Nonparametric approaches to assessing variable importance using health data

Charlie Wolock
University of Pennsylvania

January 2025

Acknowledgement

The work I'm presenting today has been performed in collaboration with...

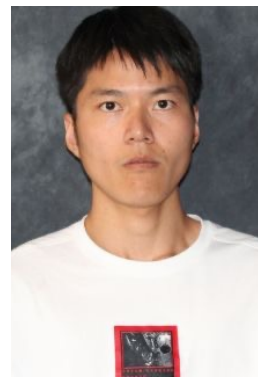

Peter Gilbert
Fred Hutch


Marco Carone
UW


Noah Simon
UW


Yong Chen
UPenn

Yang Ning
Cornell


Jian Yan
Cornell

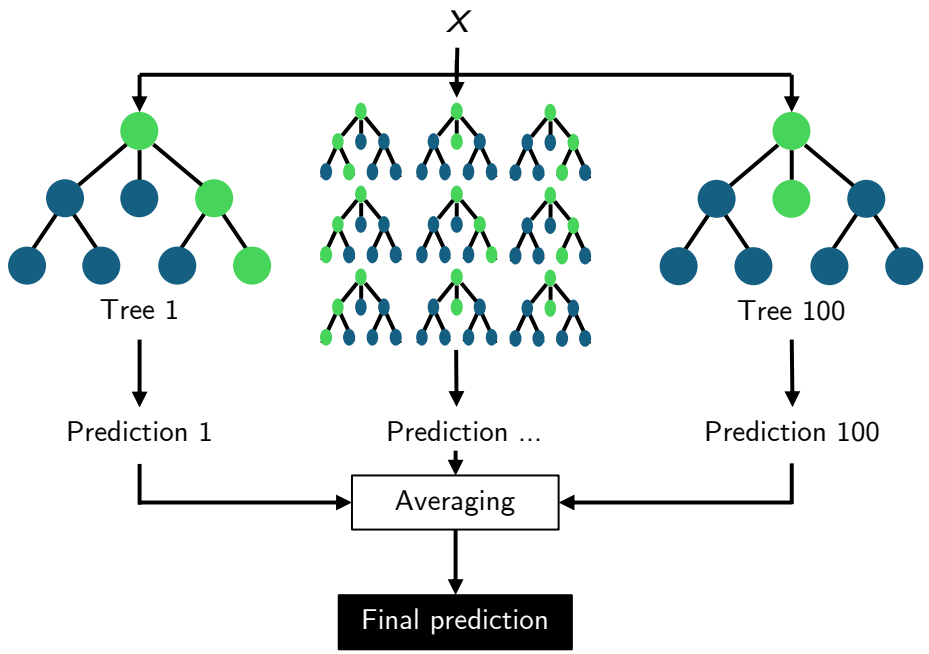

Yates Coley
KPWHRI


Brian Williamson
KPWHRI

Variable importance: what does it mean?

Variable importance is the quantification of the contribution of a feature (or group of features) toward a learning task.

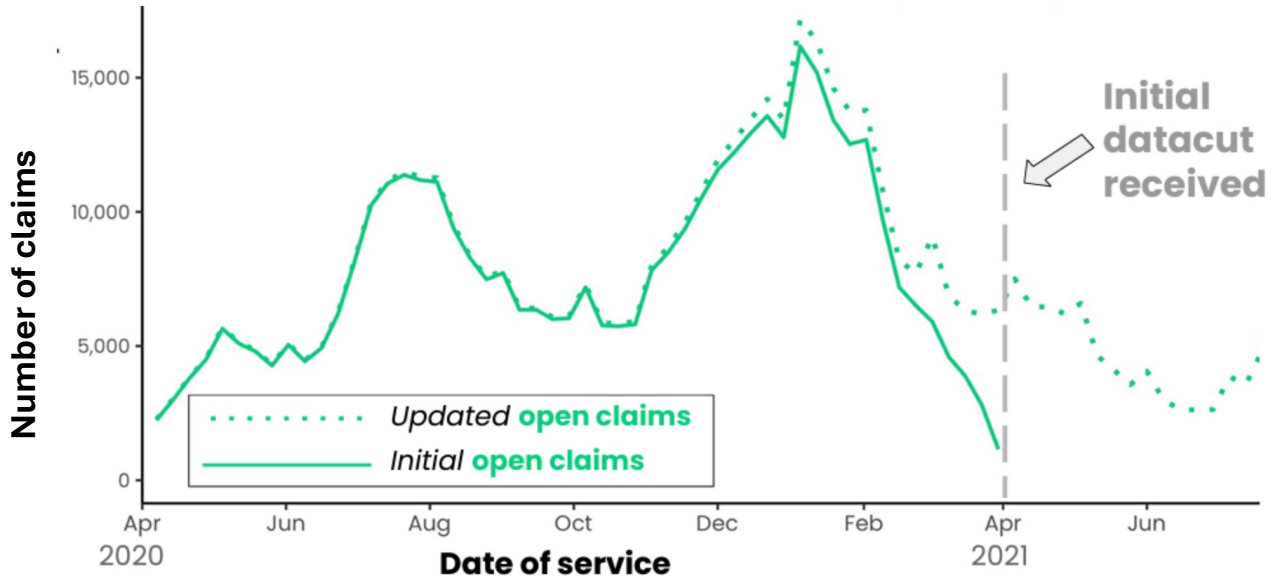
- Predicting an outcome
- Classifying an observation
- Determining a treatment rule

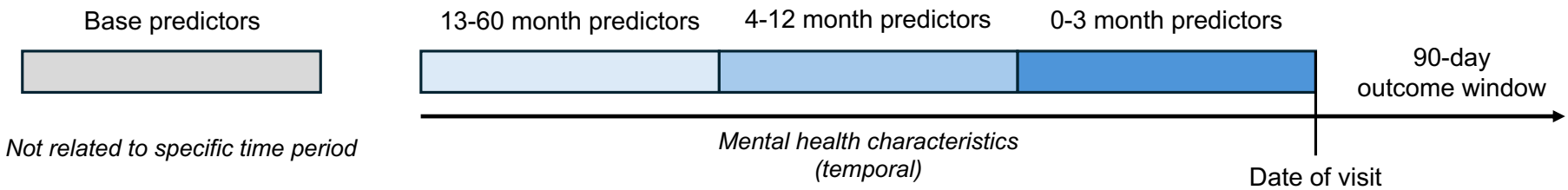


Example 1: suicide prevention

Motivation: Clinical risk prediction models for suicide and self-harm use data from EHR and insurance claims available at the time of a healthcare visit.

- Data reflecting recent events (prescriptions, diagnoses, self-harm events) may not be available in real time.



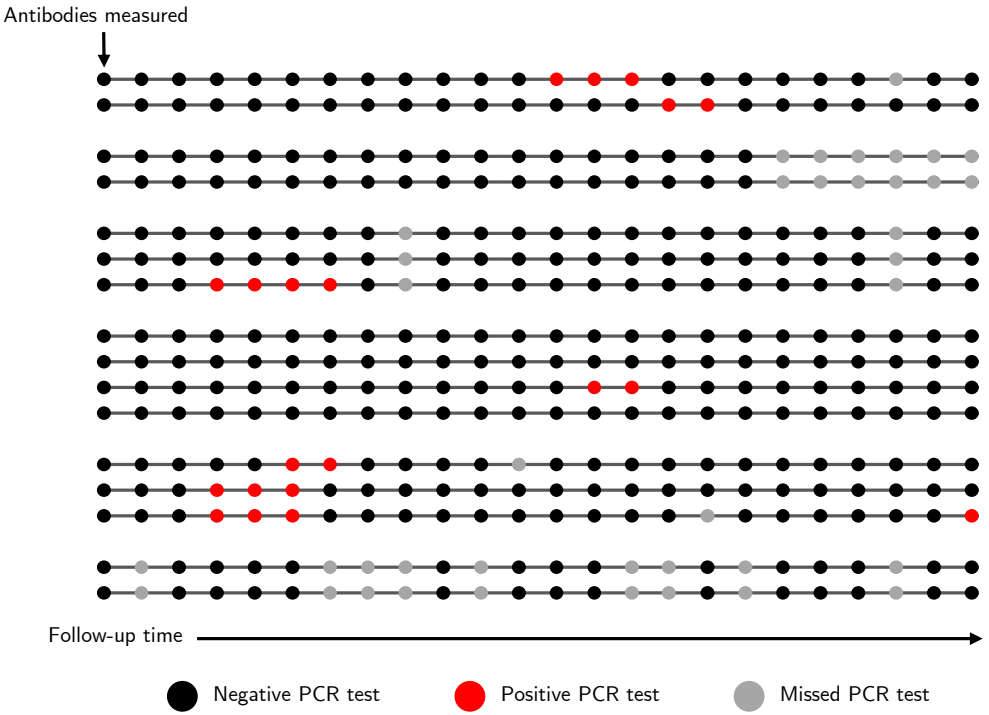


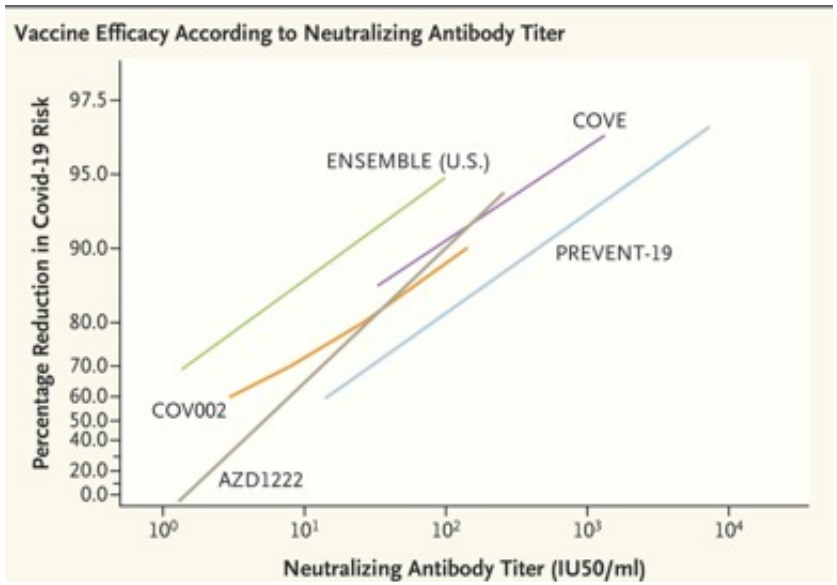
Weckstein et al. "Data lag in a large open and closed claims dataset: Navigating the completeness-timeliness tradeoff."
Wolock et al. "Importance of variables from different time frames for predicting self-harm using health system data."

Example 2: SARS-CoV-2 correlates of immunity

Motivation: Identify binding and neutralizing antibodies that predict protection against SARS-CoV-2 infection.

- Hypothesis generation for a formal correlates of immunity analysis

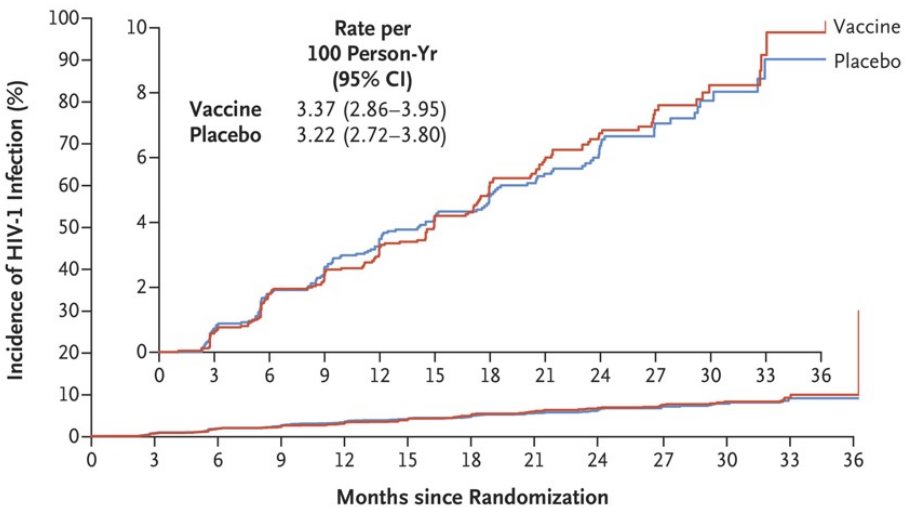




Gilbert et al. "A Covid-19 Milestone Attained — A Correlate of Protection for Vaccines."
Hoffman et al. "Correlates of Protection Against SARS-CoV-2 Infection in Children."

Example 3: HIV risk prediction

From 2016 to 2019, the HIV Vaccine Trials Network conducted a trial to investigate the efficacy of a recombinant canarypox vaccine targeted at HIV-1 subtype C (prevalent in sub-Saharan Africa) in adults aged 18-35.



Secondary objective: learn about factors that predict risk of HIV acquisition.

Risk models include:

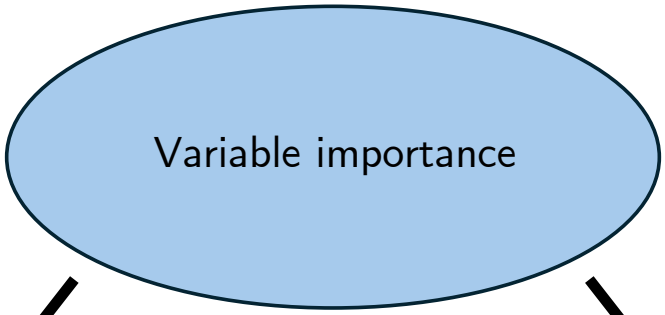
Age	Prevalent STIs
Sex	Behavioral characteristics
BMI	Partner characteristics
Housing	Geography

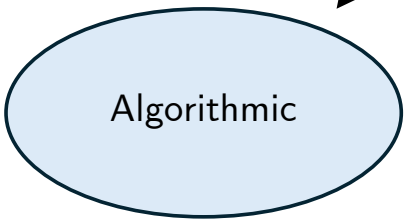
What is the relative importance of these features?

- In particular, how much do we gain from relatively "expensive" predictors?

Gray et al. "Vaccine efficacy of ALVAC–HIV and bivalent subtype C gp120–MF59 in adults."
Wolock, Gilbert, Simon, Carone. "Assessing variable importance in survival analysis using machine learning."

Types of variable importance

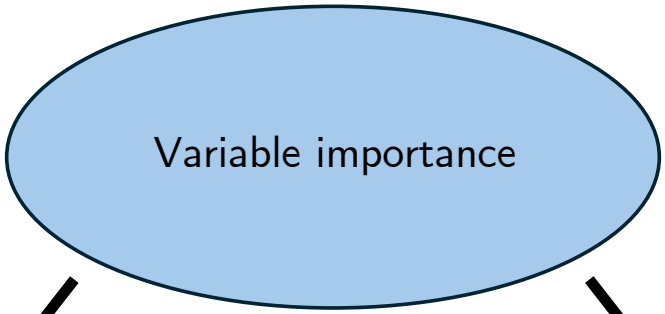




- Mean decrease in node impurity (random forests) (Breiman, 2001)
- Linear regression coefficient

- Permutation importance (Breiman, 2001; Fisher et al., 2019)
- Exclusion importance or "leave-one-covariate-out" (LOCO) (Lei et al., 2018)

Types of variable importance



Algorithmic

- Mean decrease in node impurity (random forests) (Breiman, 2001)
- Linear regression coefficient

Algorithm-agnostic

- Permutation importance (Breiman, 2001; Fisher et al., 2019)
- Exclusion importance or "leave-one-covariate-out" (LOCO) (Lei et al., 2018)

Predictiveness

The **ideal data unit** is $(X, T) \sim \mathbb{P}_0$, which lies in a nonparametric model $\mathcal{M}$.

- T is the outcome of interest.
- $X = (X_1, \ldots, X_p)$ are the features, predictors, or covariates.

We use f to denote a generic prediction function.

- The performance of f under sampling from distribution $\mathbb{P}$ is quantified by the predictiveness $\mathbb{V}(f, \mathbb{P})$, e.g., minus mean squared error.

$$\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \xrightarrow{f} f(X) \xrightarrow{\text{compare to } T} -\{f(X) - T\}^2 \xrightarrow{\text{summarize over } \mathbb{P}} -E_{\mathbb{P}}\left[\{f(X) - T\}^2\right]$$

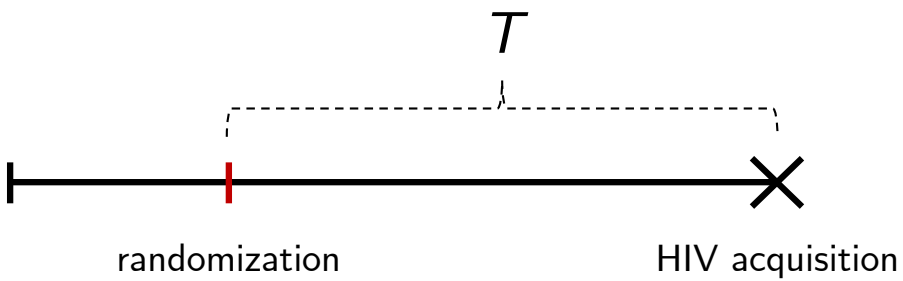
In this case, $\mathbb{V}(f, \mathbb{P}) = -E_{\mathbb{P}}\left[\{f(X) - T\}^2\right]$.

Note: Throughout this talk, 'blackboard' font represents the ideal data world, regular font the observed data world.

$$\underset{\text{ideal data}}{\mathbb{P}} \longleftrightarrow \underset{\text{obs. data}}{P}$$

Focusing on survival analysis

In each of the settings we consider, the outcome of interest T is the time between an initiating event and a terminating event.

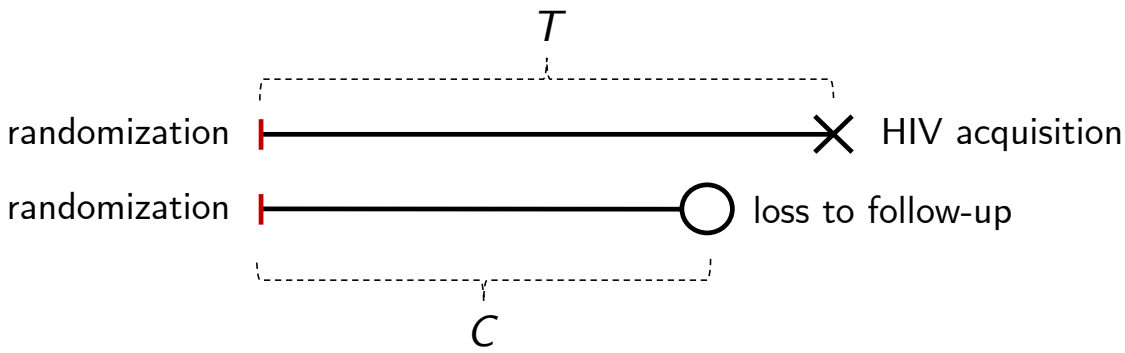


For now, we work on the scale of follow-up time, so the initiating event occurs at time $t = 0$ for all individuals.

Some common predictiveness measures in the survival setting:

- AUC at time τ: $\mathbb{V}(f, \mathbb{P}_0) = \mathbb{P}_0 \{f(X_1) > f(X_2) \mid T_1 \leq \tau, T_2 > \tau\}$
  (Heagerty and Zheng, 2005)
- Brier score at time τ: $\mathbb{V}(f, \mathbb{P}_0) = -E_{\mathbb{P}_0}[\{f(X) - \mathbb{1}(T \leq \tau)\}^2]$
  (Brier, 1950)
- C-index: $\mathbb{V}(f, \mathbb{P}_0) = \mathbb{P}_0 \{f(X_1) > f(X_2) \mid T_1 \leq T_2\}$
  (Harrell, 1982)

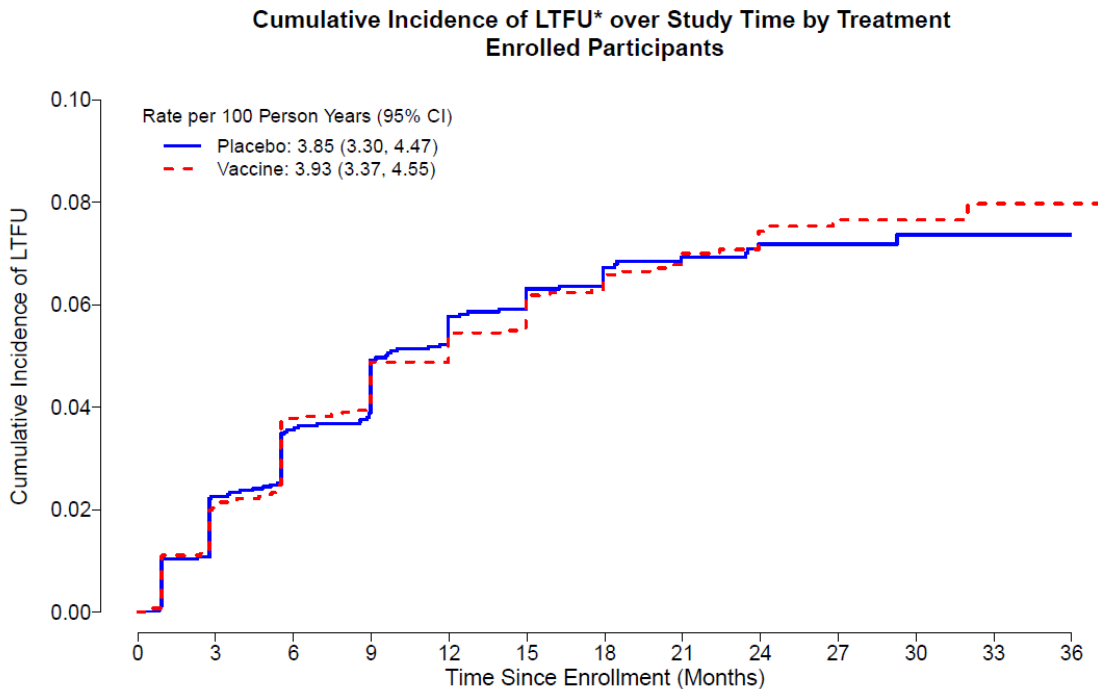
The key role of (informative) censoring



We observe $Y := \min(T, C)$ and $\Delta := \mathbb{1}(T \leq C)$.

Two types of censoring:

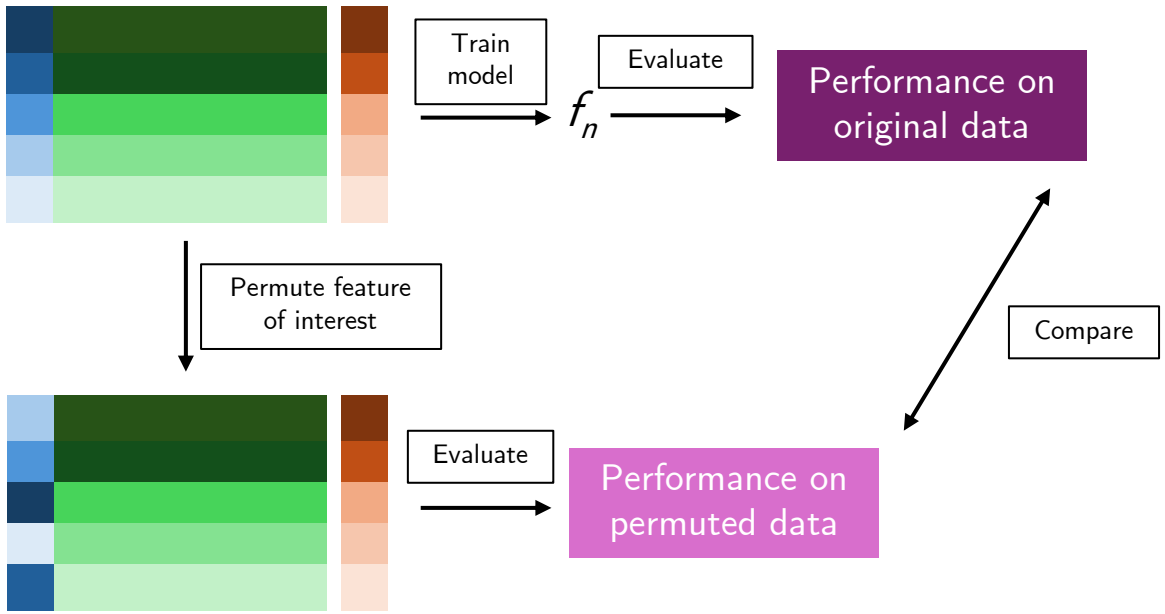
1. Study termination (administrative)
2. Participant drop-out



The current landscape

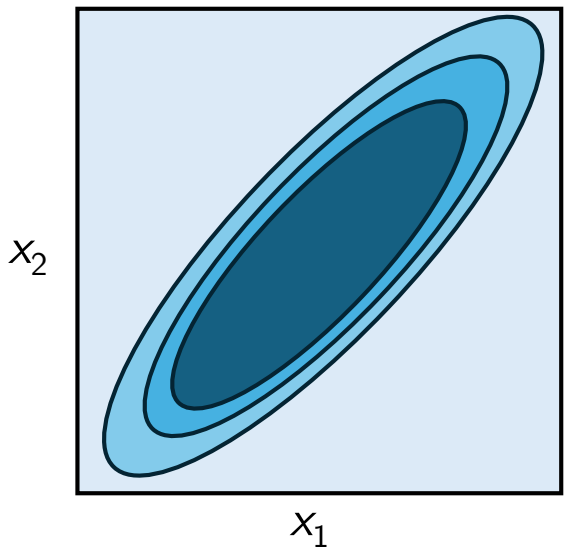
What methods exist for estimating variable importance under right censoring?

1. **Algorithmic:** Parametric/semiparametric models, e.g., coefficients from a Cox model (Cox, 1972).
   - Dependent on correct model specification
   - Difficult to compare across features
   - Not immediately clear how to handle interactions, groups of features, correlated features, etc.
   - Interpretation not necessarily linked to predictiveness
2. **Algorithm-agnostic:** Permutation (Breiman, 2001; Fisher et al., 2019).



Perils of permutation importance

- Lacking methods for estimation and inference under informative censoring.
- "Extrapolation bias": For correlated features, this approach requires generating predictions in regions with little (or even zero) probability mass under the joint distribution of the features (Hooker et al., 2021; Wang et al., 2024; Verdinelli and Wasserman, 2024).



Exclusion importance — comparing models trained using different subsets of features — does not suffer this same issue.

The oracle prediction function

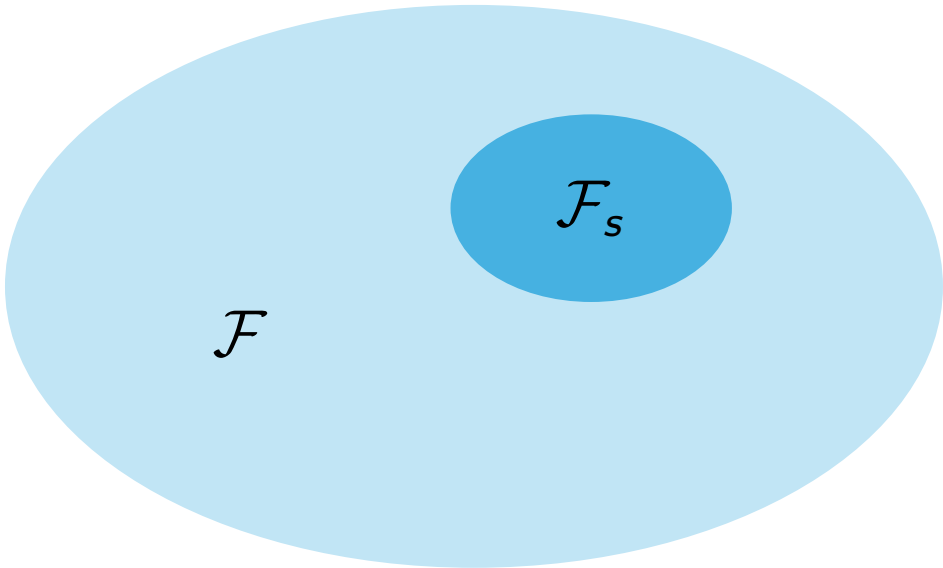
Goal: assess the importance of X_s, where $s \subset \{1, \ldots, p\}$, relative to the full predictor vector X. In the exclusion paradigm, we consider two special prediction functions:

$f_0 := \text{argmax}_{f \in \mathcal{F}} \mathbb{V}(f, \mathbb{P}_0)$

- The 'full' oracle, optimal in an unrestricted class of prediction functions

$f_{0,s} := \text{argmax}_{f \in \mathcal{F}_s} \mathbb{V}(f, \mathbb{P}_0)$

- The 'reduced' oracle, optimal in the class of prediction functions that does not use X_s



We define the **importance of X_s relative to X** as $\mathbb{V}(f_0, \mathbb{P}_0) - \mathbb{V}(f_{0,s}, \mathbb{P}_0)$.

- Rather than this subtractive notion, we could also consider an additive approach where X_s is added to a 'base' set of predictors.
- It may also be of interest to normalize by $\mathbb{V}(f_0, \mathbb{P}_0)$; estimation and inference can be handled using the delta method.

Roadmap

The parameter $\mathbb{V}(\mathbb{f}_0, \mathbb{P}_0) - \mathbb{V}(\mathbb{f}_{0,s}, \mathbb{P}_0)$ is a sensible quantification of the importance of X_s relative to X.

Taking this as our parameter of interest, we next focus on

1. identification;
2. estimation;
3. inference.

Identification

Due to right censoring, the predictiveness measure $\mathbb{V}$ is not a functional of the observed data distribution.

Many predictiveness measures have a common form that we can exploit:

$$\begin{aligned}\mathbb{V}(\mathbb{f}_0, \mathbb{P}_0) &= E_{\mathbb{P}_0}\left(\omega\left[\{\mathbb{f}_0(X_1), T_1\}, \ldots, \{\mathbb{f}_0(X_m), T_m\}\right]\right) \\ &= \int \cdots \int \omega\left[\{\mathbb{f}_0(x_1), t_1\}, \ldots, \{\mathbb{f}_0(x_m), t_m\}\right] \prod_{j=1}^{m} \mathbb{H}_0(dx_j, dt_j).\end{aligned}$$

where ω is a known kernel function and $\mathbb{H}_0$ is the joint cdf of (X, T) under $\mathbb{P}_0$.

With a slight abuse of notation, we write $\mathbb{V}(\mathbb{f}_0, \mathbb{P}_0)$ as $\mathbb{V}(\mathbb{f}_0, \mathbb{H}_0)$.

Identification

If T and C are independent **within strata defined by** X, then $\mathbb{H}_0$ is identified by the observed data distribution.

- In contrast, many existing methods for evaluating predictiveness with censored data make a stronger marginal independence assumption, simplifying estimation.

Under conditionally independent censoring, the joint cdf $\mathbb{H}_0$ can be identified pointwise — for some values of (x_0, t_0) — as

$$\mathbb{H}_0(x_0, t_0) = \int \mathbb{1}(x \leq x_0) G_0(t_0 \mid x) Q_0(dx) := H_0(x_0, t_0) \ ,$$

with

- G_0 the product-integral mapping applied to the conditional cumulative hazard function of T given X under $\mathbb{P}_0$;
- Q_0 the marginal cdf of X under $\mathbb{P}_0$.

Plug-in estimation

The identified parameter $V(f_0, H_0) - V(f_{0,s}, H_0)$ depends on the nuisance functions $(f_0, f_{0,s}, G_0)$.

- We want to use flexible (machine learning) methods, e.g., random survival forests (Ishwaran et al., 2008), survival Super Learner (Westling et al., 2024), survival stacking (Wolock et el., 2024).

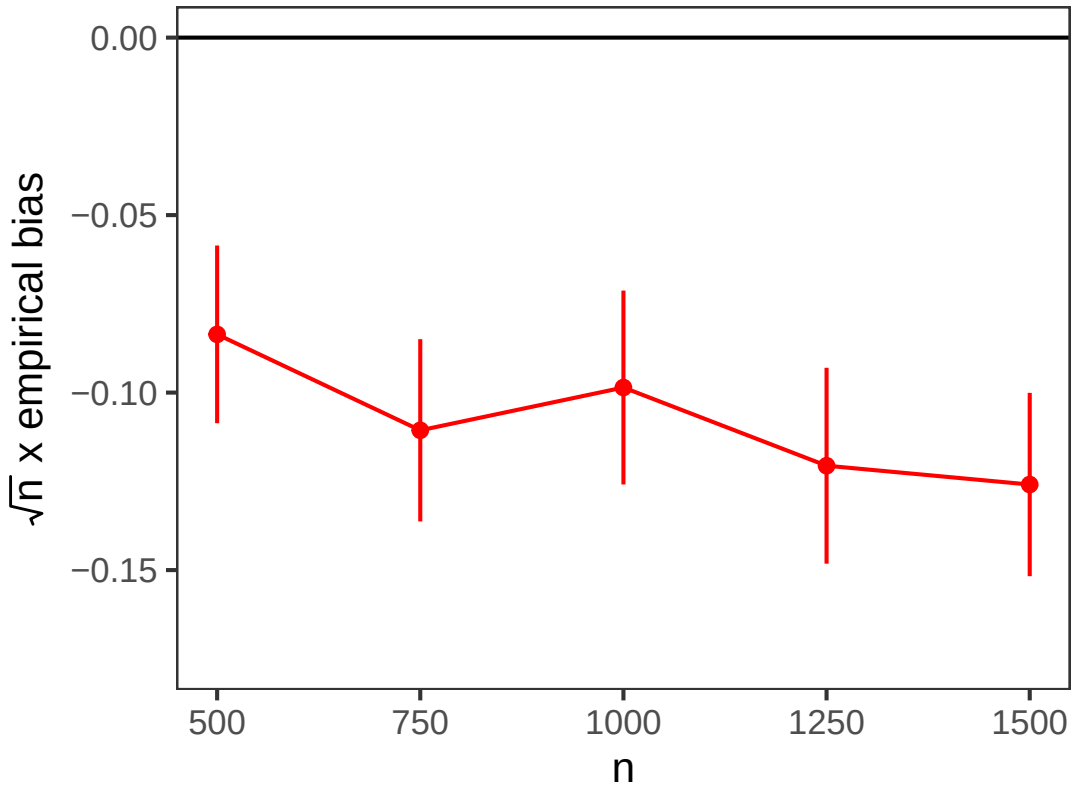
A first attempt:

1. Estimate the nuisance functions $(f_0, f_{0,s}, G_0)$ using machine learning estimators $(f_n, f_{n,s}, G_n)$.
2. Estimate Q_0 using the empirical distribution Q_n.
3. Set $H_n := (G_n, Q_n)$, and plug in estimated components:

$$V(f_n, H_n) - V(f_{n,s}, H_n) \ .$$

Plug-in estimation

The bias of the plug-in estimator tends to zero at a rate slower than $n^{-\frac{1}{2}}$.



This is due to the fact that H_n is not targeted to the parameter of interest. Interestingly, the estimation of f_0 and $f_{0,s}$ does not contribute to the excess first-order bias (Williamson et al., 2023).

Debiased estimation

How can we recover $n^{-\frac{1}{2}}$ asymptotics while using flexible nuisance estimators?

1. One-step estimator (Pfanzagl, 1982)
2. TMLE (van der Laan and Rubin, 2006)

Consider a parameter mapping $P \mapsto \Psi(P)$ that is pathwise differentiable (that is, smooth) with gradient ϕ. For any estimator $\hat{P}_n$ of P_0, a first-order expansion, similar to a functional Taylor expansion, gives

$$\Psi(\hat{P}_n) - \Psi(P_0) = \mathbb{P}_n\phi(P_0) - \mathbb{P}_n\phi(\hat{P}_n) + R(\hat{P}_n, P_0) + (\mathbb{P}_n - P_0)\{\phi(\hat{P}_n) - \phi(P_0)\}$$

- : Linear term, determines first-order behavior
- : Bias term
- : Second-order remainder term
- : Empirical process term

Under some conditions, we can expect the one-step estimator $\Psi(\hat{P}_n) + \mathbb{P}_n\phi(\hat{P}_n)$ to behave approximately like $\mathbb{P}_n\phi(P_0)$.

One-step estimator(s)

In general, there are multiple possible one-step estimators:

- 'direct' debias: using the gradient of $P \mapsto V(f_0, H_P)$
- 'indirect' debias: construct targeted estimator H_n^* of H_0 using the gradient of $P \mapsto H_P$, then construct estimator $V(f_n, H_n^*)$

The gradient of $P \mapsto H_P$, evaluated at the point (x_0, t_0), is given by

$$(x, y, \delta) \mapsto$$

$$\mathbb{1}(x \leq x_0) \left[G_0(t_0 \mid x) + S_0(t_0 \mid x) \left\{ \frac{\delta \mathbb{1}_{[0,t]}(y)}{S_0(y \mid x) R_0(y \mid x)} - \int_0^{t \wedge y} \frac{L_0(du \mid x)}{S_0(u \mid x) R_0(u \mid x)} \right\} \right]$$

(G_0, S_0, L_0): cdf, survival function, cumulative hazard of T given X.

(R_0): survival function of C given X.

$\psi_{n,s}$: direct one-step $\qquad$ $\psi_{n,s}^*$: indirect one-step

The oracle prediction functions

We have thus far avoided the question of how to estimate the oracle prediction functions f_0 and $f_{0,s}$.

- Derivation of the form of $(f_0, f_{0,s})$ must be handled on a case-by-case basis.
- For commonly used choices of $\mathbb{V}$, doubly-robust estimation is possible — consistent estimation of $(f_0, f_{0,s})$ is achieved by consistent estimation of **either** G_0 **or** R_0.

AUC, Brier score, MSE for τ–restricted survival time, ...

- The oracle prediction function can be written as $E_{\mathbb{P}_0}\{h(T) \mid X = x\}$ for a function h
- Use the doubly-robust pseudo-outcome regression approach of Rubin and van der Laan (2007)

C-index

- Oracle prediction function not available in closed form
- Estimation through direct optimization of $f \mapsto V(f, H_n^*)$.

Large-sample behavior

Result:

1. When all nuisances are estimated well, $\psi_{n,s}$ and $\psi^*_{n,s}$ have identical first-order asymptotics.
   - Letting ϕ_0 denote the efficient influence curve of the variable importance parameter, we have
   $$\psi_{n,s} = \psi^*_{n,s} + o_P(n^{-\frac{1}{2}}) = \frac{1}{n}\sum_{i=1}^{n} \phi_0(X_i, Y_i, \Delta_i) + o_P(n^{-\frac{1}{2}}) \ .$$
   - Therefore, $n^{\frac{1}{2}}(\psi^*_{n,s} - \psi_{0,s}) \rightsquigarrow N(0, \sigma^2_{0,s})$ with $\sigma^2_{0,s} := \text{var}_{P_0}\{\phi_0(X, Y, \Delta)\}$.

2. As long as either G_0 or R_0 is estimated consistently, $\psi^*_{n,s}$ remains consistent, while $\psi_{n,s}$ may fail to be.

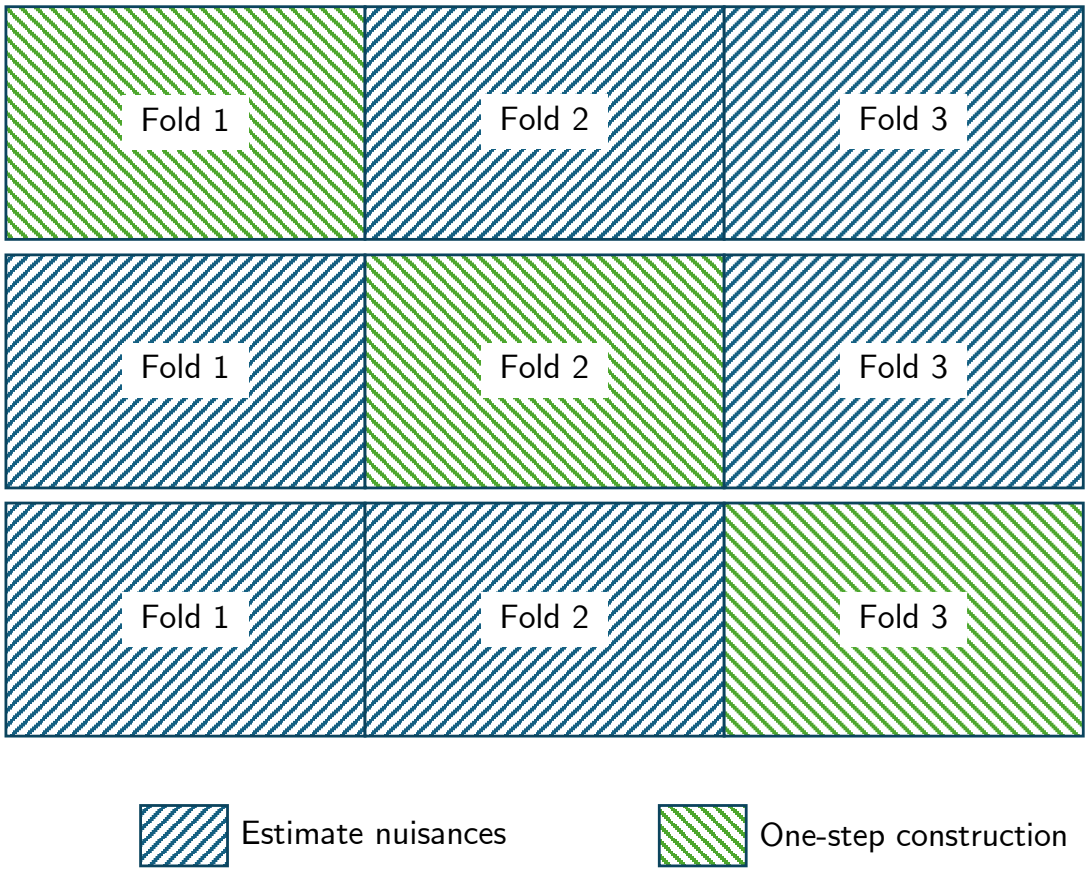
$$\psi^*_{n,s} \xrightarrow{p} \psi_{0,s} \qquad \psi_{n,s} \overset{p}{\not\to} \psi_{0,s}$$

   - Our proposed procedure enjoys doubly-robust **consistency**.
   - Doubly-robust **inference** (confidence intervals and p-values) in this setting remains an open question (Benkeser et al., 2017).

Cross-fitting

A standard regularity condition for asymptotic linearity is that $(f_n, f_{n,s}, G_n, R_n)$ are not too complex. This is called a Donsker condition.

Cross-fitting can help us avoid this condition (Zheng and van der Laan, 2011; Chernozhukov et al., 2018).

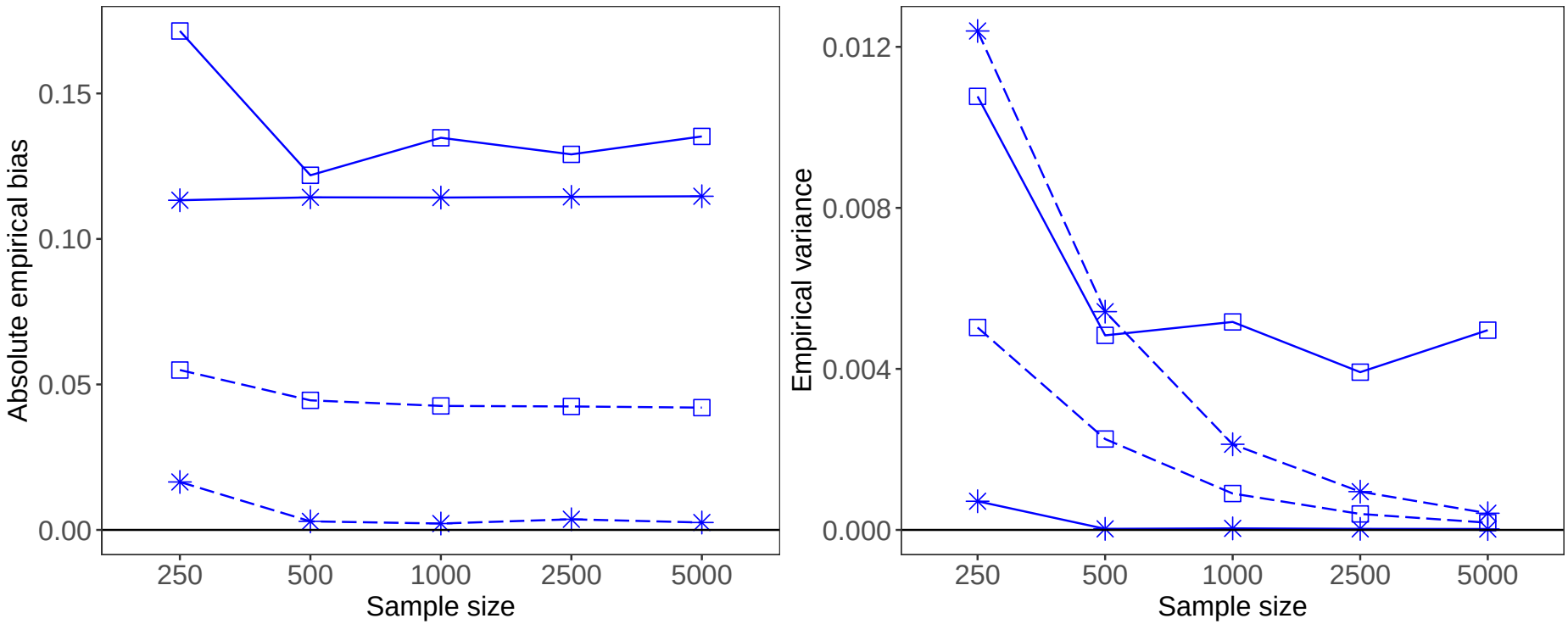


Simulations: robustness

Scenario: conditional event distribution estimator misspecified

Empirical bias and variance near zero using

- indirect debiasing
- doubly-robust pseudo-outcome estimation of $(f_0, f_{0,s})$

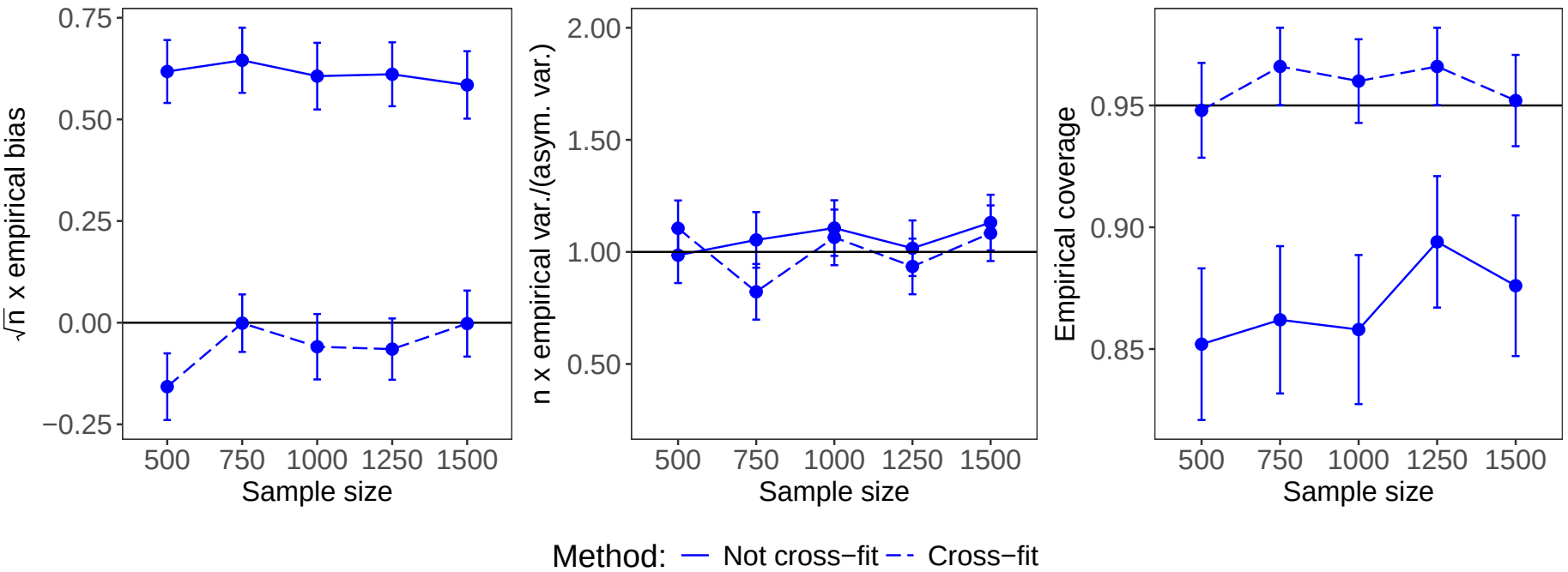


Simulations: $n^{-\frac{1}{2}}$–rate estimation and inference

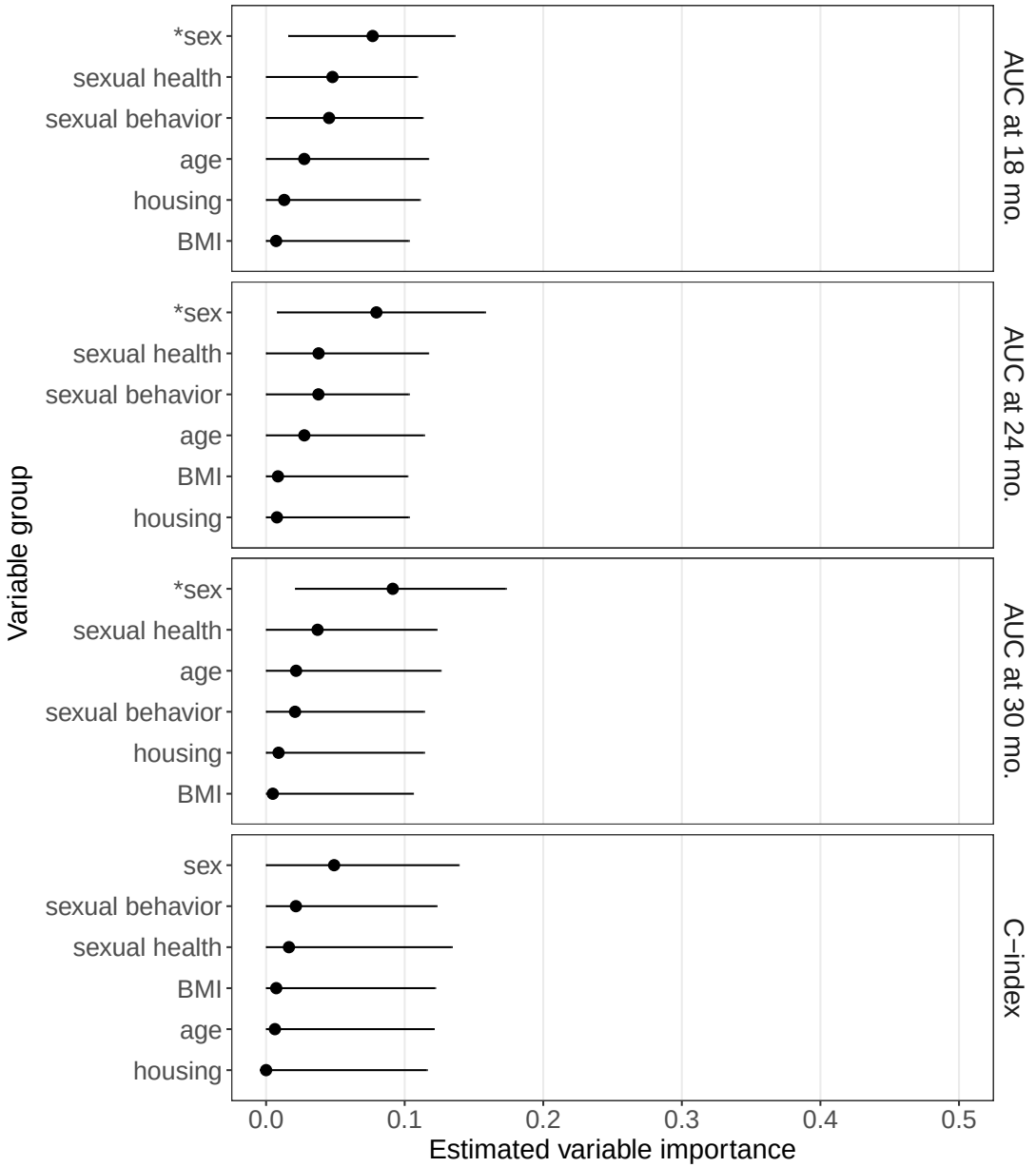
Scenario: nuisances estimated using the global survival stacking ensemble learner (Wolock et al., 2024)

The cross-fitted estimator has

- second-order bias,
- variance proportional to the nonparametric efficiency bound, and
- coverage near the nominal level.



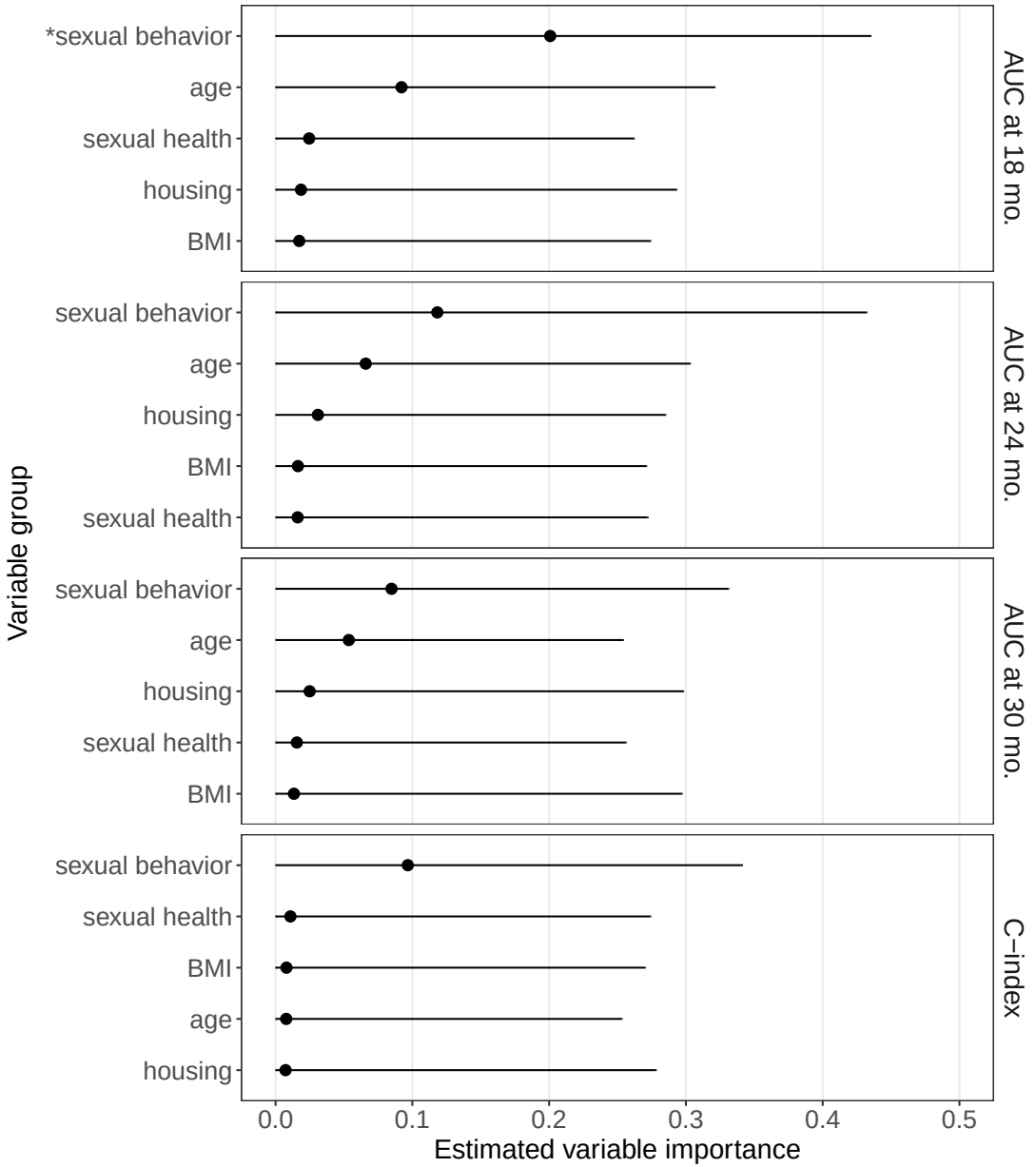
Variable importance in HVTN 702



Full cohort

- Sex assigned at birth is clearly an important predictor.
- Qualitative results are fairly stable across time horizons for AUC.

Variable importance in HVTN 702



Assigned male sex at birth

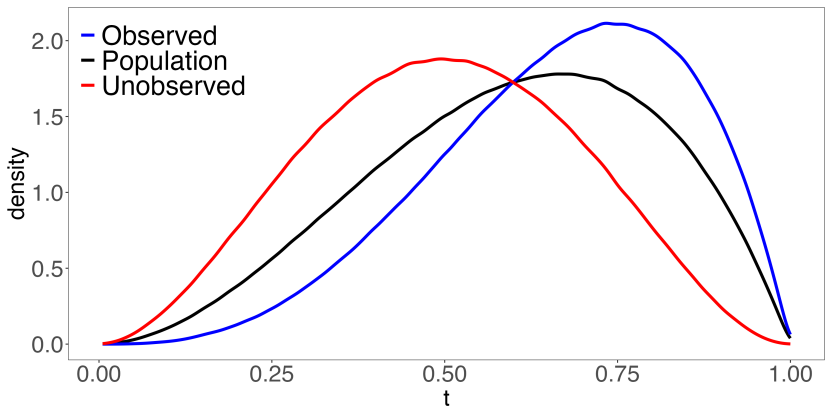
- Sexual behavior appears most important, but uncertainty is high.

Variable importance and EHR data

Assessing variable importance can be more complicated when using EHR data.

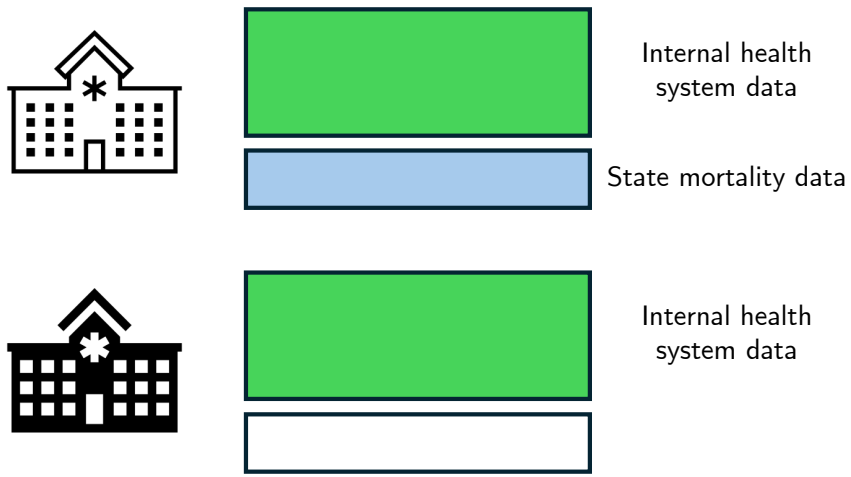
Truncation-induced selection bias:
Patients enter the study after the initiating event and before the terminating event.

Example: Cancer patients who must undergo genetic testing after diagnosis to enter the study.



Truncation induces selection bias.

Limited outcome labeling:
Ascertainment of suicide death requires linkage with state mortality records; this information is missing for certain health systems.

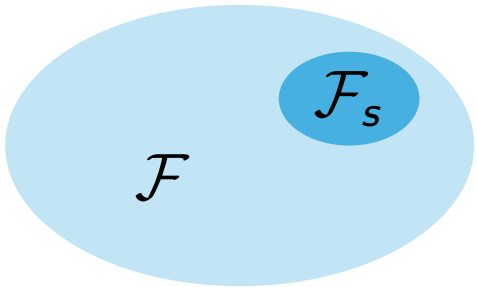


Morenz, Wolock, Carone. "Debiased machine learning for counterfactual survival functionals based on left-truncated right-censored data."
Wolock, Yan, Ning, Chen. "Transfer learning for model-free variable importance." (in progress)

Opportunities: inference under the null

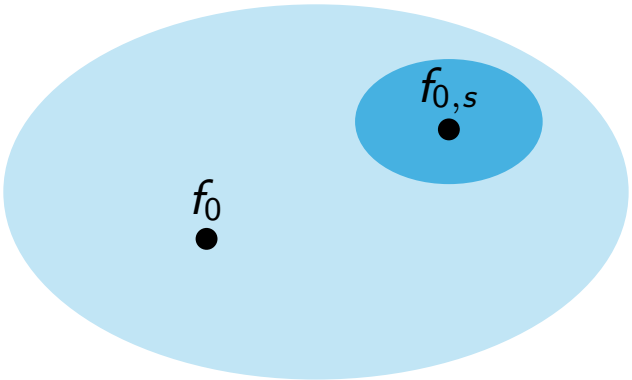
Recall: We compare the predictiveness of

- $f_0 \in \mathcal{F}$, an unrestricted class of prediction functions;
- $f_{0,s} \in \mathcal{F}_s$, the class of prediction functions that does not use X_s.

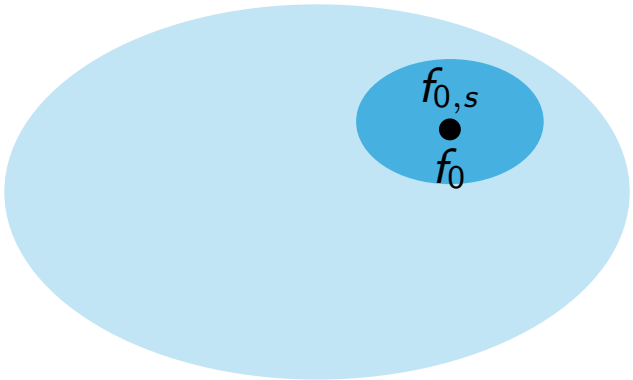


Suppose we wish to test the null hypothesis $H_0 : \psi_{0,s} = 0$.

$H_1 : \psi_{0,s} > 0$



$H_0 : \psi_{0,s} = 0$

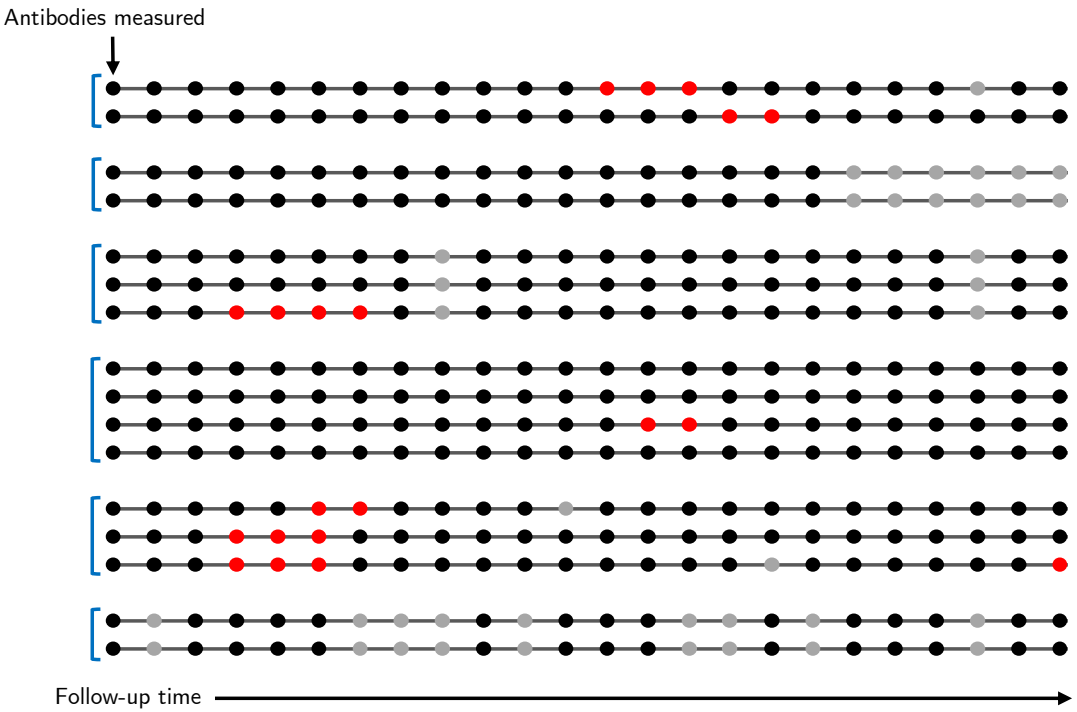


$V(f_n, H_n^*)$ and $V(f_{n,s}, H_n^*)$
have identical influence functions.

See Dai et al. (2022) and Hudson (2023) for work in this area.

Opportunities: clustered data

CASCADIA study: How important are levels of neutralizing and binding antibodies, measured at baseline, for predicting risk of SARS-CoV-2 infection?



- Recruitment is by household
  1. Data units correlated
  2. Variable # of individuals per household
- Open questions:
  1. How to debias?
  2. How to make inference?

Wrapping up

A nonparametric variable importance analysis can help provide insight into a given prediction task and inform future data collection practices.

Our proposed framework

- is nonparametric and algorithm-agnostic;
- accommodates censoring informed by measured covariates;
- encompasses commonly used predictiveness measures;
- provides doubly-robust estimation and calibrated statistical inference while allowing for flexible nuisance estimation.

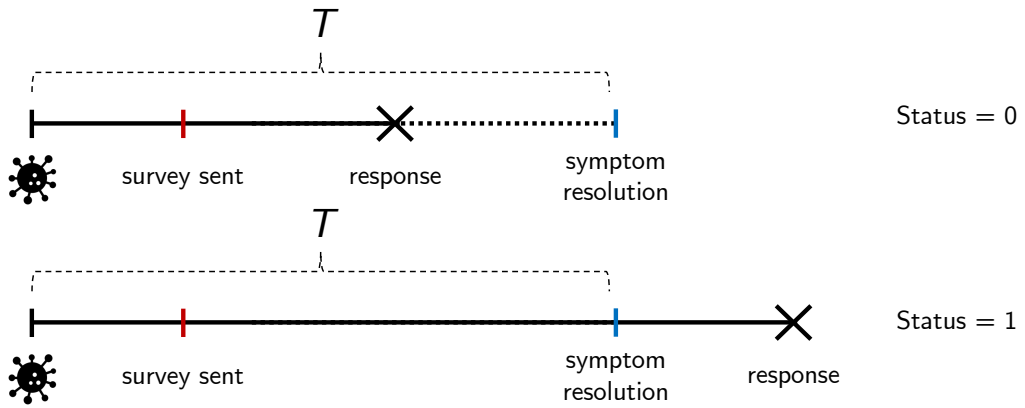
Interesting and practically important work remains to be done, including

- improved procedures for inference under the null;
- efficient estimation and inference with correlated data.

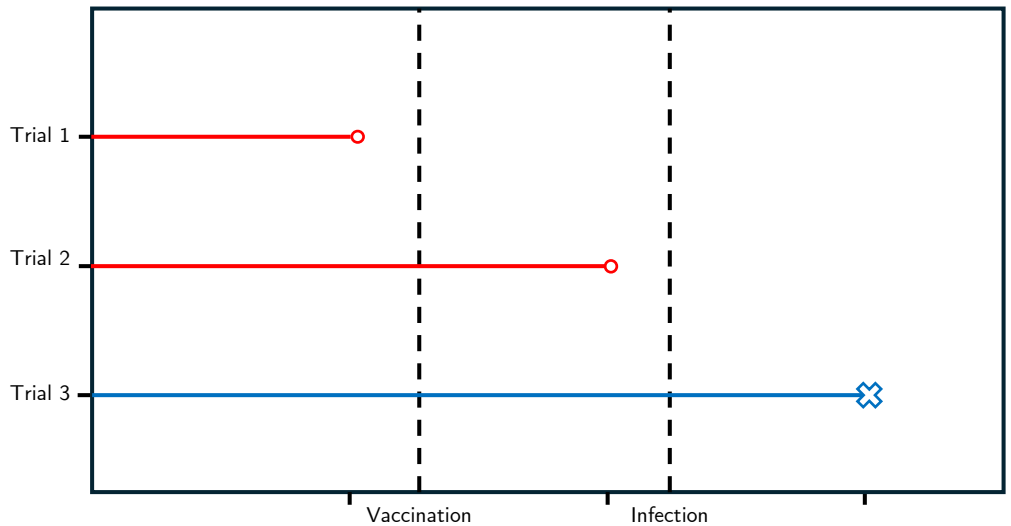
See the `survML` package on CRAN for implementation of the methods discussed today.

Other work

Estimating symptom duration following SARS-CoV-2 infection using current status data



Estimating causal effects from EHR data with underreported exposure



Wolock et al., "Investigating symptom duration using current status data: a case study of post-acute COVID-19 syndrome."
Wolock et al., "Estimating causal effects from electronic health records data with underreported exposure." (in progress)

Conclusion

Thank you for listening!

References

Babu et al. (2023). CASCADIA: a prospective community-based study protocol for assessing SARS-CoV-2 vaccine effectiveness in children and adults using a remote nasal swab collection and web-based survey design. *BMJ Open*, 13(7):e071446.

Benkeser et al. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Brier (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Chernozhukov et al. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68.

Cox (1972). Regression models and life-tables. *JRSSB*, 34(2):187–202.

Dai et al. (2022). Significance tests of feature relevance for a black-box learner. *IEEE Transactions on Neural Networks and Learning Systems*.

Fisher et al. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *JMLR*, 20(177):1–81.

Gilbert et al. (2022). A Covid-19 milestone attained—a correlate of protection for vaccines. *NEJM*, 387(24):2203–2206.

References

Gray et al. (2021). Vaccine efficacy of ALVAC-HIV and bivalent subtype C gp120–MF59 in adults. *NEJM*, 384(12):1089–1100.

Harrell et al. (1982). Evaluating the yield of medical tests. *JAMA*, 247:2543–6.

Heagerty and Zheng (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61:92–105.

Hooker, Mentch, and Zhou (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31:1–16.

Hudson, A. (2023). Nonparametric inference on non-negative dissimilarity measures at the boundary of the parameter space. *arXiv:2306.07492*.

Ishwaran et al. (2008). Random survival forests. *AoAS*, 2(3):841 – 860.

Lei et al. (2018). Distribution-free predictive inference for regression. *JASA*, 113(523):1094–1111.

Mayr and Schmid (2014). Boosting the concordance index for survival data - a unified framework to derive and evaluate biomarker combinations. *PLoS ONE*, 9.

Morenz*, Wolock*, and Carone (2024). Debiased machine learning for counterfactual survival functionals based on left-truncated right-censored data. *arXiv:2411.09017.*

Pfanzagl (1982). *Contributions to a general asymptotic statistical theory*. Springer.

References

Rubin and van der Laan (2007). A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics*, 3.

van der Laan and Rubin (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Verdinelli and Wasserman (2024). Decorrelated variable importance. *JMLR*, 25(7):1–27.

Wang et al. (2024). Targeted Learning for Variable Importance. *arXiv:2411.02221*.

Westling et al. (2024). Inference for treatment-specific survival curves using machine learning. *JASA*, 119(546):1541–1553.

Williamson, Gilbert, Simon, and Carone (2023). A general framework for inference on algorithm-agnostic variable importance. *JASA*, 118(543):1645–1658.

Wolock et al. (2024a). Importance of variables from different time frames for predicting self-harm using health system data. *Journal of Biomedical Informatics*.

Wolock et al. (2024b). Investigating symptom duration using current status data: a case study of post-acute COVID-19 syndrome. *arXiv:2407.04214*.

Wolock, Gilbert, Simon, and Carone (2024c). Assessing variable importance in survival analysis using machine learning. *Biometrika*.

Wolock, Gilbert, Simon, and Carone (2024d). A framework for leveraging machine learning tools to estimate personalized survival curves. *JCGS*, 33(3):1098–1108.

Zheng and van der Laan (2011). Cross-validated targeted minimum-loss-based estimation. In van der Laan and Rose, editors, *Targeted Learning: Causal Inference for Observational Data*, pages 459–474. Springer.