Using causal machine learning to explore heterogeneous responses to policies

Noemi Kreif, PhD
Research Fellow
Centre for Health Economics
University of York, UK

RSS Interpretable Machine Learning & Causal Inference Workshop
15/12/2020

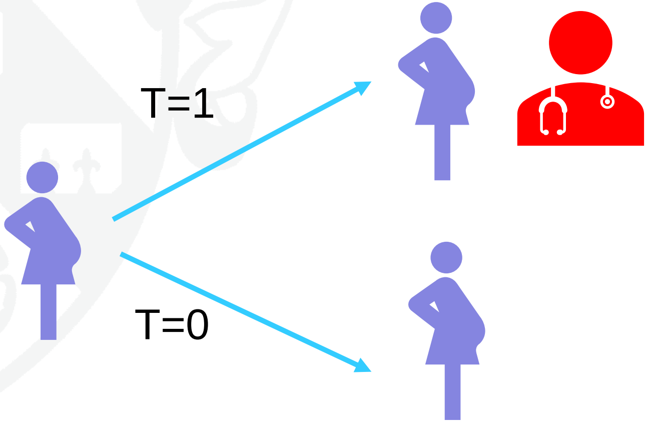
Outline

Objective: to demonstrate how causal machine learning can support research in health policy evaluation

- Target: estimating heterogeneous policy effects
- Method: “Causal Forests” (Athey et al. 2019)
- Application: evaluation of the impact of public health insurance on maternal health care utilisation in Indonesia

Motivation

- Most questions in the health and social sciences are of causal nature
  - Did a new a cancer drug improve survival of patients?
  - Did introducing sugar tax reduce obesity?
  - Did introducing universal health insurance improve access to health care ?



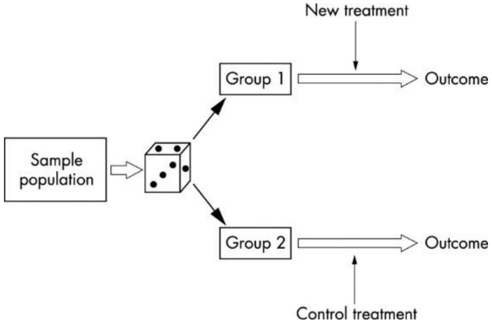
Ideally want to compare outcome in two worlds, one of which is counterfactual

“fundamental problem of causal inference”

Motivation

- How we tend to address the fundamental problem of causal inference?

  – **Randomise!**



The 2019 Nobel Memorial Prize in Economics Sciences was awarded to Abhijit Banerjee, Esther Duflo, and Michael Kremer "for their experimental approach to alleviating global poverty."

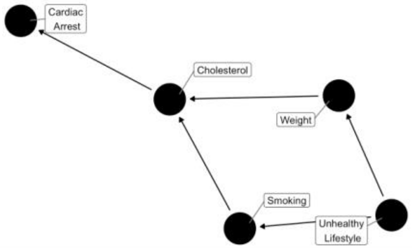
The research questions

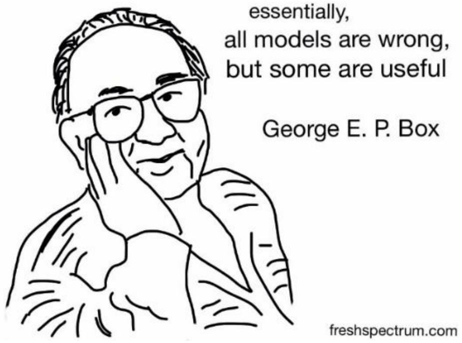
- How we tend to address the fundamental problem of causal inference in observational studies?

  – **Make (untestable) assumptions!**
    - Using external knowledge, theory

  +

  – **Fit statistical models**
    - E.g. to adjust for differences between treated and control populations





Motivation

- Because of these challenges, policy evaluations often stop at *average* effects

- Policy maker needs information on heterogeneity in the treatment effects, to answer question such as
  - Did the policy work for a given group?
  - Who were the (relative) winners and losers?
  - How could the design of future programmes be improved?

- Pre-specified subgroup analysis restrictive…
  - Non-randomised evaluations rarely pre-specified -> "cherry picking"
  - Can use the data to learn about what drives differential responses to a policy
  - Requires flexible approaches -> Machine learning can help?
  - Recently a very active area of methodological research in causal inference (vanDerWeele et al. 2019, Kunzel et al. 2019, Athey, Wager et al 2019, etc…)

Case study: the heterogeneous impacts of health insurance

Gradual expansion of Health insurance in Indonesia

- **Contributory** heath insurance since the 1970s
- **Subsidised health insurance** for the poor since the 1990a
- 20% of population still uninsured

Jaminan Kesehatan Nasional

Questions:

1) Does health insurance improve access to health services on average?
2) Which type of health insurance worked better
3) How do these impacts vary among populations subgroups?
   - poor versus rich
   - high versus low educated
   - rural versus urban
   - Other dimensions?

- Data: Survey of ~10,000 births: health insurance (treatment), and skilled birth attendance (outcome) information, ~50 covariates

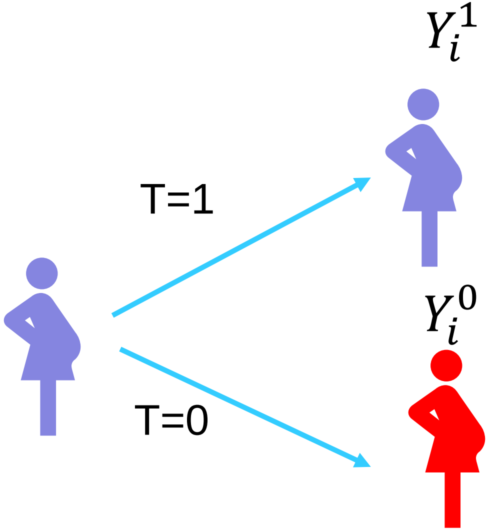
Methods: potential outcomes and causal estimands

Potential outcomes

$$Y^1 = \mathrm{Y}$$
$$Y^0 = ?$$

Giving birth assisted by a professional

- if insured
- uninsured



Individual level causal effect

$$Y_i^1 - Y_i^0$$

Methods: potential outcomes and causal estimands

Potential outcomes

$Y^1 = \mathrm{Y}$
$Y^0 = ?$

Causal estimand

(involves counterfactuals)

e.g. ATE
$E[Y^1 \text{-} Y^0]$

Average treatment effect

- Average benefit from everyone having insurance vs. no one having it

Methods: potential outcomes and causal estimands

Potential outcomes

$Y^1 = \mathrm{Y}$
$Y^0 = ?$

Causal estimand

(involves counterfactuals)

e.g. ATT
$E[Y^1\text{-}Y^0|W = 1]$

Average treatment effect among the treated (ATT)

- How much those who had health insurance have benefitted?

Methods: potential outcomes and causal estimands

Potential outcomes

$Y^1 = \mathrm{Y}$
$Y^0 = ?$

Causal estimand

(involves counterfactuals)

e.g. ATC
$E[Y^1\text{-}Y^0|W = 0]$

Average treatment effect among the controls (ATC)

- How much those who did not have health insurance would have benefitted from having insurance?

Methods: potential outcomes and causal estimands

Potential outcomes

$Y^1 = \mathrm{Y}$
$Y^0 = ?$

Causal estimand

e.g. CATE(X)

Conditional average treatment effect (CATE) function:

$$\tau(x) = \mathrm{E}[\mathrm{Y}_i(1) - \mathrm{Y}_i(0) | X_i = \mathrm{x}]$$

Methods: potential outcomes and causal estimands

Potential outcomes

$Y^1 = \mathrm{Y}$
$Y^0 = ?$

Causal estimand

e.g. CATE(X)

Conditional average treatment effect (CATE) function:

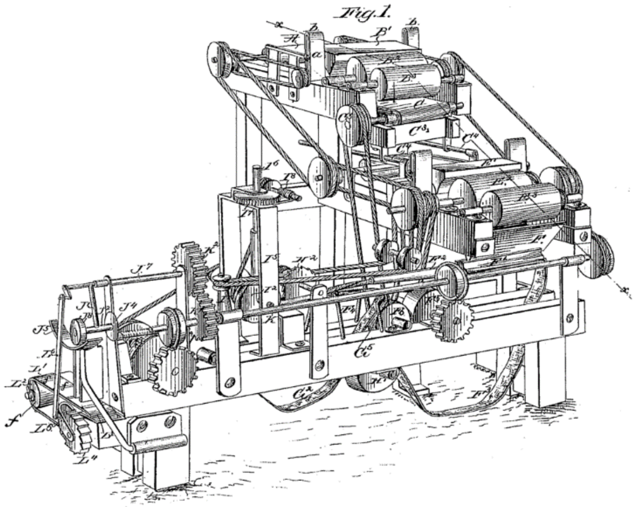
$$\tau(x) = \mathrm{E}[\mathrm{Y}_i(1) - \mathrm{Y}_i(0) | X_i = \mathrm{x}]$$

- e.g. Pre-specified subgroups of interest: wealth (quintiles), education, rural status
- High dimensional if many (multi-valued, continuous) Xs -> challenge

The CATE estimand

Woman's characteristics

- Age
- Wealth
- Education
- Region
- Birth order
- Etc.



Predicted, individual specific gain from having health insurance

Methods: potential outcomes and causal estimands

Potential outcomes

$Y^1 = \mathrm{Y}$
$Y^0 = ?$

Causal estimand

e.g. CATE(X)

Identifying assumptions

- No unmeasured confounders

$$Y^1, Y^0 \perp W \mid X$$

X: demographic, socioeconomic variables, availability of health services in community, birth year and province indicators

- Overlap (no characteristics perfectly predict insurance status)

Methods: potential outcomes and causal estimands

Potential outcomes

$Y^1 = \text{Y}$
$Y^0 = ?$

→

Causal estimand

e.g. CATE(X)

→

Identifying assumptions

→

Estimators

(functions of observed data)

- Many estimators of average treatment effects aim to adjust for x covariates
  - Regression, propensity score methods, double-robust methods
  - Machine learning has been playing an increasing role in the construction of estimators of treatment effects

“Causal Machine learning” combines key strengths of the two fields

	Machine learning for prediction	Causal inference
Can we observe the “ground truth”?	Yes	No (“fundamental problem of causal inference) -assumptions
Inference (standard errors)	Not a priority	Priority/well developed
Model selection	Transparent Data adaptive	Based on “theory” (?) Can be subjective

Inspiration: Athey S. The impact of machine learning on economics. 2018

Causal Machine learning

(1) ML for variable selection for confounding adjustment (e.g. double-lasso Belloni et al. 2014)

(2) ML to estimate "nuisance parameters" (propensity scores, regression functions)

- targeted learning (van der Laan and Rose, 2011), double/debiased machine learning (Chernozhukov et al, 2018)

(3) Modify loss function ML algorithms to minimise bias in causal parameters of interest

- E.g. Causal Forests (Athey et al. 2019), R-learning (Nie and Wager, 2017)

Causal Forest to estimate CATEs

(Nie and Wager 2017, Athey et al. 2019)

Motivation: partially linear model

$Y_i = f(X_i) + W_i\tau + \varepsilon_i$ for now assume τ homogenous

residualise Y_i and W_i

$W_i^{res} = W_i - \mathrm{p}(X_i)$ where $\mathrm{p}(X_i) = \mathrm{E}[W_i | X_i]$ (the propensity score)

$Y_i^{res} = Y_i - \mathrm{m}(X_i)$ where $\mathrm{m}(X_i) = \mathrm{E}[Y_i | X_i]$

- Nuisance parameters $\mathrm{p}(X_i)$ and $\mathrm{m}(X_i)$ estimated by machine learning

Causal Forest to estimate CATEs
(Nie and Wager 2017, Athey et al. 2019)

τ can be estimated from the simple linear regression

$Y_i^{res} = \tau\, W_i^{res} + \varepsilon_i$ -> $\hat{\tau} = \frac{\sum\{W_i - E[W_i|X_i]\}\{Y_i - E[Y_i|X_i]\}}{\sum\{W_i - E[W_i|X_i]\}^2}$

- Consistent, asymptotically linear
- Cross-fitting allows for the use of a wide range of ML algorithms

Double/debiased machine learning estimator described in Chernozhukov et al. 2018

Causal Forest to estimate CATEs

(Nie and Wager 2017, Athey et al. 2019)

Extension of the partially linear model:

$Y_i = f(X_i) + W_i\tau(X) + \varepsilon_i$ $\tau(X)$ heterogenous

τ can be estimated from the simple linear regression in a small neighbourhood $N(X)$

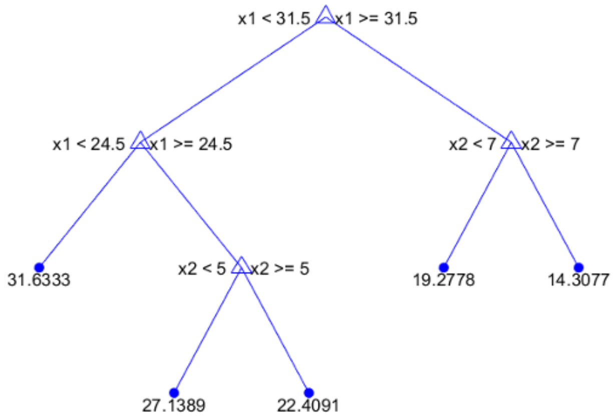
$Y_i^{res} = \tau(X)\, W_i^{res} + \varepsilon_i$ -> $\widehat{\tau(X)} = \frac{\sum\{W_i - E[W_i|X_i]\}\{Y_i - E[Y_i|X_i]\}}{\sum\{W_i - E[W_i|X_i]\}^2}$

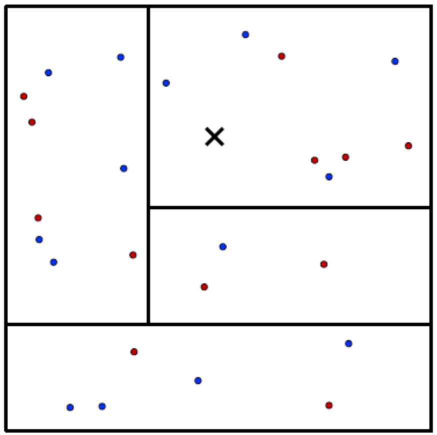
sums over $x \in N(x)$

How to choose N(X)?

Using an approach based on random forests -> **Causal Forest**

Random forests for prediction (Breiman 2001)

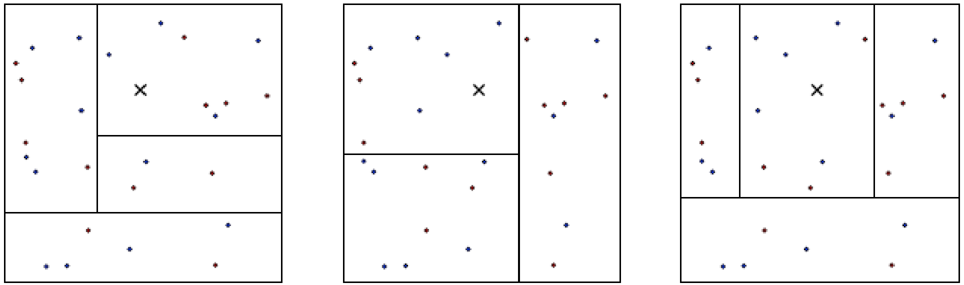




Regression tree predicts the outcome of observation with X covariates based on average outcomes in a “leaf” of a tree, with similar Xes

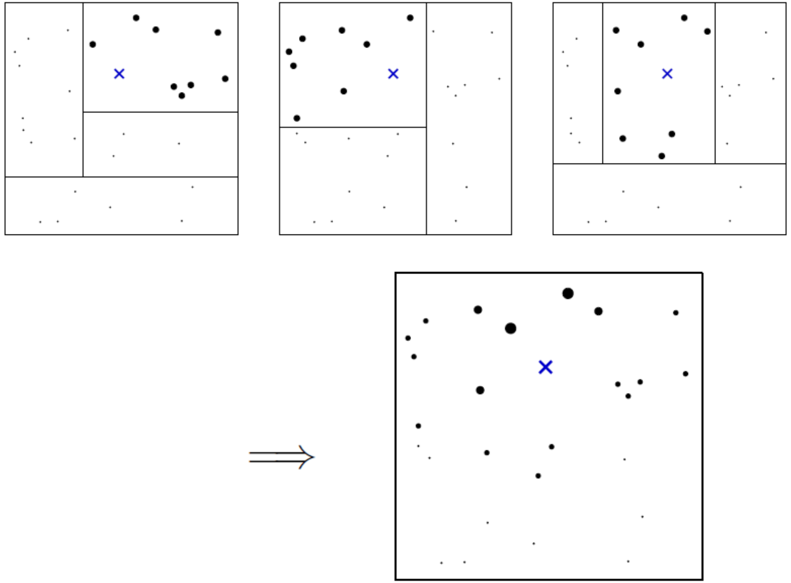
Tree structure (partitions) selected to minimise root mean squared prediction error (RMSE), in a new sample

Random forests for prediction (Breiman 2001)



- To improve estimation performance, many trees built, on subsamples of the data and subsets of the covariates

Random forests for prediction (Breiman 2001)



Combine trees into a forest:

“Neighbouring observations” get different weights in the final predictions, based on the frequency they have been selected to be on the same leaf as X

Causal Forests for CATEs

(Wager and Athey 2018, Athey et al. 2019)

- Causal Forests modify the splitting criterion of random forest to maximise the treatment effect heterogeneity as opposed to minimising prediction RMSE
- "Causal Tree"
  - Treatment effects estimated on a partitions of the data (lm yres ~ wres)
  - Choose splits to maximise differences between estimated τ
- Do this many times -> Causal Forest
  - Save weights $\alpha_i(X)$: how often observation i was used to estimate treatment effect at X

Causal Forests for CATEs

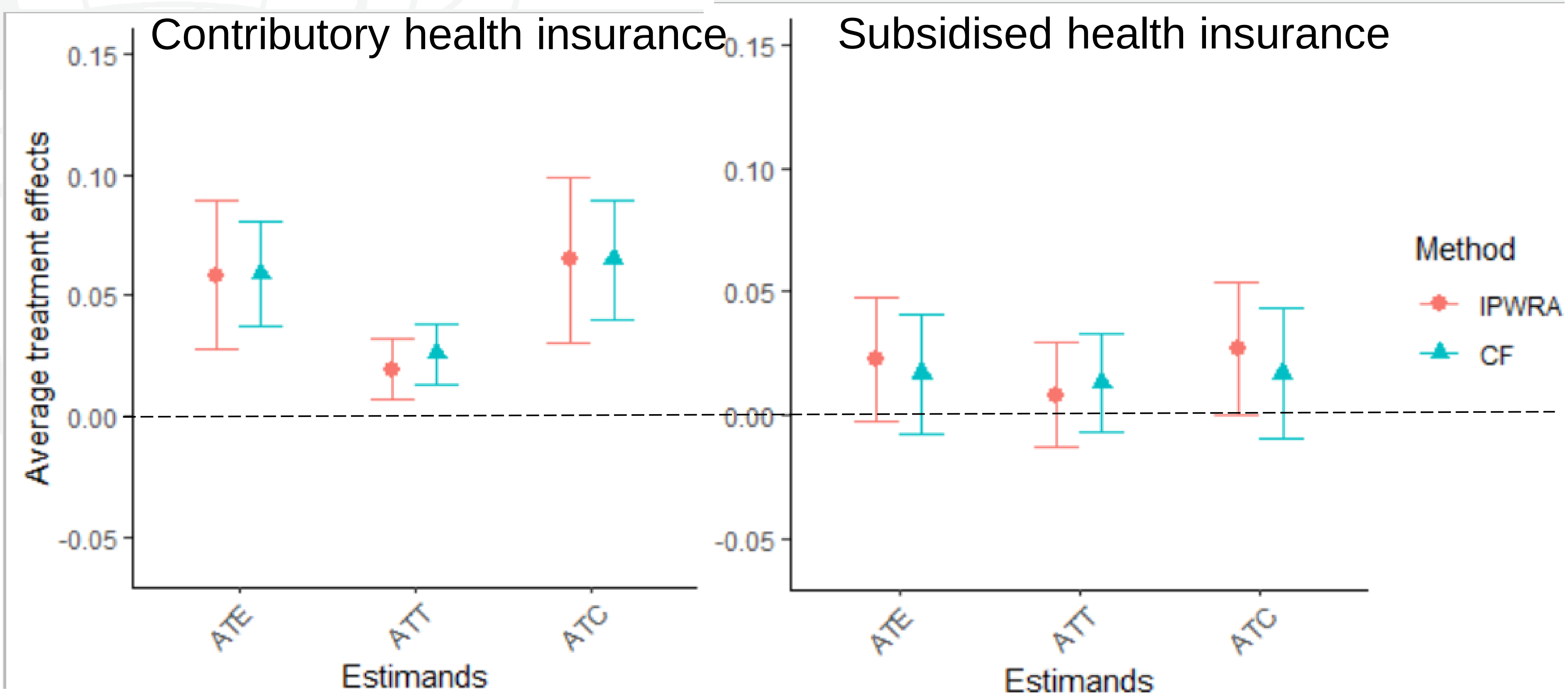
(Wager and Athey 2018, Athey et al. 2019)

- Weights "plugged in" the residual on residual regression, resulting in

$$\widehat{\tau(X)} = \frac{\sum \alpha_i(X)\{W_i - E[W_i|X_i]\}\{Y_i - E[Y_i|X_i]\}}{\sum \alpha_i(X)\{W_i - E[W_i|X_i]\}^2}$$

- Asymptotic normality of estimator, inference based on resampling from forests

Average treatment effects: traditional and ML methods give similar results



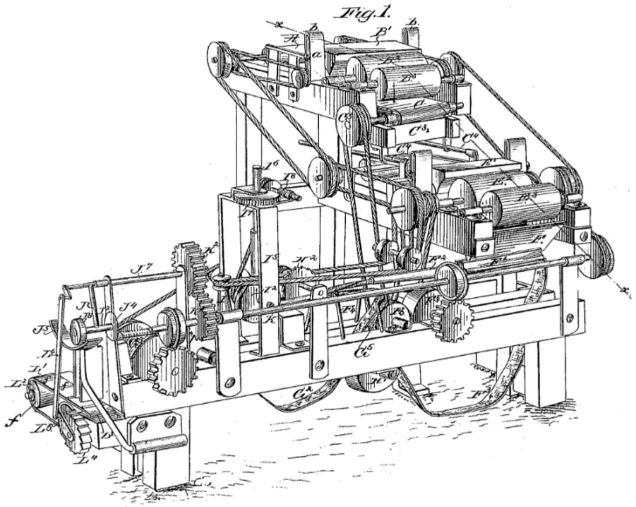
Results: variable importance from the Causal Forests

	Subsidised HI		Contributory HI	
Ranking	Variable importance measure	Variable	Variable importance	Variable
1	0.126	Birth order >=3	0.127	Province East Java
2	0.085	Birth year 2012	0.123	Higher education
3	0.084	Age >=31	0.083	Wealth quantile 4
4	0.075	Past covariates imputed	0.069	Province South Kalimantan
5	0.066	Cash transfer	0.066	Rural community
6	0.065	Poor card	0.060	Wealth quantile 5
7	0.063	Birth year 2014	0.055	Province West Sumatra
8	0.062	Birth order =2	0.049	Private practice in community
9	0.054	Province West Nusa Tenggara	0.048	Senior education
10	0.046	Natural disaster	0.045	Province Banten

The CATE estimand

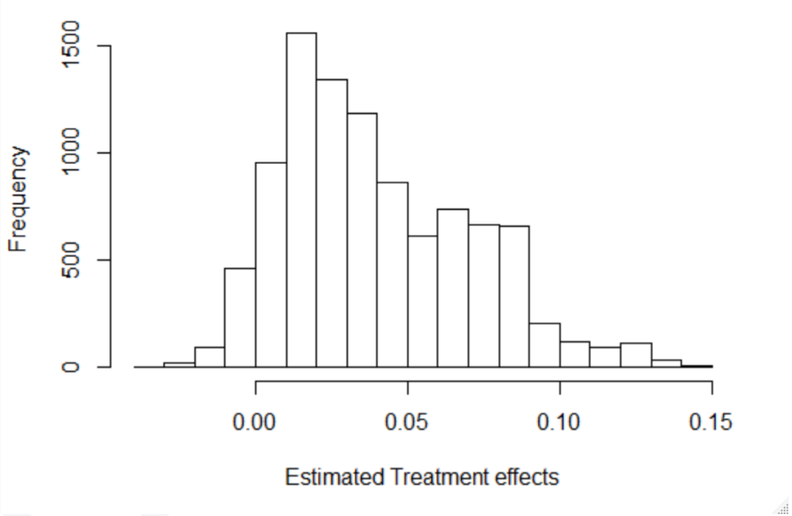
Woman's characteristics

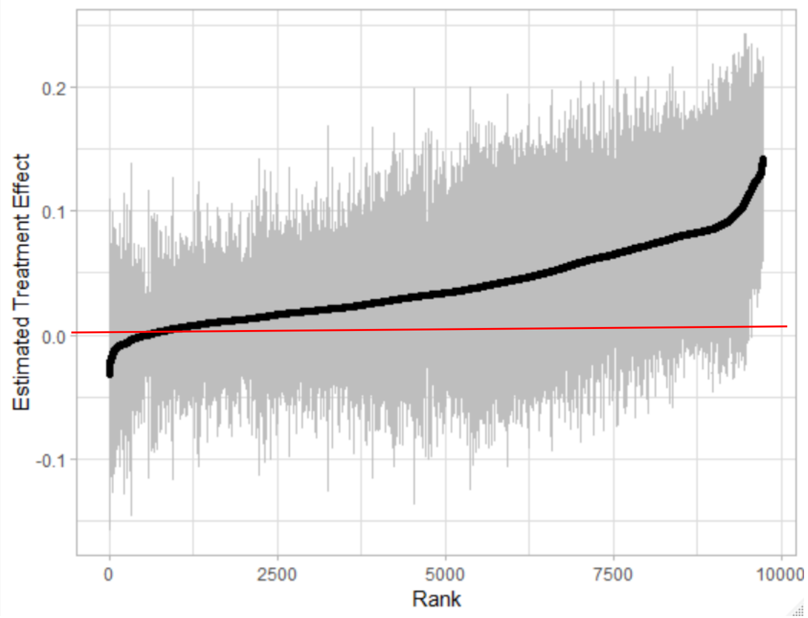
- Age
- Wealth
- Education
- Region
- Birth order
- Etc.



Predicted individual specific gain from having health insurance

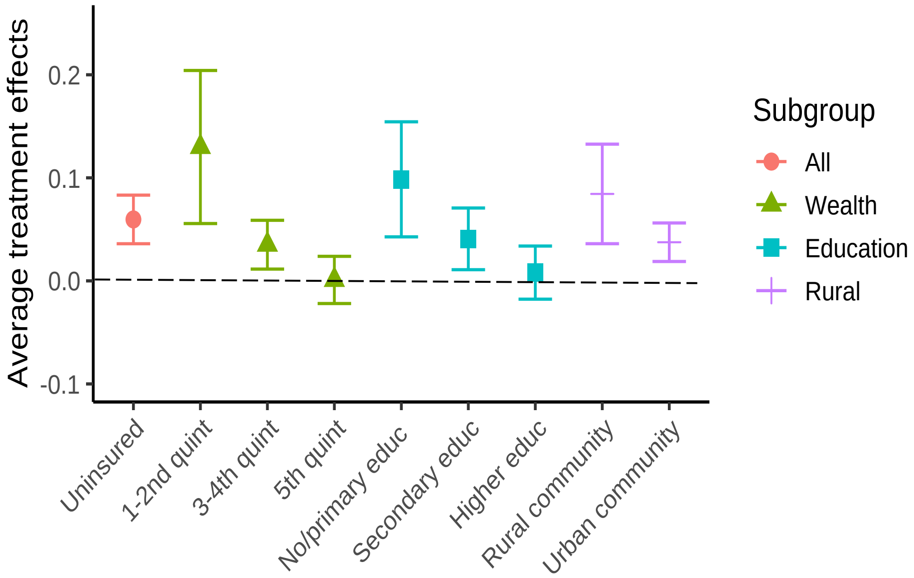
Distribution of estimated individual level treatment effects from CF (contributory health insurance)



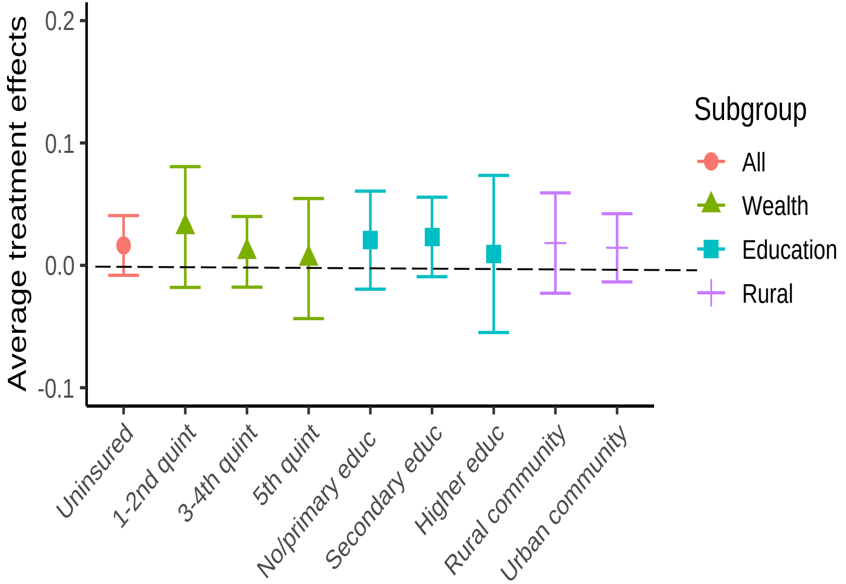


Pre-specified subgroups CATCs from causal forests

Contributory health insurance

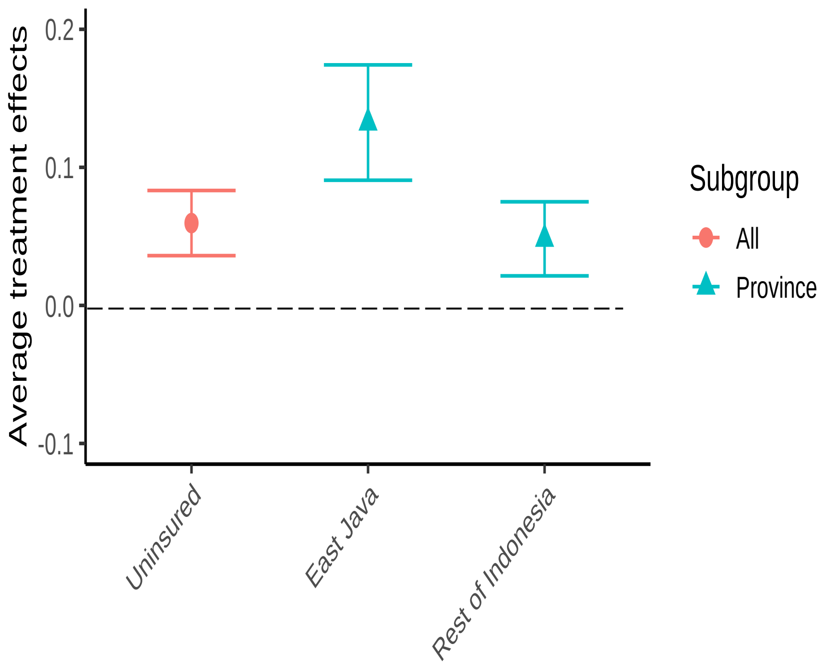


Subsidised health insurance

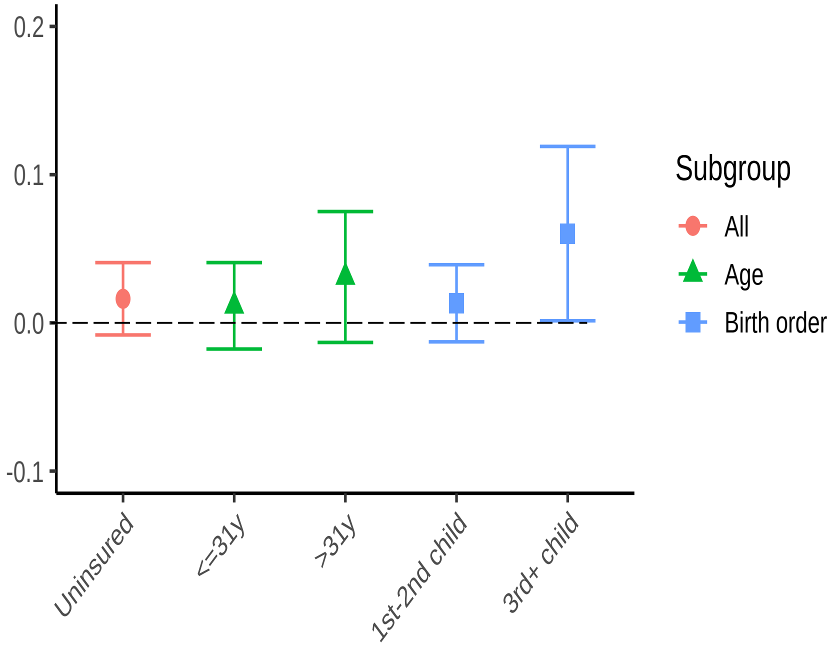


(Some) “Discovered” subgroups CATCs from causal forests

Contributory health insurance



Subsidised health insurance



Discussion

- Crucial developments in linking ML and causal inference frameworks in health and social sciences
- Causal ML can help learning about treatment effect heterogeneity
- For Indonesian health insurance expansion CF uncovers heterogeneity in treatment effects, for contributory HI (pro poor)
- Null results of subsidised HI can be explained by not effective HI (due to supply side constraints)
- Future avenues: learn optimal policy allocation rules, respecting constaints
- Challenge in health and social sciences: strong assumptions of no unobserved confounding
  - ML developed for instrumental variable estimation and panel data settings

Noemi.Kreif@York.ac.uk

@nkreif

Medical Research Council

New Investigator Resarch Grant: “Tailoring health policies to improve outcomes using machine learning, causal inference and operations research methods”

**Who Benefits from Health insurance?
Uncovering Heterogeneous Policy Impacts Using Causal Machine Learning**

Noemi Kreif, Andrew Mirelman, Rodrigo Moreno-Serra, Taufik Hidayat, Karla DiazOrdaz, Marc Suhrcke

CHE Research Paper 173

References

- VanderWeele TJ, Luedtke AR, van der Laan MJ, Kessler RC. Selecting optimal subgroups for treatment using many covariates. Epidemiology. 2019 May 1;30(3):334-41.
- Athey S, Tibshirani J, Wager S. Generalized random forests. The Annals of Statistics. 2019;47(2):1148-78
- Luedtke AR, Van Der Laan MJ. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. Annals of statistics. 2016 Apr;44(2):713.
- Chernozhukov V, Demirer M, Duflo E, Fernandez-Val I. Generic machine learning inference on heterogenous treatment effects in randomized experiments. National Bureau of Economic Research; 2018 Jun 7.
- Athey S, Wager S. Estimating Treatment Effects with Causal Forests: An Application. arXiv preprint arXiv:1902.07409. 2019 Feb 20.
- Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association. 2018 Jul 3;113(523):1228-42.

References

1. Nie X and Wager S. Quasi-oracle estimation of heterogeneous treatment effects. arXiv preprint arXiv:1712.04912 , 2017.
2. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, Volume 21, Issue 1, 1 February 2018, Pages C1–C68
3. Athey S, Tibshirani J, Wager S. Generalized random forests. The Annals of Statistics. 2019;47(2):1148-78
4. Künzel, Sören R., et al. "Metalearners for estimating heterogeneous treatment effects using machine learning." *Proceedings of the national academy of sciences* 116.10 (2019): 4156-4165.
5. Belloni A, Chernozhukov V, Hansen C. Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies. 2014 Apr 1;81(2):608-50.

Tuning parameters

Tuning parameter	grf package argument in causal_forest() function	Values (subsidised HI analysis)	Values (contributory HI analysis)
Fraction of the data used to build each tree	sample.fraction	0.472	0.500
Number of variables tried for each split	mtry	21	21
Minimum number of observations in each tree leaf	min.node.size	1	5
The fraction of data used for determining splits	honesty.fraction	0.620	0.500
Prunes the estimation sample tree such that no leaves are empty	honesty.prune.leaves	TRUE	TRUE
Maximum imbalance of a split	alpha	0.091	0.05
Controls how harshly imbalanced splits are penalized	Imbalance.penalty	0.061	0