox models.
  - Split into training and test sets of equal size.
  - Use bootstrap to estimate uncertainty.

Table: 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.

Summary

- Many practical problems in prediction modeling involve counterfactuals.
- Here, we considered cases where predictions under hypothetical interventions were desired but only training data from observational sources were available.
- Common measures of model performance can be identified, but only under causal assumptions.
- Key insight: counterfactual performance can be evaluated independent of whether model is "correct".
- Here we used simple train/test split but same principle should apply to cross–validation or bootstrapping.
- In the paper, we extend this framework to other measures of model performance such as AUC and calibration as well as time–varying treatments.

Concluding remarks and future directions

Conclusions

- We shouldn't shy away from the fact that many things we want to predict are inherently counterfactual.
- Outside ideal conditions, doing so will often require untestable assumptions.
  - But by embracing this opens up all the work in causal inference that has been done to weaken them/propose alternatives fit to purpose (e.g. instrumental variables, proximal inference, etc.)
- Those of a causal persuasion entering the world of prediction can benefit from some of its ethos.
  - E.g. we can evaluate the performance of a counterfactual prediction independent of whether the underlying model is correct.
  - There may be situations where variance matters more than bias.
- There are also many prediction problems which could benefit from adopting a more causal perspective.

Appendix

Table: Monte carlo simulation results comparing g–formula and landmark approaches.

	MSE(Δk, k^*)			AUC(Δk, k^*)		
k^*	g–formula	landmark	landmark (lags)	g–formula	landmark	landmark (lags)
Scenario 2: Competing risk prediction						
0	**0.102**	0.104	0.104	**0.893**	0.892	0.892
	(0.006)	(0.007)	(0.007)	(0.014)	(0.014)	(0.014)
3	**0.097**	0.099	0.099	**0.915**	0.913	0.911
	(0.006)	(0.006)	(0.006)	(0.013)	(0.013)	(0.013)
6	**0.089**	0.090	0.091	**0.925**	0.923	0.921
	(0.006)	(0.006)	(0.006)	(0.012)	(0.012)	(0.012)

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

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

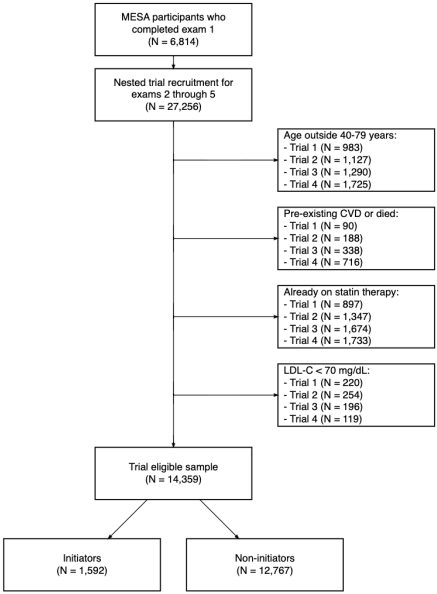
Intervention	Risk	95% CI	RR	95% CI	% intervened on
Framingham Offspring Cohort					
Never treat	7.6 %	(7.0%, 8.2%)	ref		13%
Natural course[1]	7.1%	(7.0%, 8.2%)	0.95	(7.0%, 8.2%)	0%
Always treat	6.0%	(7.0%, 8.2%)	0.79	(7.0%, 8.2%)	87%
Framingham Original Cohort					
Never treat[2]	11.2%	(7.0%, 8.2%)	ref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

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

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

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

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


MESA participants who completed exam 1 (N = 6,814)
Nested trial recruitment for exams 2 through 5 (N = 27,256)
Age outside 40-79 years:
- Trial 1 (N = 983)
- Trial 2 (N = 1,127)
- Trial 3 (N = 1,290)
- Trial 4 (N = 1,725)
Pre-existing CVD or died:
- Trial 1 (N = 90)
- Trial 2 (N = 188)
- Trial 3 (N = 338)
- Trial 4 (N = 716)
Already on statin therapy:
- Trial 1 (N = 897)
- Trial 2 (N = 1,347)
- Trial 3 (N = 1,674)
- Trial 4 (N = 1,733)
LDL-C < 70 mg/dL:
- Trial 1 (N = 220)
- Trial 2 (N = 254)
- Trial 3 (N = 196)
- Trial 4 (N = 119)
Trial eligible sample (N = 14,359)
Initiators (N = 1,592)
Non-initiators (N = 12,767)

Paper 2: Benchmarking analysis

ITT:

- Method 1: Pooled logistic regression conditional on specified baseline covariates in each nested trial and baseline assignment.
- Method 2: g–estimation of SNAFTM where non–adherence is ignored (baseline assignment only).

Adherence–adjusted:

- Method 1 (IPCW):
  - Weights estimated using pooled logistic regression for being censored.
  - Final HR calculated using weighted pooled logistic regression.
- Method 2 (SNAFTM):
  - g–estimation of SNAFTM with time–varying treatment.
  - Propensity scores estimated using pooled logistic regression.

Models condition on pre–treatment covariates from previous examination (e.g. for trial starting based on exam 2 we use covariates from exam 1). We account for competing risks by using subdistribution estimators and loss to follow up using inverse probability weighting.

Paper 2: Benchmarking sensitivity analysis

Table: Estimates of intention to treat effect of statin initation under different adjustment sets in emulated target trial for Benchmarking, Multi–Ethnic Study of Atherosclerosis, 2000 to 2010.

Model	HR	95% CI	P–value
unadjusted	1.03	(0.81, 1.31)	0.8
demographics[a]	0.90	(0.71, 1.14)	0.4
demographics[a] and risk factors[b]	0.71	(0.55, 0.91)	0.007
demographics[a], risk factors[b], and medications[c]	0.69	(0.53, 0.89)	0.004

CI = Confidence Interval; HR = Hazard Ratio

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

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

[c] Anti–hypertensive use, insulin use, daily aspirin use, anti–depressant use, vasodilator use, anti–arrvhtmic use

Paper 2: Benchmarking sensitivity analysis

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

Model	HR	95% CI	P–value
randomly selected mo.	0.69	(0.53, 0.89)	0.004
last exam + 1 mo.	0.62	(0.60, 0.65)	<0.001
current exam – 1 mo.	0.66	(0.64, 0.69)	<0.001

CI = Confidence Interval; HR = Hazard Ratio

Paper 2: Sensitivity analysis unmeasured confounding

- A limitation is that exchangeability is likely violated or only approximately true.
- In presence of unmeasured confounding, we can use modified version of sensitivity analysis suggested in Robins et al 2000.
- Suppose that the amount of unmeasured confounding were known in the sense that the degree of dependence between $T^{\overline{a}=0}$ and the conditional probability of treatment were known on the log–odds scale.
- Then we could solve for the parameter ω and function $q(k, \overline{X}_k, \overline{A}_{k-1}, T^{\overline{a}=0})$ in the logistic regression:

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

- Since ω and $q(k, \overline{X}_k, \overline{A}_{k-1}, T^{\overline{a}=0})$ are unknown, we could instead vary them over a plausible range of values and functional forms and examine the influence of unmeasured confounding on our resulting counterfactual prediction models.

Proofs