Evidence-Based Medicine Meets Causal Inference

Presentation @ Econometric Institute Internal Seminar

Max Welz [1] Kevin ten Haaf [2]

[1] Erasmus University, Dept. of Econometrics

[2] Erasmus MC, Dept. of Public Health

July 21, 2020

Outline

Outline

Introduction (1/3)

Some Context

- Joint work with the Department of Public Health at Erasmus.
- Interdisciplinary project between statistics/econometrics and medicine.
- Initiated as my Master's thesis in econometrics in which I used state-of-the-art causal inferential methods to identify causal heterogeneity in Erasmus MC's NELSON trial
  - NELSON (de Koning et al., 2020) is the world's second largest lung cancer screening trial.
  - My thesis' results suggest that standard medical methods for the analysis of RCTs may only capture heterogeneity insufficiently.

➔ Show clinicians the added value of incorporating the results of modern causal inferential models in their decision-making.

Introduction (2/3)

Terminology

- Evidence-based medicine: the "conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett et al., 1996).
  - ➔ Reliable methods to quantify evidence from medical trials are required, especially if there is heterogeneity across subgroups in the trials: Discern signal from noise!
- Causal inference: the "Identification of the cause or causes of a phenomenon, by establishing covariation of cause and effect, a time-order relationship with the cause preceding the effect, and the elimination of plausible alternative causes" (Shaughnessy et al., 2000).
- The term treatment refers to an economic treatment (which is not necessarily medical) here.

Introduction (3/3)

Brief Summary of the Project

- Medical models for trial evaluations are quite different to the models from causal inference that econometricians are used to.
- We demonstrate how modern models from causal inference can yield valuable additional insights for evidence-based and personalized medicine.
- We propose to consider the results of an array of medical and causal inferential methods before making medical decisions.
- Target audience: clinicians. We want to raise awareness of the importance of using "good" models and therefore aim at a general medical journal.
- Relevance: for instance design of future national lung cancer screening programs in order to minimize costs and maximize efficacy (only treat groups that benefit sufficiently).

Outline

General Setup: The Potential Outcomes Framework

- Suppose there are n independent samples $(Y_i, \mathbf{X}_i, W_i)$, indexed by i.
  - The outcome of interest is denoted Y_i, the covariates are captured in p-vector $\mathbf{X}_i$, and W_i is a binary treatment indicator where $W_i = 1$ indicates an assignment to the treatment group and $W_i = 0$ to the control group.
- Given a realization $\mathbf{X}_i = \mathbf{x}$, the DGP can generally be viewed as

$$Y_i(W_i) = \theta(\mathbf{x})W_i + \nu(\mathbf{x}) + \varepsilon_i,$$

  where
  - $\theta(\mathbf{x})$ is the causal parameter: the causal effect of W_i on Y_i given $\mathbf{x}$.
  - $\nu(\mathbf{x})$ is the nuisance parameter which captures the raw effect of $\mathbf{x}$ on Y_i (excluding the effect of treatment W_i on Y_i). Usually highly nonlinear!
  - ε_i is an independent noise term that follows some zero-mean distribution.
- In practice, we only ever observe one of the two potential outcomes:

$$Y_i = \begin{cases} Y_i(1) & \text{if } W_i = 1, \\ Y_i(0) & \text{if } W_i = 0. \end{cases} \implies \text{The fundamental problem!}$$

Functions of the Causal Parameter

$$Y_i = \begin{cases} Y_i(1) & \text{if } W_i = 1, \\ Y_i(0) & \text{if } W_i = 0, \end{cases} \quad \text{where } Y_i(W_i) = \theta(\mathbf{X}_i)W_i + \nu(\mathbf{X}_i) + \varepsilon_i.$$

- An important quantity is the average treatment effect (ATE): $\text{ATE} = n^{-1} \sum_{i=1}^{n} \theta(\mathbf{X}_i)$.
  - If we are in an RCT (that is, random assignment of treatment W_i), an unbiased estimate of the ATE is given by $\mathbb{E}[Y_i|W_i = 1] - \mathbb{E}[Y_i|W_i = 0]$.
  - RCTs are standard in medicine, so obtaining the ATE is not a problem.
- Usually we are interested in estimating group-level ATEs or determining whether there is heterogeneity in $\theta(\mathbf{X}_i)$ and if yes, by what covariates that heterogeneity is driven by: subgroup analysis.
  - This is where models from econometrics and medicine start to differ considerably!

Subgroup Analysis in Econometrics

In econometrics, we usually estimate the group-level ATE as a function of per-individual estimates of the causal parameter. For instance,

- Chernozhukov et al. (2018) propose a method for quantifying causal group-level heterogeneity in RCTs that does not rely on strong assumptions.
- Athey et al. (2019) use a random-forest-based GMM-type estimator that yields asymptotically consistent estimates of the causal parameter. A "doubly-robust" estimator of group-level ATEs can then be applied on the individual-level estimates of the causal parameter (Athey and Wager, 2019).

Common characteristics of modern econometric causal inference:

1. Nonparametric (no assumptions on the shape of the DGP).
2. Able to yield a direct estimate of the group-level ATE.
3. Sometimes even an asymptotic theory.

Subgroup Analysis in Medicine (1/14)

Overview

Kent et al. (2018) distinguish between three types of subgroup analyses in medicine:

1. One-variable-at-a-time;
2. Risk modeling;
3. Effect modeling.

Each of these methods assumes that outcome Y_i is a binary outcome indicator, that is:

$$Y_i = \begin{cases} 1 & \text{if individual } i \text{ gets the outcome within the observation period,} \\ 0 & \text{otherwise.} \end{cases}$$

We therefore henceforth assume without loss of generality that Y_i is a binary mortality indicator.

Subgroup Analysis in Medicine (2/14)

Type 1: One-variable-at-a-time

Procedure:

1. Partition the samples in subgroups based on their realization of some scalar covariate X_i which is contained in $\mathbf{X}_i$.
2. Then, given a subgroup, calculate

$$RateRatio = \frac{(\#\text{Deaths for } W_i = 1)/(\#\text{Person-years for } W_i = 1)}{(\#\text{Deaths for } W_i = 0)/(\#\text{Person-years for } W_i = 0)}.$$

   (The number of person-years is the sum of the number of years each individual spent in the observation period.)
3. Since this is essentially a Poisson counting process, use a Poisson distribution to calculate confidence intervals of the rate ratio.

Subgroup Analysis in Medicine (3/14)

Type 1: One-variable-at-a-time

- Obviously, one-variable-at-a-time analyses are problematic from a causal inference point of view, but they are the the most often used way of reporting subgroup analyses in medicine.
- Some issues: risks for false-negative and false-positive results due to low power for statistical interactions, weak prior theory on potential effect modifiers, and multiplicity. For example, a lower mortality in a subgroup could be caused by the effect of a different variable.

➔ Can lead to *very* wrong policy recommendations (and already have...).

- The medical literature acknowledges that better methods are needed (Kent et al., 2018; Kent et al., 2020).

Subgroup Analysis in Medicine (4/14)

Overview of Predictive Modeling Approaches

- Motivated by the limitations of one-variable-at-a-time analyses, the medical literature suggests to apply predictive modeling for the task of subgroup analysis instead.
- It suggests to fit two separate models (recall that Y_i is binary):
  - One to estimate baseline risk $\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i)$;
  - One to estimate post-treatment risk $\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i, W_i = w_i)$.
- This is done to assess by how much the treatment intervention changes the mortality risk.
- Two main performance measures:
  - Calibration: Ability to match the estimated level of risk closely to the observed level of risk, across different risk-levels;
  - Discrimination: Ability of the model to distinguish between those with the outcome ($Y_i = 1$) and those without the outcome ($Y_i = 0$).
- Two main approaches: risk modeling and effect modeling.

Subgroup Analysis in Medicine (5/14)

Type 2: Risk Modeling

In risk modeling, we fit the two models, but use the estimated coefficients of the first model to fit the second.

1. Stage 1: Estimate $(p+1)$ coefficient vector $\boldsymbol{\beta}$ in the linear model

$$\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) = g\left(\boldsymbol{\beta}^\top \mathbf{x}_i^*\right),$$

   where $\mathbf{x}_i^* = (1, \mathbf{x}_i^\top)^\top$ and $g : \mathbb{R} \to (0,1)$ the logistic function.

2. With $\hat{\boldsymbol{\beta}}$, define the linear predictor $\hat{lp}_i = \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i^*$.

3. Stage 2: Estimate $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^\top$ in a separate linear model

$$\mathbb{P}\Big(Y_i = 1 \,\Big|\, \mathbf{X}_i = \mathbf{x}_i, W_i = w_i, \hat{lp}_i\Big) = g\Big(\hat{lp}_i + \gamma_0 + \gamma_1 w_i + \gamma_2 (w_i \cdot \hat{lp}_i)\Big).$$

4. With $\hat{\boldsymbol{\gamma}}$, define post-treatment risk
$\hat{\mathbb{P}}_i = g\Big(\hat{lp}_i + \hat{\gamma}_0 + \hat{\gamma}_1 w_i + \hat{\gamma}_2 (w_i \cdot \hat{lp}_i)\Big).$

5. Define the reversed treatment assignment W_i^{rev} by

$$W_i^{rev} = \begin{cases} 1 & \text{if } W_i = 0, \\ 0 & \text{if } W_i = 1. \end{cases}$$

Subgroup Analysis in Medicine (7/14)

Type 2: Risk Modeling

6. Use W_i^{rev} to calculate $\hat{\mathbb{P}}_i^{rev} = g\left(\hat{lp}_i + \hat{\gamma}_0 + \hat{\gamma}_1 w_i^{rev} + \hat{\gamma}_2 (w_i^{rev} \cdot \hat{lp}_i)\right)$, given $W_i^{rev} = w_i^{rev}$.
7. Calculate absolute and relative predicted benefits $\widehat{apb}_i$ and $\widehat{rpb}_i$, respectively, by

$$\widehat{apb}_i = \begin{cases} \hat{\mathbb{P}}_i^{rev} - \hat{\mathbb{P}}_i & \text{if } W_i = 1, \\ \hat{\mathbb{P}}_i - \hat{\mathbb{P}}_i^{rev} & \text{if } W_i = 0, \end{cases}$$

$$\widehat{rpb}_i = \begin{cases} \hat{\mathbb{P}}_i / \hat{\mathbb{P}}_i^{rev} & \text{if } W_i = 1, \\ \hat{\mathbb{P}}_i^{rev} / \hat{\mathbb{P}}_i & \text{if } W_i = 0. \end{cases}$$

➔ $\hat{\mathbb{P}}_i^{rev}$ intends to estimate the counterfactual post-treatment risk.

Subgroup Analysis in Medicine (8/14)

Type 3: Effect Modeling

Effect modeling is closely related to risk modeling with the difference that the two fitted models are completely independent of each other.

1. Baseline risk: Estimate $(p+1)$ coefficient vector $\boldsymbol{\beta}$ in the linear model

$$\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) = g\left(\boldsymbol{\beta}^\top \mathbf{x}_i^*\right),$$

where $\mathbf{x}_i^* = (1, \mathbf{x}_i^\top)^\top$ and $g : \mathbb{R} \to (0, 1)$ the logistic function.

Subgroup Analysis in Medicine (9/14)

Type 3: Effect Modeling

2. Specify $\mathcal{S} \subseteq \{1, \ldots, p\}$, which is a set of covariates that are suspected to be treatment effect modifiers/moderators.
3. Consider the linear model

$$\mathbb{P}\Big(Y_i = 1 \,\Big|\, \mathbf{X}_i = \mathbf{x}_i, W_i = w_i\Big) = g\left(\boldsymbol{\alpha}^\top \mathbf{x}_i^* + \sum_{\{j : j \in \mathcal{S}\}} \delta_j w_i x_{ij}\right)$$

   and gather its coefficients in the $(k+1)$ coefficient vector $\boldsymbol{\gamma} = (\boldsymbol{\alpha}^\top, \{\delta_j\}_{j\in\mathcal{S}}^\top)^\top = (\gamma_0, \gamma_1, \ldots, \gamma_k)^\top$, where $k = p + |\mathcal{S}|$. The δ_j are the coefficients of interaction terms $w_i x_{ij}$.
4. Fit $\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i, W_i = w_i)$ by means of penalized logistic regression (lasso here) and obtain $\hat{\boldsymbol{\gamma}}$.

Subgroup Analysis in Medicine (10/14)

Type 3: Effect Modeling

5. Let $\mathcal{U} = \{j : \hat{\gamma}_j \neq 0, j \neq 0\} \subseteq \{1, \ldots, k\}$ be the set of variables retained by the lasso, and $\mathcal{R} = \mathcal{U} \setminus \{j : j \in \mathcal{U}, j > p\}$ be the set of the retained variables that are *not* interaction effects.
6. Fit two separate ordinary logistic regression models:
   - An unrestricted model of Y_i on variable set $\mathcal{U}$,
   - A restricted model of Y_i on variable set $\mathcal{R}$.
7. Using the restricted and unrestricted model, perform a likelihood ratio test on joint significance of the coefficients of the interaction effects. If we reject joint significance, use the restricted model as final model. Otherwise, use the unrestricted model as final model.
8. Note: We use three separate random non-overlapping sets of samples for the three tasks of estimation, evaluation, and hypothesis testing, as suggested by Wasserman and Roeder (2009).

Subgroup Analysis in Medicine (11/14)

Type 3: Effect Modeling

9. Use the final effect model to obtain predictions of the post-treatment risk, $\hat{\mathbb{P}}_i = \hat{\mathbb{P}}(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i, W_i = w_i)$, of all samples.
10. Use reversed treatment assignment W_i^{rev} to obtain predictions of the counterfactual risk: $\hat{\mathbb{P}}_i^{rev} = \hat{\mathbb{P}}(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i, W_i = w_i^{rev})$.
11. Analogously to risk modeling, calculate absolute and relative predicted benefits $\widehat{apb}_i$ and $\widehat{rpb}_i$, respectively, by

$$\widehat{apb}_i = \begin{cases} \hat{\mathbb{P}}_i^{rev} - \hat{\mathbb{P}}_i & \text{if } W_i = 1, \\ \hat{\mathbb{P}}_i - \hat{\mathbb{P}}_i^{rev} & \text{if } W_i = 0, \end{cases}$$

$$\widehat{rpb}_i = \begin{cases} \hat{\mathbb{P}}_i / \hat{\mathbb{P}}_i^{rev} & \text{if } W_i = 1, \\ \hat{\mathbb{P}}_i^{rev} / \hat{\mathbb{P}}_i & \text{if } W_i = 0. \end{cases}$$

Subgroup Analysis in Medicine (12/14)

Processing the Predicted Benefits

Suppose we have obtained the quantities of absolute and relative predicted benefits $\widehat{apb}_i$ and $\widehat{rpb}_i$, respectively, from a predictive model. Thereupon, we're interested in evaluating heterogeneity in the treatment effectiveness.

1. Group the samples into q groups based on the quantiles of the baseline risk $\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i)$, which was estimated in the first modeling stage. Denote by $G_i \in \{1, \ldots, q\}$ the group membership.
2. For each group $G_i = g$, calculate the *observed benefit* $\widehat{ob}_g$:

$$\begin{aligned}\widehat{ob}_g =& |\{i : G_i = g, W_i = 1\}|^{-1} \sum_{\{i:G_i=g,W_i=1\}} Y_i - \\ & |\{i : G_i = g, W_i = 0\}|^{-1} \sum_{\{i:G_i=g,W_i=0\}} Y_i,\end{aligned}$$

which is essentially an estimate of the group-level ATE.

Subgroup Analysis in Medicine (13/14)

Processing the Predicted Benefits

3. Furthermore, for each group $G_i = g$, calculate the group-level *predicted benefit* by taking the within-group mean. For instance, for the absolute predicted benefits $\widehat{apb}_i$,

$$\widehat{apb}_g = |\{i : G_i = g\}|^{-1} \sum_{\{i:G_i=g\}} \widehat{apb}_i.$$

➔ Intends to average out the noise of the noisy individual estimates $\widehat{apb}_i$.

Recall now that a model is considered to be "good" if it is calibrated well, that is, if $\widehat{ob}_g \approx \widehat{apb}_g$ for all $g = 1, \ldots, q$. Assess by calibration plots in which the $\widehat{ob}_g$ and $\widehat{apb}_g$ are plotted against each other.

Subgroup Analysis in Medicine (14/14)

Comments on the Predictive Models

- Aside from the two general assumptions of unconfoundedness ($(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i | \mathbf{X}_i$) and sufficient overlap ($\mathbb{P}(W_i = 1 | \mathbf{X}_i = \mathbf{x}) \in (0, 1)$), the two types of predictive models rely on the linearity assumption on the DGP.

➔ Might be too restrictive for valid causal inference.

- The requirement *observed benefit* $\approx$ *predicted benefit* relies on meaningful grouping based on baseline risk $\mathbb{P}(Y_i = 1 | \mathbf{X}_i = \mathbf{x})$. But this is unknown and estimating it with linear models might fail to capture heterogeneity.
- No underlying statistical theory. However, they resemble to some extent *generic machine learning* (Chernozhukov et al., 2018), so maybe it is possible to derive one. . . .

Outline

A Simple Data Generating Process

In order to demonstrate how predictive models work, we apply them on $n = 10,000$ i.i.d. samples generated by the following simple process.

- $\mathbf{X}_i \sim \mathcal{N}_5(\mathbf{0}, \mathbf{I}), \varepsilon_i \sim \mathcal{N}(0, 0.5^2), \boldsymbol{\beta} = (0.2, 0.5, -0.3, 0.7, -0.1, 0.4)^\top$;
- Set $z_i = (1, \mathbf{x}_i^\top) \cdot \boldsymbol{\beta} + \varepsilon_i$ for $\mathbf{X}_i = \mathbf{x}_i$;
- Calculate $\pi_i(0) = \mathbb{P}(Y_i = 1 | \mathbf{X_i} = \mathbf{x}_i, W_i = 0) = \text{logistic}(z_i)$;
- Set $\pi_i(1) = \mathbb{P}(Y_i = 1 | \mathbf{X_i} = \mathbf{x}_i, W_i = 1) = 0.7 \cdot \pi_i(0)$ to model a constant relative treatment effect across all treated individuals;
- Draw $Y_i(W_i) \sim \text{Bernoulli}(\pi_i(W_i))$ for $W_i = 0, 1$;
- Draw $W_i \sim \text{Bernoulli}(0.5)$ and determine the observed outcome

$$Y_i = \begin{cases} Y_i(1) & \text{if } W_i = 1, \\ Y_i(0) & \text{if } W_i = 0. \end{cases}$$

➔ Y_i is a binary mortality indicator and treatment W_i reduces the relative mortality risk by 30%.

➔ $ATE = n^{-1} \sum_{i=1}^{n} \left(\pi_i(1) - \pi_i(0)\right) \approx -0.162$.

Performance of the Predictive Models

Performance on a representative generated dataset. Note that the models produce positive estimates for the predicted benefits to facilitate interpretation. Further recall $ATE \approx -0.162$ and W_i induces a relative mortality reduction of 30%.

The risk model:

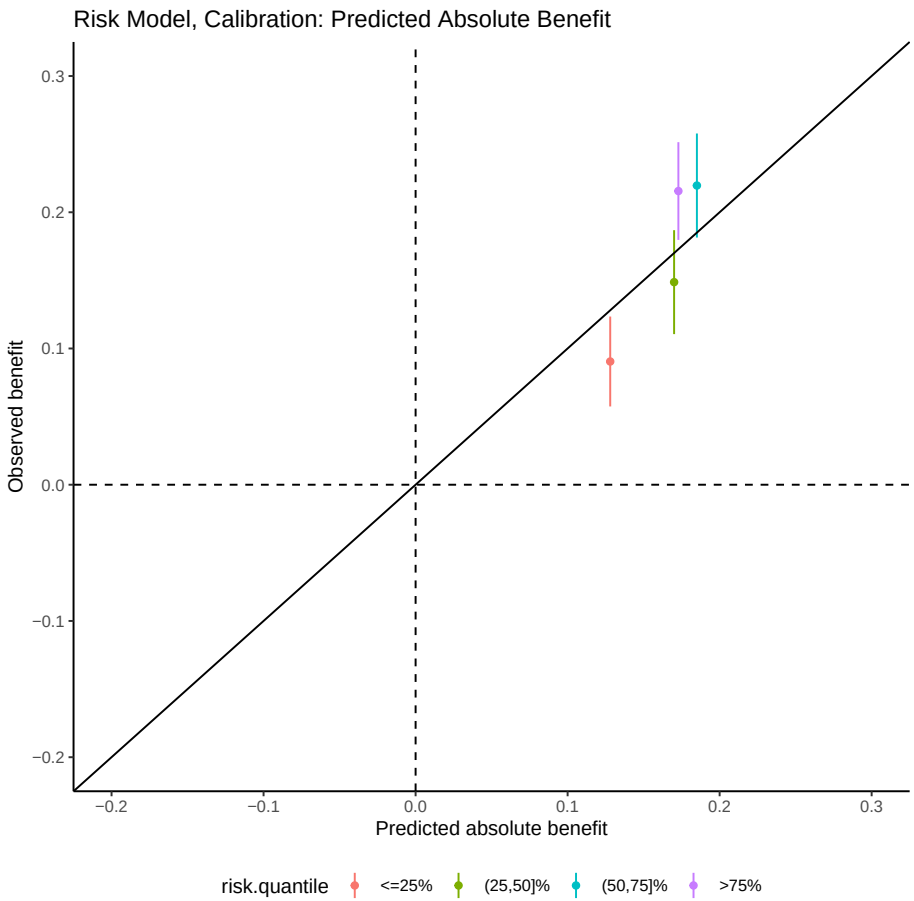
- $n^{-1} \sum_{i=1}^{n} \widehat{apb}_i = 0.164$,
- $1 - n^{-1} \sum_{i=1}^{n} \widehat{rpb}_i = 0.32$.

The effect model:

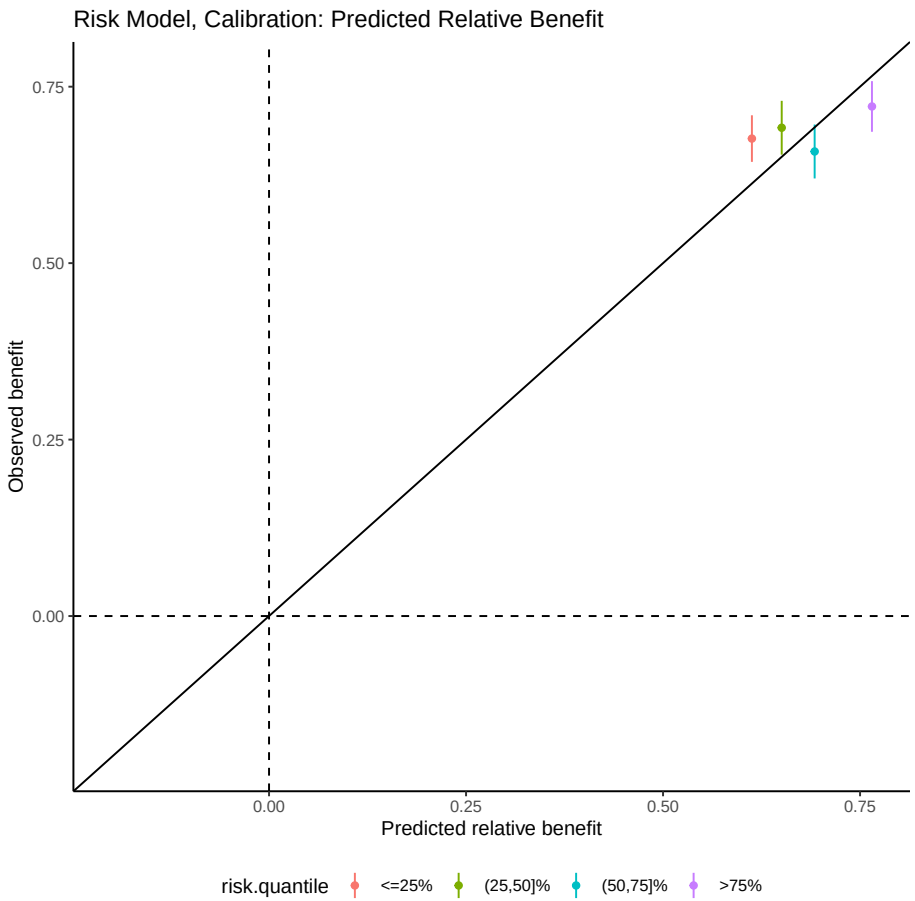
- $n^{-1} \sum_{i=1}^{n} \widehat{apb}_i = 0.162$,
- $1 - n^{-1} \sum_{i=1}^{n} \widehat{rpb}_i = 0.325$.

The estimated coefficients of both models furthermore correctly suggest that there are no interaction effects.

Calibration Plots of the Risk Model



(a) Group-level absolute benefit



(b) Group-level relative benefit

Left panel: group means are close to the 45° line ➔ Good calibration.

Right panel: Constant relative benefits ➔ No heterogeneity.

Calibration Plots of the Causal Forest

- A "causal" random forest (Athey et al., 2019) estimates the causal parameter of interest $\theta(\mathbf{x}_i)$, given $\mathbf{X}_i = \mathbf{x}_i$, for each individual $i = 1, \ldots, n$. This is a nonparametric method!
- Since Y_i is binary, $\theta(\mathbf{x}_i)$ is interpreted as treatment-induced reduction in mortality risk.

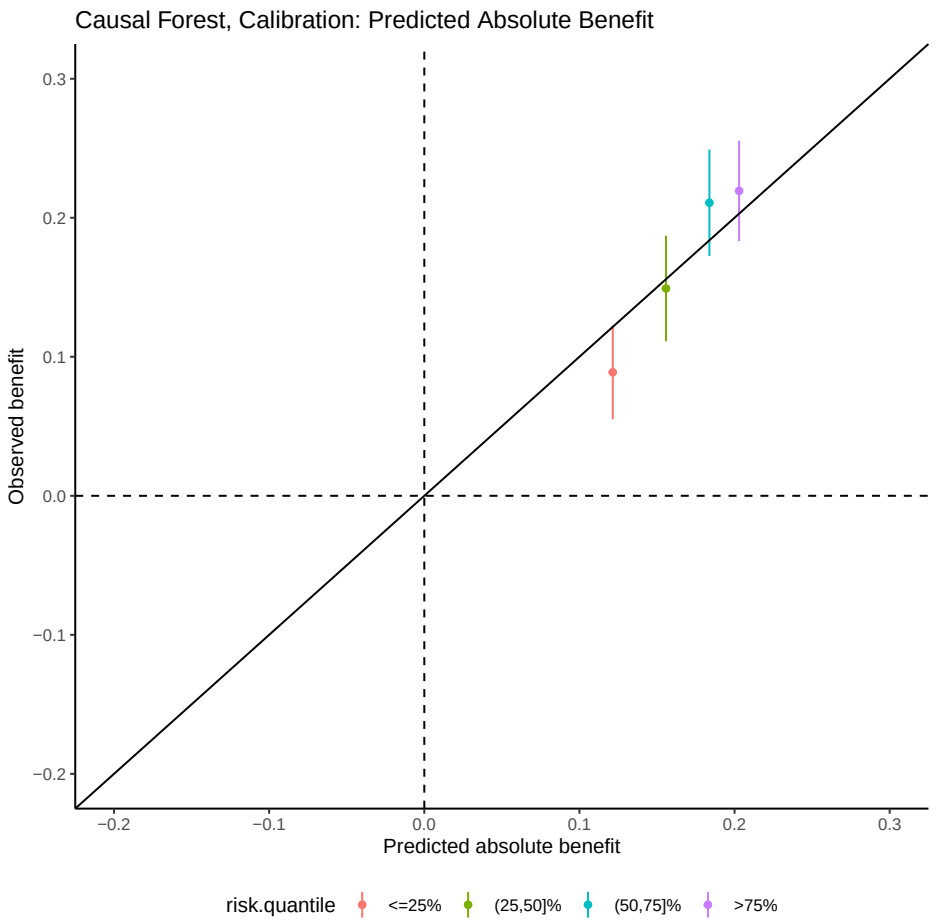
➔ This corresponds to the absolute predicted benefit!

- Unfortunately, the causal forest does not allow for the estimation of the relative predicted benefit.

➔ Evaluate heterogeneity by, for instance, plotting $\hat{\theta}(\mathbf{x}_i)$ against each covariate.

- Estimate baseline risk through ordinary regression forest of Y_i on $\mathbf{X}_i$ (nonparametric again).

Calibration Plot of the Causal Forest



Good calibration! Also, accurate estimation of ATE: estimated to be -0.166 and plotting the $\hat{\theta}(\mathbf{x}_i)$ against each covariate correctly does not indicate heterogeneity.

Calibration Plot of Generic Machine Learning

Working generic machine learning (Chernozhukov et al., 2018) into this framework is still work in progress! Same goes for survival models.

Outline

Conclusion

To-Do-List:

- Conceive a large-scale simulation with a nonlinear DGP. Preliminary results suggest that the predictive models break down, whereas the modern causal inferential methods do not.
- Investigate how the modern methods perform in relatively small sample sizes.
- Finalize an R package which provides flexible and user-friendly functions for the estimation and analysis of predictive models in medicine.
- Apply the predictive models on the data of the NELSON trial.

Main point: We want to demonstrate to clinicians that they should consider several different models before they engage in medical decision making (evidence-based medicine!). By doing so, they get a better understanding of the factors driving treatment effect heterogeneity and therefore become more certain that recommending an alternative treatment for a group might be the "right thing" to do.

Thanks for you attention! Any questions?

References

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized Random Forests. *The Annals of Statistics*, 47(2):1148–1178.

Athey, S. and Wager, S. (2019). Estimating Treatment Effects with Causal Forests: An Application. *Observational Studies*, 5:36–51.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2018). Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments. Technical Report 24678, National Bureau of Economic Research, Cambridge, Massachusetts.

Kent, D. M., Steyerberg, E., and van Klaveren, D. (2018). Personalized Evidence Based Medicine: Predictive Approaches to Heterogeneous Treatment Effects. *The British Medical Journal*, 363.

Kent, D. M., van Klaveren, D., Paulus, J. K., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P., Patrick-Lake, B., Morton, S., Pencina, M., et al. (2020). The Predictive Approaches to Treatment Effect Heterogeneity (PATH) Statement: Explanation and Elaboration. *Annals of Internal Medicine*, 172(1):W1–W25.

de Koning, H. J., van der Aalst, C. M., de Jong, P. A., Scholten, E. T., Nackaerts, K., Heuvelmans, M. A., Lammers, J.-W. J., Weenink, C., Yousaf-Khan, U., Horeweg, N., et al. (2020). Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *New England Journal of Medicine*, 382(6):503–513.

References (cont.)

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence-Based Medicine: What it is and What it isn't. *The British Medical Journal*, 312(7023):71–72.

Shaughnessy, J. J., Zechmeister, E. B., and Zechmeister, J. S. (2000). *Research Methods in Psychology*. McGraw-Hill.

Wasserman, L. and Roeder, K. (2009). High Dimensional Variable Selection. *The Annals of Statistics*, 37(5A):2178–2201.