Prediction under hypothetical interventions: evaluation of counterfactual performance using longitudinal observational data

Nan van Geloven

n.van_geloven@lumc.nl
Department of Biomedical Data Sciences
Leiden University Medical Center, the Netherlands

joint work with Ruth Keogh (LSHTM, UK)

Oct 9th 2023

Why do we need prediction under interventions?

1. For **informing individual treatment decisions** we would like to know
   - an individual's expected outcome if they were to receive the treatment
   - an individual's expected outcome if they were not to receive the treatment

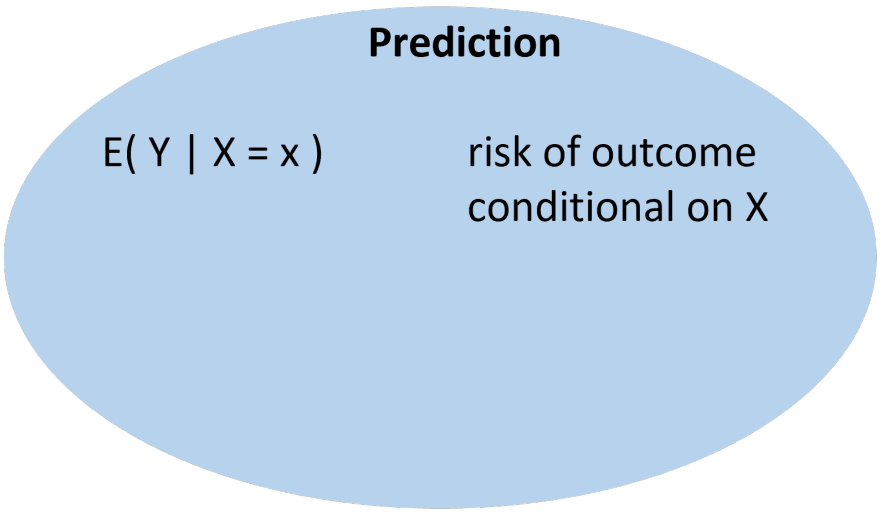
Why do we need prediction under interventions?

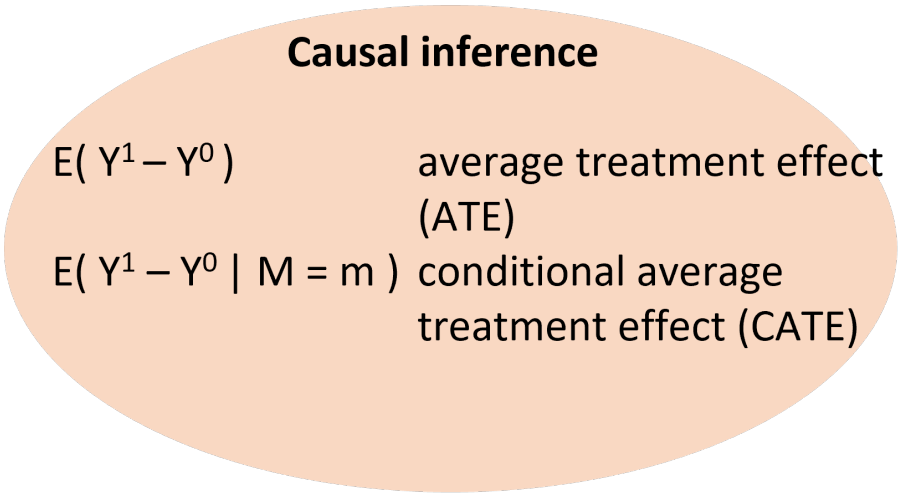
1. For **informing individual treatment decisions** we would like to know
   - an individual's expected outcome if they were to receive the treatment
   - an individual's expected outcome if they were not to receive the treatment
2. Prediction under **dataset shift**. When treatment policy is different in deployment than in development setting we would like to know
   - individuals' expected outcomes if treatment were to be administered as in deployment setting

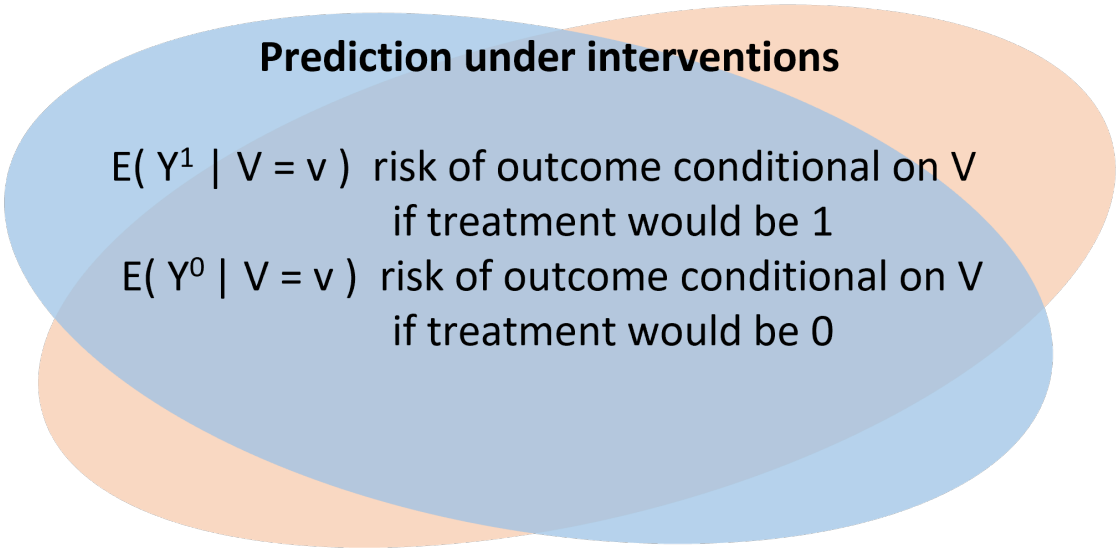
Why do we need prediction under interventions?

1. For **informing individual treatment decisions** we would like to know
   - an individual's expected outcome if they were to receive the treatment
   - an individual's expected outcome if they were not to receive the treatment
2. Prediction under **dataset shift**. When treatment policy is different in deployment than in development setting we would like to know
   - individuals' expected outcomes if treatment were to be administered as in deployment setting

Unless estimated from randomized studies, these expected outcomes (risks) are counter to the fact for a subset of the individuals in the development data set.


Prediction
E(Y | X = x)
risk of outcome conditional on X


Causal inference
E(Y1 – Y0)
average treatment effect (ATE)
E(Y1 – Y0 | M = m)
conditional average treatment effect (CATE)


Prediction under interventions
E(Y1 | V = v)
risk of outcome conditional on V if treatment would be 1
E(Y0 | V = v)
risk of outcome conditional on V if treatment would be 0

Development of predictions under interventions

Predictions under interventions: estimates of risk under different possible treatments/interventions, while also accounting for other patient characteristics that are predictive of the outcome.

- By secondary analysis of randomized trial data
- Combining observational data with estimates of treatment effects from trials
- From observational data using e.g. MSM-IPTW[1,2], Cens-IPW[2] or g-formula[3]

[1] Sperrin et al. 2018
[2] van Geloven et al. 2020
[3] Dickerman et al. 2022

Evaluating performance of predictions under interventions

- Assess how well the predictions match observed outcomes in a (new) dataset, e.g. to inform model selection
- Challenge in observational validation data sets: outcomes under treatment strategy of interest are not observable for all patients.
- Aim of this work: propose methods for evaluation of counterfactual predictive performance for time-to-event outcomes

Previous work

- Pajouheshnia et al. (2017): studied O:E ratio and c-index estimated by IPW for point treatment and binary outcome

Previous work

- Pajouheshnia et al. (2017): studied O:E ratio and c-index estimated by IPW for point treatment and binary outcome
- Sperrin et al. (2018): studied predictive performance in the subset of patients who did not initiate statins during follow up -> selected validation sample

Previous work

- Pajouheshnia et al. (2017): studied O:E ratio and c-index estimated by IPW for point treatment and binary outcome
- Sperrin et al. (2018): studied predictive performance in the subset of patients who did not initiate statins during follow up -> selected validation sample
- Review by Lin et al. (2021) found 0/13 models assessed performance: "The most pressing problem to address for predictions under hypothetical interventions is model validation."

Previous work

- Pajouheshnia et al. (2017): studied O:E ratio and c-index estimated by IPW for point treatment and binary outcome
- Sperrin et al. (2018): studied predictive performance in the subset of patients who did not initiate statins during follow up -> selected validation sample
- Review by Lin et al. (2021) found 0/13 models assessed performance: "The most pressing problem to address for predictions under hypothetical interventions is model validation."
- Boyer et al. (sept 2023): model performance for time-varying treatment and binary outcome using IPW, conditional loss function and a doubly robust approach for squared error loss

This work (https://arxiv.org/abs/2304.10005)

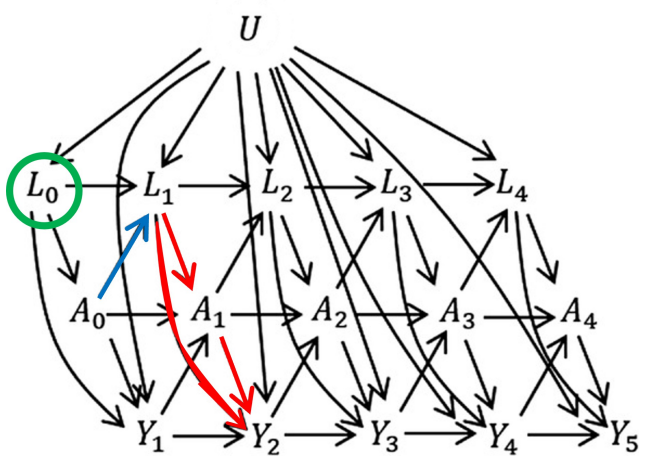
Prediction under hypothetical interventions: evaluation of performance using longitudinal observational data

RUTH H. KEOGH & NAN VAN GELOVEN†

Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, UK
Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, NL
ruth.keogh@lshtm.ac.uk, n.van_geloven@lumc.nl
† The two authors contributed equally.

- Validation of predictions under sustained treatment strategies using observational data with time-to-event outcome
- Extensions of performance measures: calibration, discrimination (c-index and AUCt), Brier score
- Simulations
- Application: mortality risk for liver patients when receiving or not receiving a transplant

Observational data structure (Keogh et al 2021)



U	unobserved covariate
L_0	baseline covariates used when estimating risk
L_k	time-dependent confounders
A_k	treatment status at visits $k = 0, 1, 2, ...$
$a = a_0, a_1, \ldots$	treatment pattern over time
T, D	(continuous) time to event plus status

Model development

- factual risk:

$$R(\tau|L_0) = P(T \leq \tau|L_0)$$

- risk under intervention a:

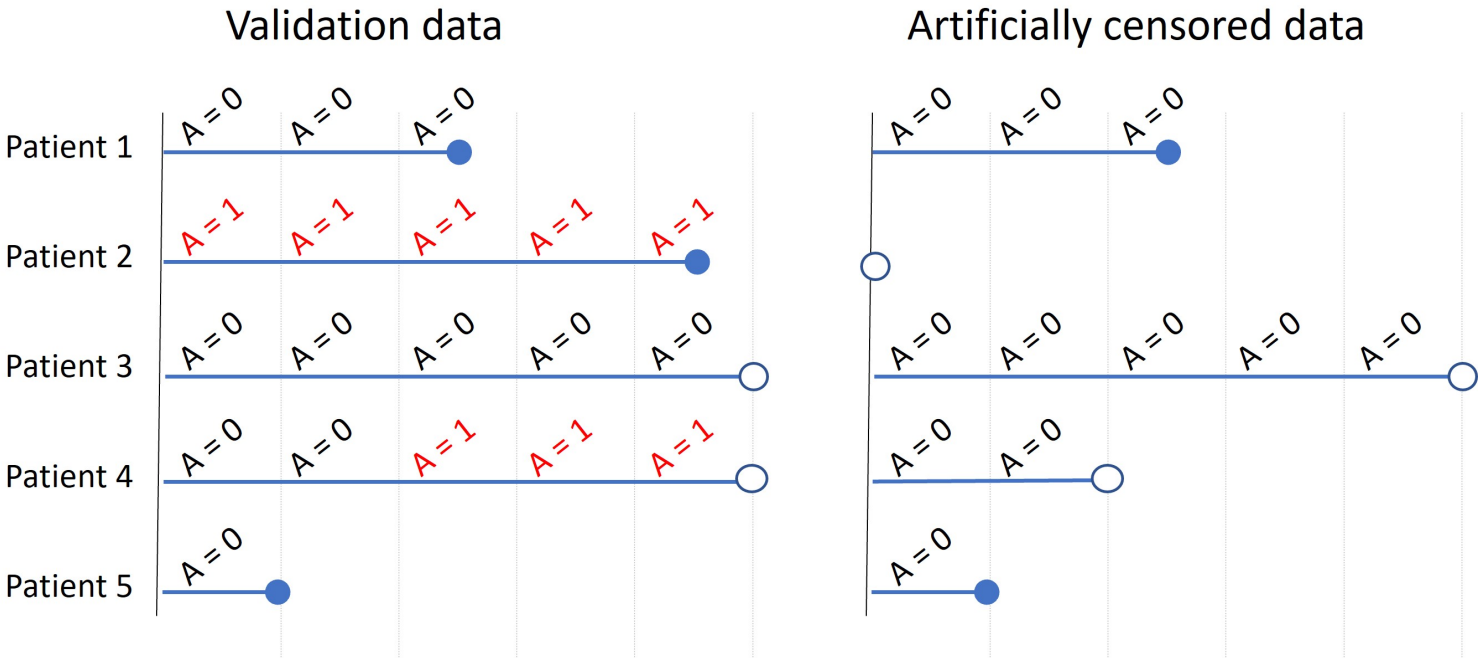
$$R^a(\tau|L_0) = P(T^a \leq \tau|L_0)$$

where T^a is counterfactual T if an individual would follow a

- We assume a model for untreated risk $a = (0, 0, 0, 0, 0)$ has been developed and we want to assess performance of estimates $\hat{R}^a(\tau|L_0)$ in a new dataset

Mimic the treatment strategy under which predictions are made

Artificially censor individuals when they deviate from the strategy of interest, for instance if $a = (0, 0, \cdots)$



Artificially censored data: $(\tilde{T}_a, \tilde{D}_a)$

Use IPCW to address this artificial censoring

- Let G be the conditional survival function of the artificial censoring times:

$$G_a(t|L) = \prod_{s=0}^{\lfloor t \rfloor} \Pr(A_s = a | A_{s-1} = a, \bar{L}_s)$$

where $\bar{L}_s = L_0, \ldots, L_s$ is the covariate history up to s

Use IPCW to address this artificial censoring

- Let G be the conditional survival function of the artificial censoring times:

$$G_a(t|L) = \prod_{s=0}^{\lfloor t \rfloor} \Pr(A_s = a | A_{s-1} = a, \bar{L}_s)$$

  where $\bar{L}_s = L_0, \ldots, L_s$ is the covariate history up to s
- Weighing by G_a^{-1} forms a population in which all individuals had followed the strategy under evaluation
- under the assumptions of consistency, conditional sequential exchangeability, positivity and correct model specification of $\hat{G}_a$

Calibration measures

Do estimated risks match "observed" outcomes?

- observed versus expected risk split up in subgroups defined by expected risk (calibration curve)
- "observed versus expected ratio" based on risks
- "observed versus expected ratio" based on number of events

Observed outcomes estimated by weighted Kaplan-Meier or weighted Nelson-Aalen

Discrimination measures

Are higher risks assigned to individuals who experience the event earlier?

- c-index

$$C_\tau^a = P(\hat{R}_i^a(\tau) > \hat{R}_j^a(\tau) | T_i^a < T_j^a, T_i^a \leq \tau)$$

- cumulative dynamic AUC(t)

$$AUC^a(t) = P(\hat{R}_i^a(t) > \hat{R}_j^a(t) | T_i^a \leq t, T_j^a > t),$$

Proposed estimator for C-index

$$\hat{C}^a(\tau) = \frac{\sum_{i=1}^n \sum_{j=1}^n I(\hat{R}_i^a(\tau) > \hat{R}_j^a(\tau)) \text{comp}_{aij}^{(1)}(\tau) \hat{W}_{aij}^{(1)}}{\sum_{i=1}^n \sum_{j=1}^n \text{comp}_{aij}^{(1)}(\tau) \hat{W}_{aij}^{(1)}}$$

with $\hat{W}_{aij}^{(1)} = \hat{G}_{ac}^{-1}(\tilde{T}_{ai}^{-}|L_i)\hat{G}_{ac}^{-1}(\tilde{T}_{ai}|L_j)$

and $G_{ac}^{-1}(t|L) = G_a^{-1}(t|L) \times G_c^{-1}(t)$ combines weights for artificial censoring with weigths for 'standard' (non-informative) censoring.

Proposed estimator for C-index

$$\hat{C}^a(\tau) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I(\hat{R}_i^a(\tau) > \hat{R}_j^a(\tau))\text{comp}_{aij}^{(1)}(\tau)\hat{W}_{aij}^{(1)}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \text{comp}_{aij}^{(1)}(\tau)\hat{W}_{aij}^{(1)}}$$

with $\hat{W}_{aij}^{(1)} = \hat{G}_{ac}^{-1}(\tilde{T}_{ai}^{-}|L_i)\hat{G}_{ac}^{-1}(\tilde{T}_{ai}|L_j)$

and $G_{ac}^{-1}(t|L) = G_a^{-1}(t|L) \times G_c^{-1}(t)$ combines weights for artificial censoring with weigths for 'standard' (non-informative) censoring.

Extension of Gerds et al. (2013)
We make a similar extension for AUC(t)

Brier score

Expected squared difference between event indicator and estimated risk

$$E[(I(T^a \leq t) - \hat{R}^a(t))^2]$$

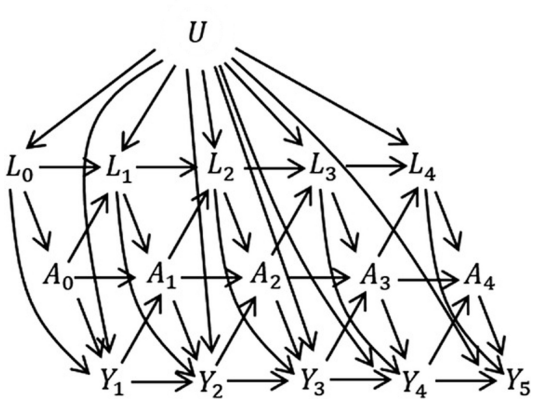
Proposed estimator:

$$\hat{BS}^a(t) = \frac{1}{n}\sum_{i=1}^{n}((I(\tilde{T}_{ai} \leq t) - \hat{R}_i^a(t))^2 W_{ai})$$

with $W_{ai} = \frac{I(\tilde{T}_{ai} \leq t, \tilde{D}_{ai}=1)}{\hat{G}_{ac}(\tilde{T}_{ai}|L_i)} + \frac{I(\tilde{T}_{ai} > t)}{\hat{G}_{ac}(t|L_i)}$.

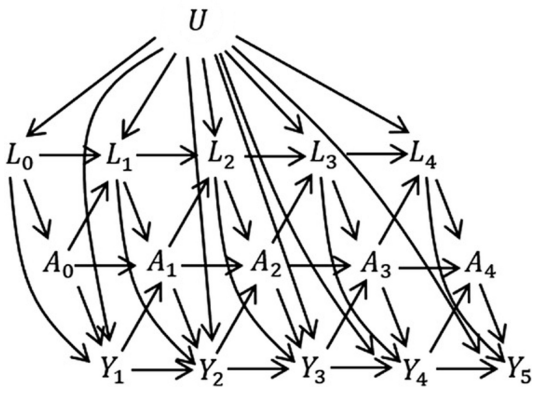
Simulation setup

development data



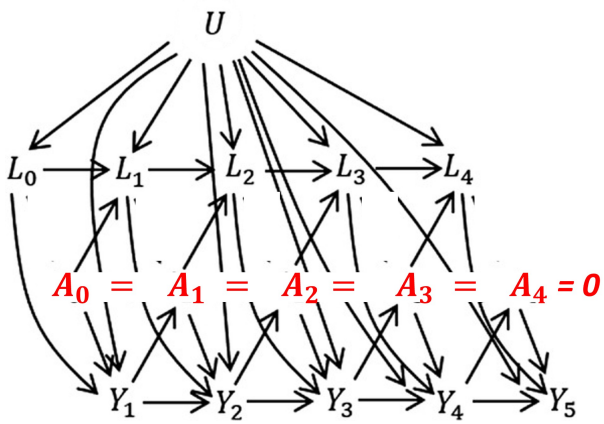
obtain counterfactual prediction model (using MSM-IPTW)

validation data



evaluate predictive performance

counterfactual validation data

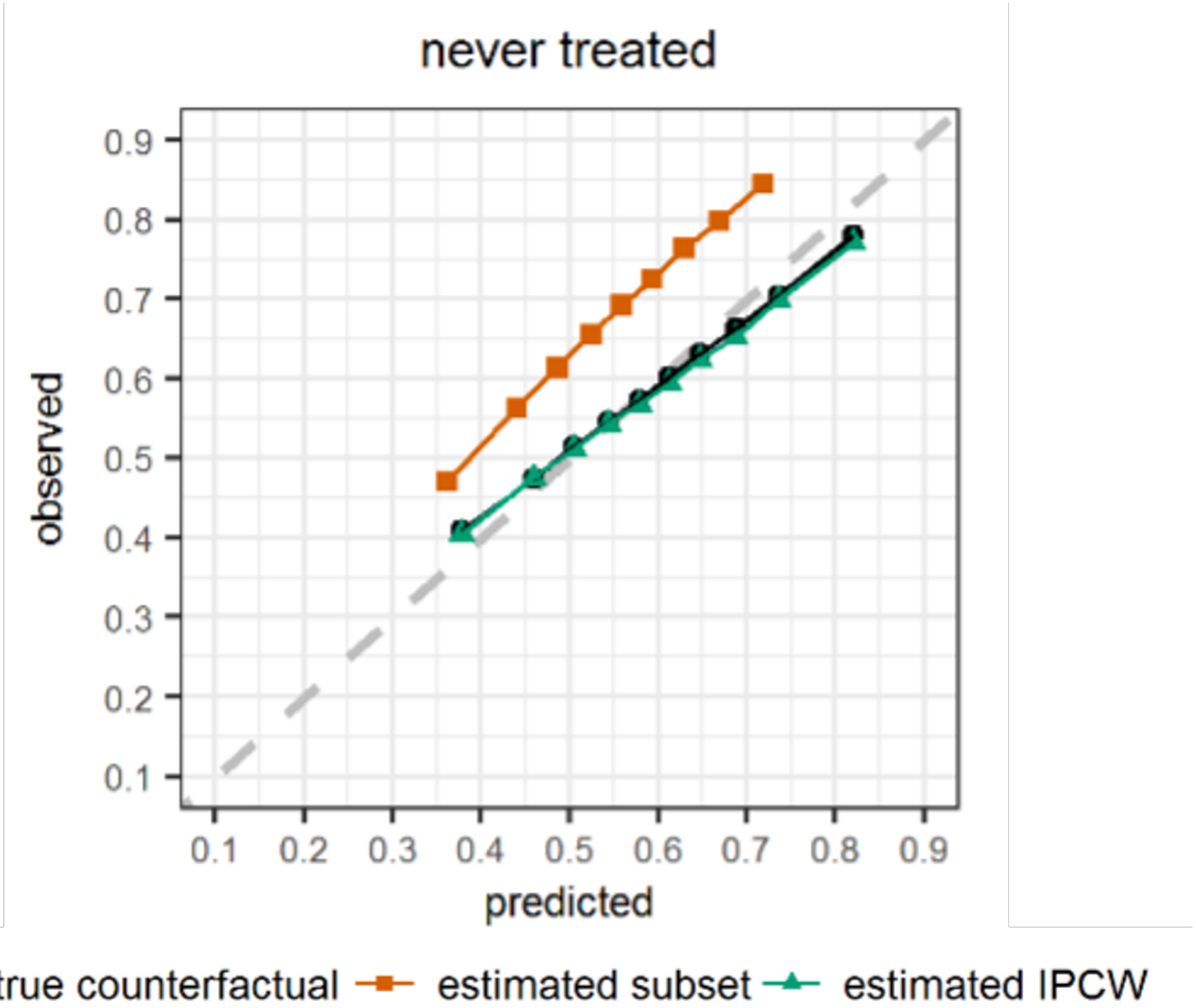


evaluate predictive performance (true values)

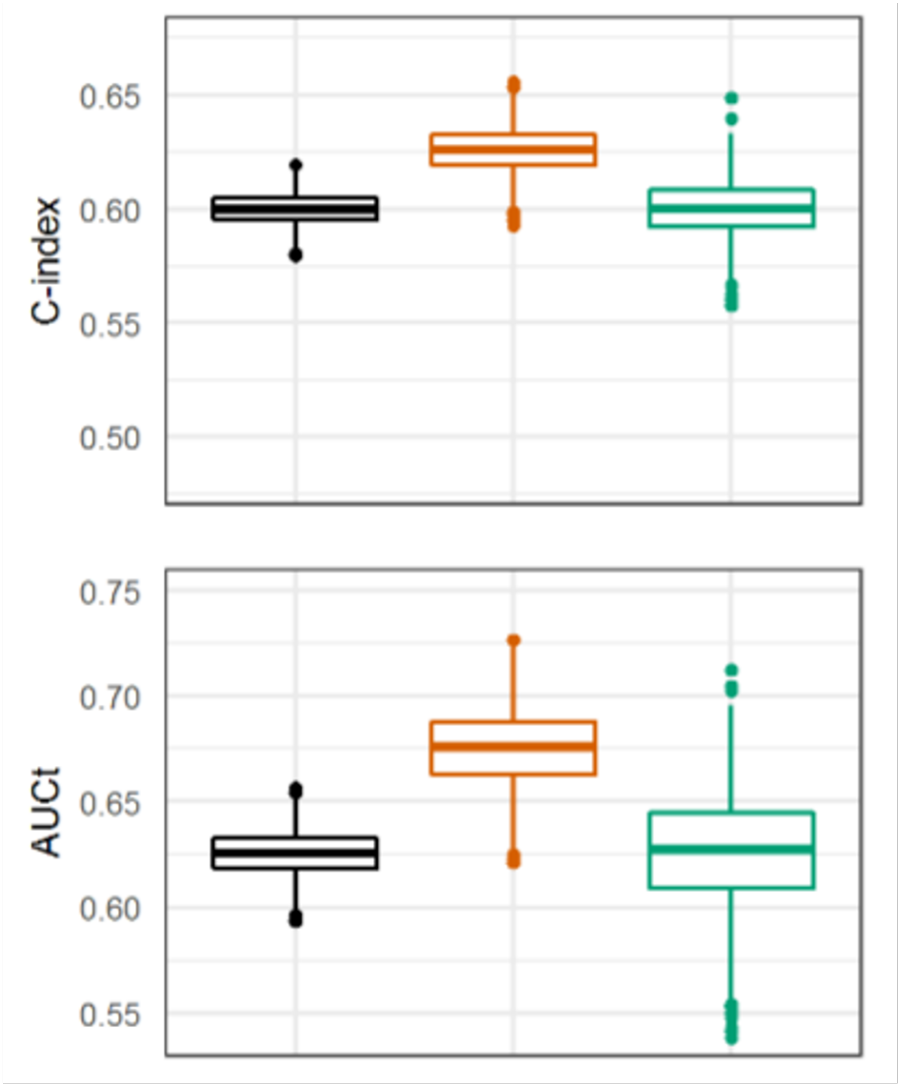
Simulation results

- Data generated and analysed using Cox proportional hazards models and Aalen additive hazards models
- Including scenarios where we expect good and bad predictive performance of predictions under interventions
  - higher baseline hazard in development data
  - measurment error when applying the development model
  - conditional Cox model $\neq$ marginal Cox model
- Conclusion: it works!
- Simulations also show the bias introduced by the 'subset' approach

Results calibration



Results discrimination



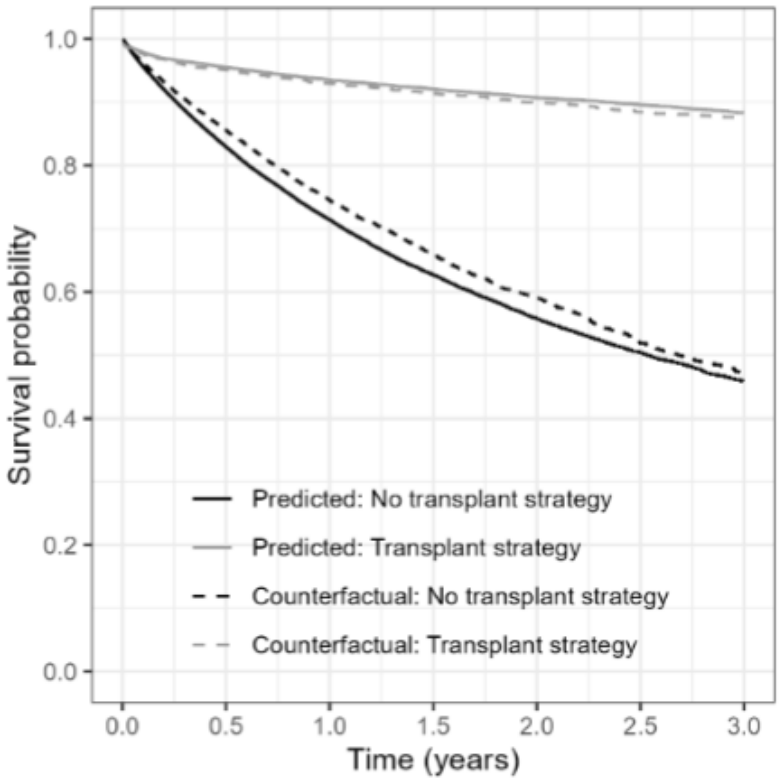
true counterfactual — estimated subset — estimated IPCW

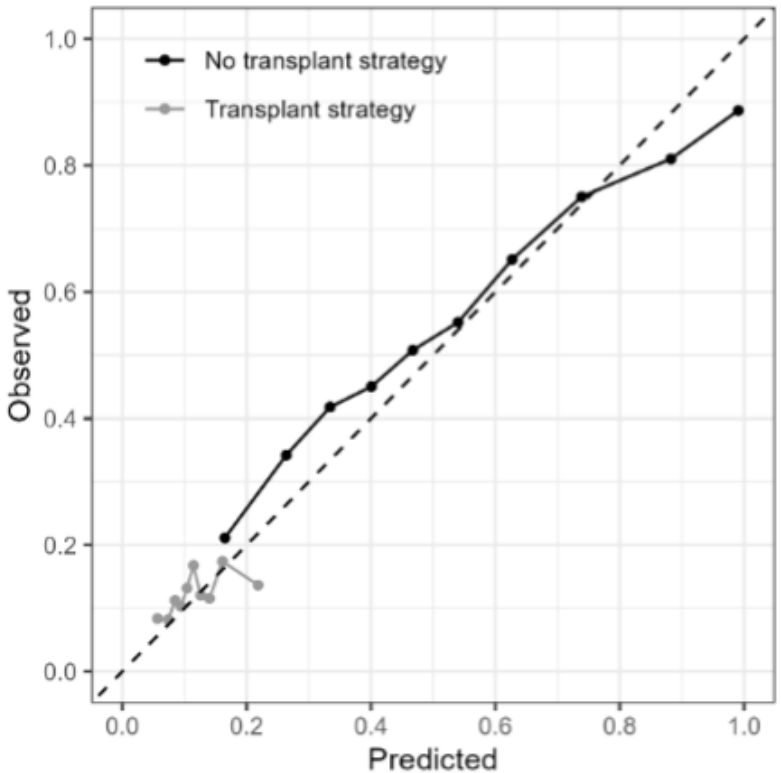
UNOS transplant data

- US data on patients waitlisted for a liver transplant from the United Network for Organ Sharing (UNOS)
- n=30203 patients (70%) used for development
- n=12987 patients (30%) used for validation
- Estimate risks of composite outcome of death or removal from waiting list due to worsening health condition up to 3 years under the interventions of:
  - receiving a liver transplant
  - not receiving a transplant

  conditional on their characteristics at moment of making the prediction (about 30 parameters)

Results transplant data I





Results transplant data II

	Strategy	
	No transplant	Transplant
Calibration: OE ratio based on risk by 3 years	0.983	1.060
Discrimination: C-index up to 3 years	0.749	0.561
Discrimination: AUCt at 3 years	0.781	0.552
Prediction error: scaled Brier score (%) at 3 years	66.8	12.0

Conclusions

- Our approach to counterfactual performance evaluation using artificial censoring + IPCW gives unbiased estimates of predictive performance when weights are correct
- Current work: what can be expected when assumptions do not hold?
- Future work:
  - work out how to combine with cross-validation / bootstrapping
  - compare to alternative proposal using g-formula (Dickerman et al 2022)
  - towards doubly robust approach
  - extend to competing risks

Invite to "Causal inference for AI in health" seminar series

Similar seminar series by causal inference researchers in Leiden/Delft/Rotterdam. Everyone is welcome.

Next meeting Oct 23 15.00 at LUMC:

- Maurice Korf: *Carefully Causal: an R function to improve causal inference in applied epidemiology*
- Jim Smit: *Asking what If? in the Intensive Care: a review of applied causal inference for time-varying treatments*
- Marta Spreafico: *Investigating positivity violations in marginal structural survival models: a study on IPTW estimator performance*

Sign up to our mailing list through this google form

References

1. Sperrin et al Stat Med 2018
2. Van Geloven et al. Eur J of Epidem 2020
3. Dickerman et al. Eur J Epidemiol 2022
4. Pajouheshnia et al. BMC Med Res Meth 2017
5. Lin et al. Diagn Prog Res 2021
6. Keogh et al. Biom J 2021
7. Efthimiou et al. Stat Med 2023
8. Boyer, arXiv Sept 2023
9. Gerds et al. Stat Med 2013

n.van_geloven@lumc.nl