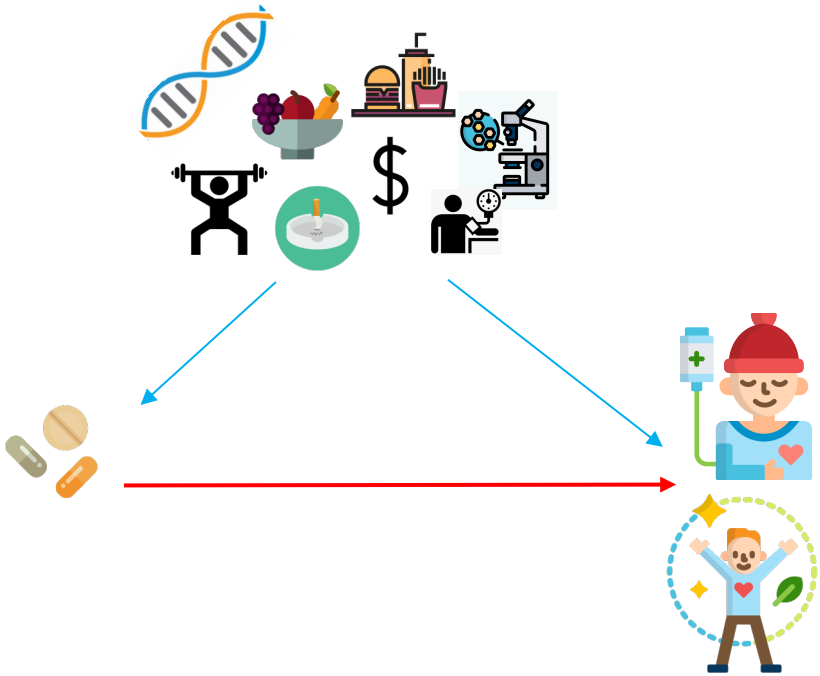
The Alan Turing Institute

Going beyond the average: causal machine learning for treatment effect heterogeneity estimation

Karla Diaz-Ordaz @karlado

Real-world evidence

- *Which treatment works?*
  - *For whom?*
- Using observational data: **confounding**
- Estimate:
  - Average causal effect (ATE)
  - *Effect-modifiers?*
    - Conditional average treatment effect (CATE)



Example: Causal effects of cancer treatment

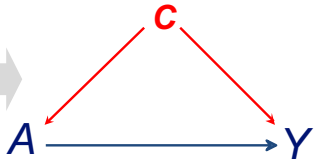
Potential outcomes

Y^1 *what-if* treated
Y^0 *what-if* not treated

Causal estimand
(involving counterfactuals)

$$ATE = E[Y^1 - Y^0],$$
or
$$CATE(x) = E[Y^1 - Y^0 | X = x],$$
for baseline (vector) X
e.g. X baseline Cancer Stage, ECOG

Identifying assumptions



- No interference
- Consistency
- Conditional exchangeability given C

Statistical estimand
(function of the observed data)

$$ATE = E[Y|A=1,C] - E[Y|A=0,C],$$

and/or

$$CATE(x) = E[Y|A=1,C,X=x] - E[Y|A=0,C,X=x]$$

- Data : 2700 lung cancer patients, 130 clinical variables (from EHRs) and 500 genomic variables
  A = Treatment : a class of cancer treatment vs all other oncology treatments (as first line)
  Y = outcome = Alive at t months since treatment initiation
  C = confounders = All baseline variables are potential controls for confounding: demographics, clinical, labs, genomics
- $t \in \{0,1,2,\dots,24\}$ months, survival time > t and not censored
- Censoring dealt with by up-weighting those observed (IPCW)

Estimation

1. Outcome regression: example Partially linear model

$$E[Y|A,C] = \beta A + m(C)$$

   – β is the parameter of interest, i.e. estimates the ATE
   – Assumes model is correctly specified: linear in A and correct function $m(C)$
   – Can be checked from data: problematic in **high-dimensions**
   – Overfitting and extrapolation

Estimation

2. Propensity scores: we model $p(C)=Pr[A \mid C]$
   - Useful in high-dim: if conditional exchangeability holds given C, then it also holds given $p(C)$.
   - Assumes positivity: $0 < Pr[A \mid C] < 1$,
   - $p(C)$ **must be correctly specified**
   - Misspecification: difficult to diagnose especially

   with poor overlap and very-high-dimensional settings

Using machine learning to attenuate model misspecification

- Direct use of machine learning estimates of either PS or outcome models leads to **regularisation bias**
- This bias is the result of **over-smoothing**
- and of choosing **overly sparse models**: mistakenly throwing out important confounders
- No valid way of calculating uncertainty; bootstrap is not valid with ML (except for a few algorithms)

Plug-in bias in machine learning

Suppose we use machine learning to estimate $m(C)$ and estimate β by

$$\hat{\beta} = \left\{\frac{1}{n}\sum A_i^2\right\}^{-1} \frac{1}{n}\sum A_i\{Y_i - \widehat{m}(C_i)\};$$

- further decomposition shows that the regularisation bias term equals (to 1st order)

$$E[A_i^2]^{-1}\,\frac{1}{n}\sum p(C_i)\{m(C_i) - \widehat{m}(C_i)\}$$

  with $p(C_i) = E[A_i|C_i]$ the true propensity score.
- The regularisation bias term is the sum n terms that do not have mean zero, divided by $\sqrt{n}$
- So why not "center" A , so that the new "regressor" $A_i - p(C_i)$ has mean zero ?
  - Estimate β by $\hat{\beta} = \left\{\frac{1}{n}\sum\{A_i - \hat{p}\,(C_i)\}^2\right\}^{-1} \frac{1}{n}\sum\{A_i - \hat{p}\,(C_i)\}\,\{Y_i - \widehat{m}(C_i)\}$

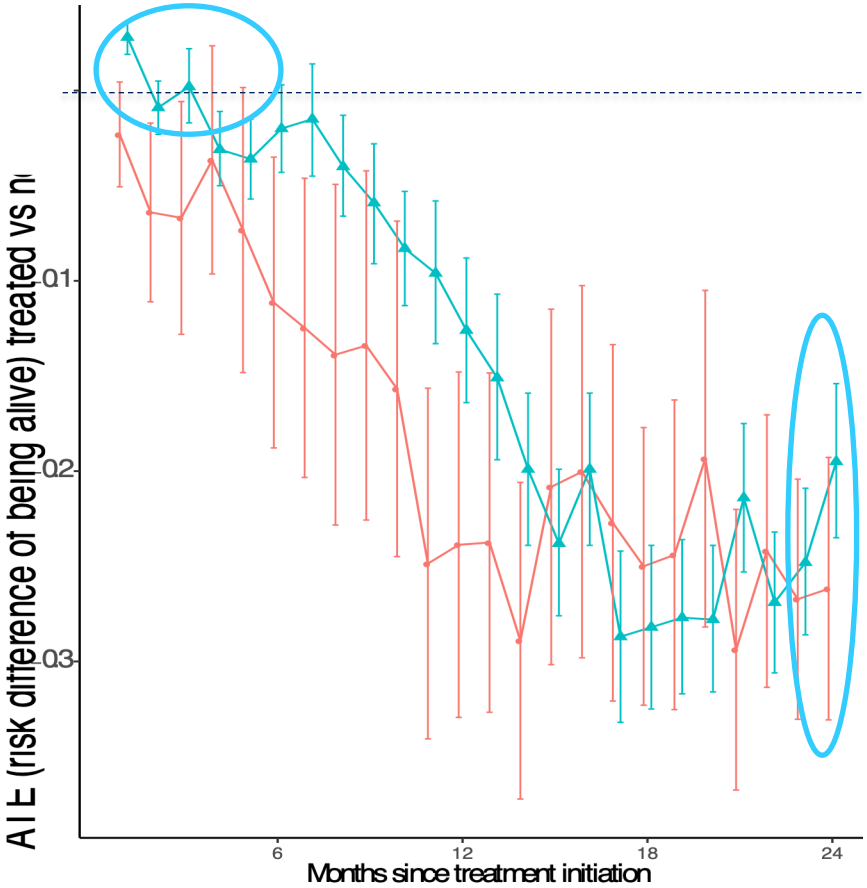
Doubly robust estimators to the rescue

- This is the “usual” doubly robust estimator for the ATE
- AIPTW uses a model for the outcome and a model for the treatment [1]
- the regularisation bias term now depends on the product of the estimation errors in $m(C_i)$ and $p(C_i)$
- Consistent if at least one model is correctly specified
- ML algorithms used to estimate $m(C_i)$ and $p(C_i)$ need to converge sufficiently fast (product of order $\sqrt{n}^{-1}$)

A bit of history

- Mathematical statisticians had studied “plug-in estimators”: obtained plugging in non-parametric estimators into statistical functionals [2,3,4,5]
- van der Laan extended this theory to construct plug-in estimators based on machine learning: “Targeted Maximum Likelihood Estimators” [6, 7]
- Chernozhukov [8] extended this theory to work under weaker conditions by invoking sample splitting leading to “Double machine learning”

Results for ATE



estimated via:

- DML that uses the usual DR estimator paired with Random Forests (RF) estimation and sample splitting (AIPTW_RF)

- TMLE which used ensemble learning (Super Learner using GML, RF, boosting) no sample splitting

Method

AIPTW_RF

TMLE

Treatment effect heterogeneity

- Perhaps the ATE is masking effects in certain populations
- We can estimate conditional ATE (CATE) given a predetermined (vector) X:

$$\tau(x) = E\left[Y_i^1 - Y_i^0 | X_i = x\right]$$

- Recent interest in using machine learning to **find drivers** of treatment effect heterogeneity
- **Problem: how to do valid confidence bands for CATE after data-driven selection of effect modifiers?**

Heterogeneous treatment effects

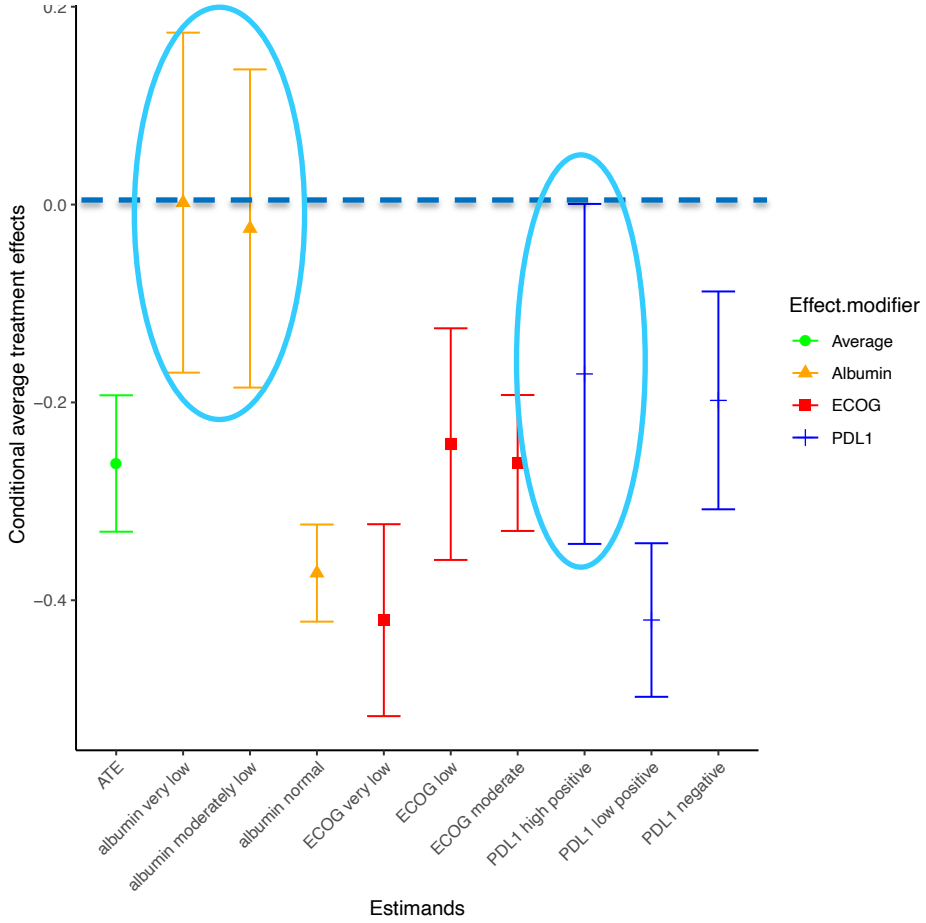
- We cannot estimate all heterogeneous effects
- But can answer: is there heterogeneity?
- What are the characteristics of those with the largest treatment effect?
- use sample splitting to obtain valid confidence bands

Causal forests [9,10]

- Train Trees to **maximise heterogeneity** in treatment effect as opposed to RMSE in outcome prediction
- **"Honesty":** in a tree, i used to select splits or estimate $\tau(x)$
- Forests are formed using weighted aggregation, weights chosen to minimise bias in $\tau(x)$
- **general test** of heterogeneity implemented
- for observational data:
  - grow outcome and PS forests
  - use these predictions to "center" both, so they are mean zero (DR framework)
  - apply CF to these "residualised" variables

Results at 24 months

- Evidence of treatment heterogeneity
- negative ATE
- some subgroups CATE is no different from 0
- The picture is unclear



Conclusions

- TMLE and DML approaches gave similar results for ATE
- CF detects heterogeneity in treatment effects, evidence only at 24m
- Interpret in light of small numbers
- Effect-heterogeneity questions require a more “homogeneous” definition of treatment
- Next steps: time-updated covariates may explain better the treatment effect heterogeneity. Need to consider treatment intensification
- Potential for remaining unobserved confounding
    - DML developed for Instrumental Variable models
        - ATE [8],
        - for CATE via g-estimator [11]

References

1. Robins and Rotnitzky (2001) “Comment on the Bickel and Kwon article, 'Inference for semiparametric models: Some questions and an answer’”, Statistica Sinica 11(4):920-936 Bickel, et al (1998). Efficient and Adaptive Estimation for Semiparametric Models. Springer, New York.
2. Newey, (1990) “Semiparametric efficiency bounds’, Journal of Applied Econometrics 5, 99–136.
3. Robins and Rotnitzky, (1995). “Semiparametric efficiency in multivariate regression models with missing data”. Journal of the American Statistical Association, 90(429), 122-129.
4. Van der Vaart, A. W. (1991): “On Differentiable Functionals,” Annals of Statistics 19, 178–204.
5. Rubin, D B., and van der Laan MJ. (2008) "Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis." The International Journal of Biostatistics 4.1.
6. van der Laan, M.J. and Rose, S., 2011. Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media.
7. Chernozhukov et al (2018) “Double/debiased machine learning for treatment and structural parameters”. The Econometrics Journal, Volume 21, Issue 1, 1 February 2018, C1–C68
8. Wager, S, and Athey S (2018). "Estimation and inference of heterogeneous treatment effects using random forests." Journal of the American Statistical Association 113.523 : 1228-1242.
9. Athey S, Tibshirani J, Wager S. (2019) “Generalized random forests”. The Annals of Statistics. 47(2):1148-78
10. DiazOrdaz K, Daniel Rh, Kreif, N “Data-adaptive doubly robust instrumental variable methods for treatment effect heterogeneity” Pre-print on arXiv https://arxiv.org/abs/1802.02821

Extra slides

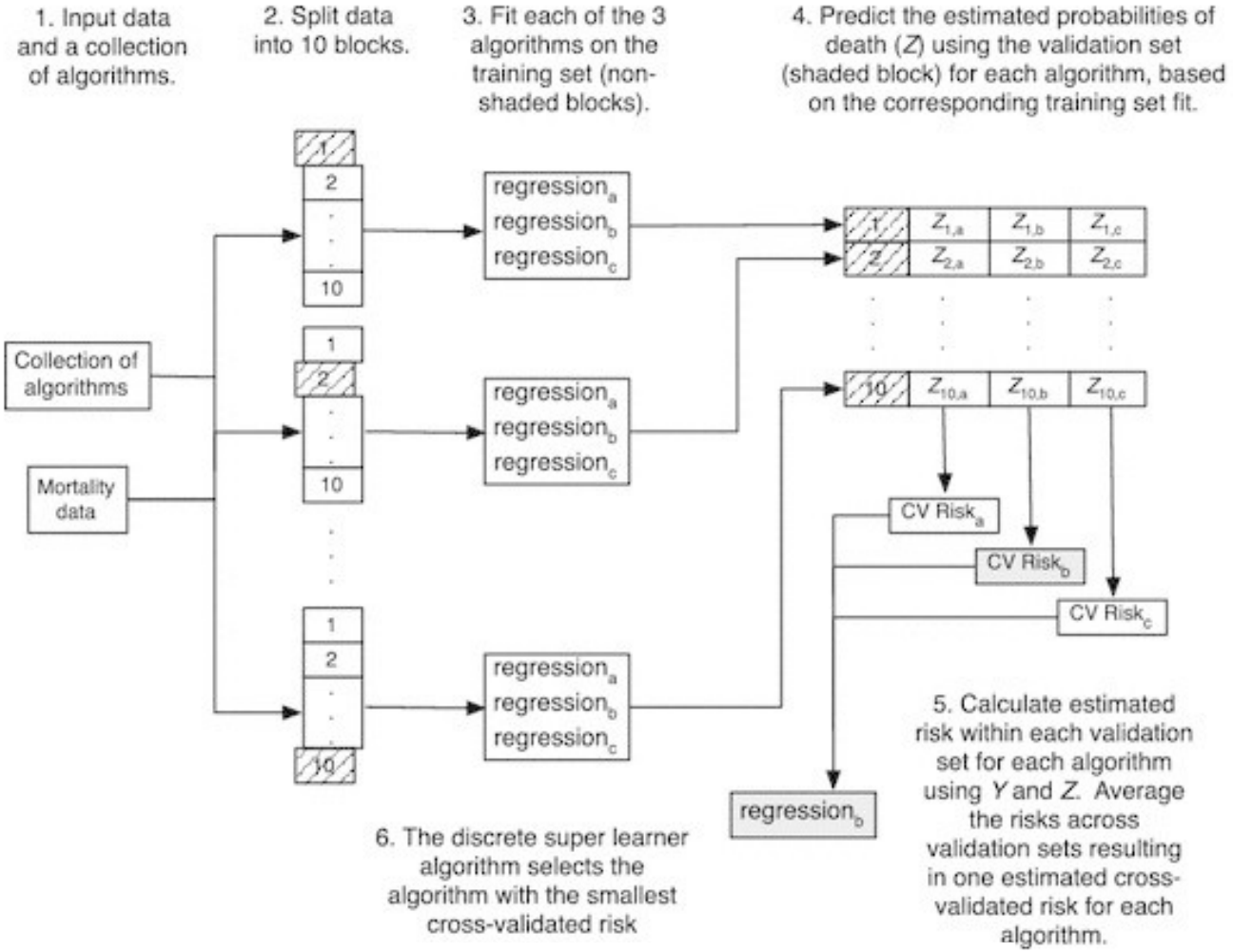
Data

- Note: all patients received active treatment
- Index date = treatment start
- Extracted baseline variables recorded prior to the index

Data

- Genomics: number of mutations in the gene and whether it was:
  - Short variants
  - Copy number variation (amplification/deletion)
  - Rearrangements
- Missing baseline values : single non-parametric imputation with Random Forest (*R Package missForest*) + missingness indicator

SuperLearner



Algorithms used in SL library in the example

Algorithm	Description
glm	GLM with logit link for PS and outcome models
gam	Generalised additive models with 2, 3, 4, and 5 knots.
Lasso	*budget c* s.t. abs. value of the regression coefficients adds up to c
tree methods	
Random forests	bootsrapped samples of the data, growing a tree using only a (random) subset of covariates for the splits Average all (reduction in variance).
boosted CART	sequence where previous tree's residuals are outcomes for next tree then add them all together
BART	Bayesian additive regression trees
Super learner	...