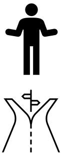
Causal prediction for medical decision making: Methods and Practice - day 3

Nan van Geloven

Evaluation of causal predictions



Recap evaluation of regular predictions

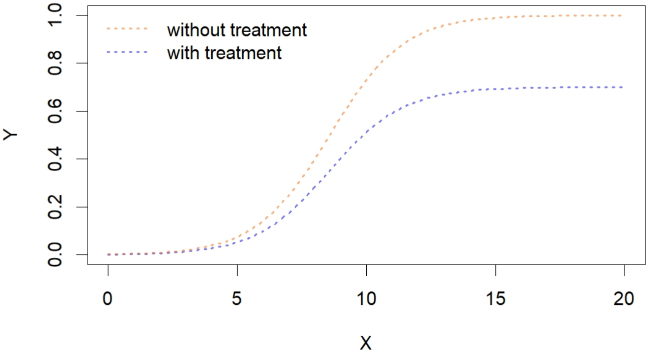
Goal: assess how well predictions match unseen observations.
Common metrics

- mean squared error / Brier score
- AUC
- calibration curve

Performance evaluation of causal predictions

- Not an issue when estimating ACE: nuisance models fitted on observed (not potential) outcomes
- Relevant for predictions under interventions to assess relation between potential outcomes and covariates:
  - model selection/internal validation: assess performance of alternative models in test sample(s)
  - out-of-sample/external validation: evaluate performance of certain model in a new dataset

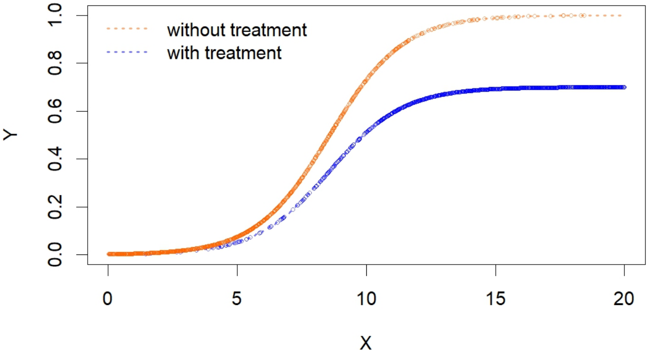
Toy example[1]



potential outcome distributions $Y^0(x)$ and $Y^1(x)$

[1] adapted from Doutreligne and Varoquaux 2023

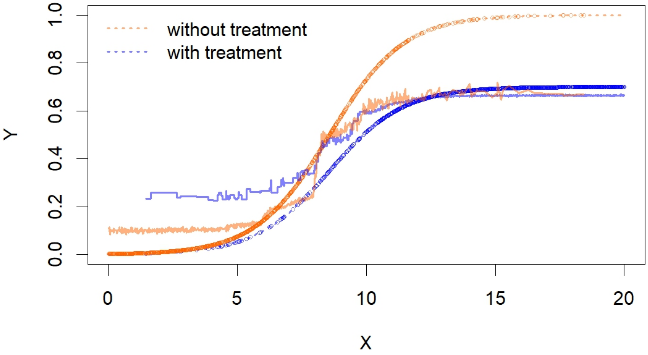
Toy example[1]



observed data points $Y(X|A=0)$ and $Y(X|A=1)$

[1] adapted from Doutreligne and Varoquaux 2023

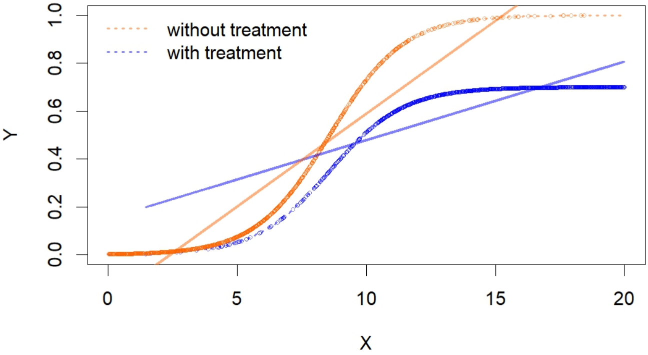
Toy example[1]



random forest:
R^2 observed outcomes = 0.92, R^2 potential outcomes = 0.77

[1] adapted from Doutreligne and Varoquaux 2023

Toy example[1]



linear model:

R^2 observed outcomes = 0.90, R^2 potential outcomes = 0.81

[1] adapted from Doutreligne and Varoquaux 2023

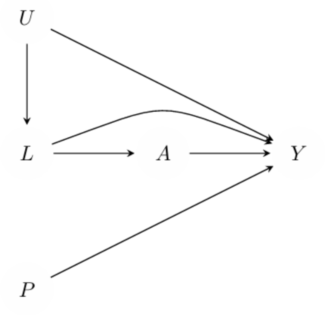
Challenge in evaluating performance of predictions under interventions

- Assess how well the predictions match "observed" outcomes in the test dataset
- Then "observed outcomes" in test dataset also need to be under the treatment strategies of interest
- These outcomes are not observable for all patients in observational data
- Need to estimate counterfactual "observed outcomes" in test dataset

For example in the Steno model

- This morning we developed a causal model predicting cvd outcomes in a scenario with and a scenario without statin initiation, using the training data
- now suppose we want to evaluate its performance in (observational) test data

Structure observational test data



P	(pure) prognostic factors
L	confounders
$X = \{L, P\}$	all covariates
$X^* = \{L^*, P^*\}$	subset of covariates in prediction model
A	treatment
Y	outcome

Recap prediction estimand

- Expected outcome under treatment *a*:

$$\begin{aligned}\mu_a(X^*) &= E(Y^a|X^*)\\ &= P(Y^a = 1|X^*) \qquad \text{(for binary Y)}\end{aligned}$$

  where Y^a is potential outcome Y if an individual would follow a and X^* are predictors in the model (may include only subset of L)
- Assume a candidate model $\hat{\mu}(X^*)$ has been developed before

Evaluation estimand

We want to assess performance of model $\hat{\mu}(X^*)$ under treatment level a in some observational test dataset D_{test}. For example mean squared error:

$$MSE^a = E[(Y^a - \hat{\mu}(X^*))^2 | D_{test} = 1)]$$

Similar challenge as before: we do not observe Y^a for all individuals in D_{test}

Two identification strategies[2]

Under the same assumptions as before (consistency, conditional exchangeability, positivity), but now in test data(!)

- inverse probability weighting:

$$MSE^a = E(\frac{I(A=a)}{P(A=a|X, D_{test}=1)}(Y-\hat{\mu}(X^*))^2|D_{test}=1)$$

- 'loss modeling':

$$MSE^a = E(E(Y-\hat{\mu}(X^*))^2|X, A=a, D_{test}=1)$$

[2] for proofs see Boyer et al. arXiv 2025

Estimation through inverse probability weighting

$$MSE^a = E\left(\frac{I(A=a)}{P(A=a|X, D_{test}=1)}(Y - \hat{\mu}(X^*))^2 | D_{test}=1\right)$$

$$\hat{MSE}^a = \frac{1}{n_{test}} \sum_{i=1}^{n} \left(\frac{I(A_i = a, D_{test}=1)}{\hat{P}(A=a|X, D_{test}=1)}(Y_i - \hat{\mu}(X_i^*))^2\right)$$

- Restrict to individuals in test set who followed treatment *a*
- calculate their regular contribution to MSE
- reweigh to extrapolate to full population
- assumes (next to identification assumptions) correct specification of the weights model

Example code MSE-IPW

Regular evaluation

```
MSE <- 1/ntest * sum((Y - predictions)^2)
```

IPW evaluation scenario no treatment

```
MSE0 <- 1/ntest * sum((Y - predictions)^2 *
(A==0) * weights)
```

IPW evaluation scenario with treatment

```
MSE1 <- 1/ntest * sum((Y - predictions)^2 *
(A==1) * weights)
```

Estimation through 'loss modelling'[3]

$$MSE_a = E(E(Y - \hat{\mu}(X^*))^2 | X, A = a, D_{test} = 1)$$

$$\hat{MSE}_a = \frac{1}{n_{test}} \sum_{i=1}^{n} I(D_{test} = 1)\hat{h}_a(X_i),$$

with $\hat{h}_a(X_i)$ an estimator for conditional loss:

$$E((Y - \hat{\mu}(X^*))^2 | X, A = a, D_{test} = 1)$$

For binary Y and $X^* = X$, you only need outcome model $E[Y|X, A = a]$

[3] Boyer et al. arXiv 2025

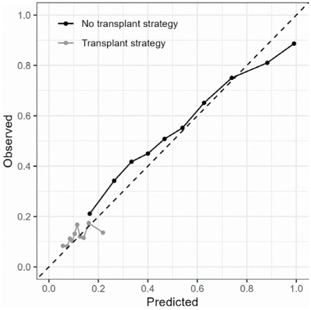
Calibration curve

Do estimated risks match "observed" outcomes?

- plot of observed outcomes by expected risk :

$$p \to P[Y^a = 1 | \hat{\mu}(x^*) = p]$$

- Estimated with IPW or through outcome modelling
- Example:



Discrimination measures

Pairs of individuals with and without event (i, j) are evaluated for whether the individual with event was assigned the higher risk by the model.

- Counterfactual AUC:

$$P(\hat{\mu}(X_i^*) > \hat{\mu}(X_j^*) | Y_i^a = 1, Y_j^a = 0)$$

- IPW estimation needs weights for the pair $i, j : w_{ij} = w_i * w_j$

References

1. Doutreligne and Varoquaux, arXiv 2023
2. Pajouheshnia et al. BMC Med Res Meth 2017
3. Boyer et al., arXiv 2025
4. Keogh and Van Geloven, Epidem 2024