Optimal decision rules with policy-relevant guarantees

Mats Stensrud

École Polytechnique Fédérale de Lausanne

Institute of Mathematics

Acknowledgements

This talk is based on joint work with

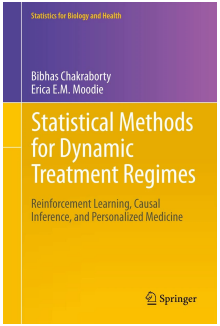
Oliver Dukes, Riccardo Brioschi, Aaron Hudson

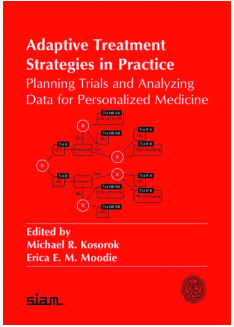
arXiv:2410.00985 (published in JASA, 2026)

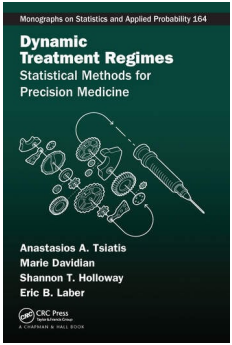
and unpublished work with

Julien Laurendeau, Oliver Dukes, and Aaron Hudson

Optimal dynamic strategies (Policy learning)

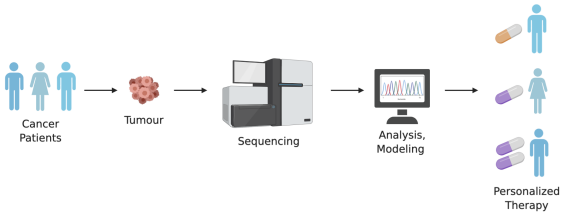








Relevance to precision medicine



The idea: tailor treatment decisions to patient characteristics

The premise: treatment effect heterogeneity

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

31 January 2019
EMA/CHMP/539146/2013
Committee for Medicinal Products for Human Use (CHMP)

Guideline on the investigation of subgroups in confirmatory clinical trials


U.S. FOOD & DRUG ADMINISTRATION

← Home / Regulatory Information / Search for FDA Guidance Documents / Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products

GUIDANCE DOCUMENT

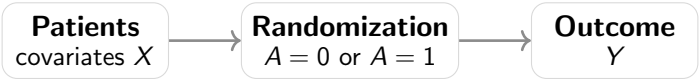
Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products

Guidance for Industry

And other problems



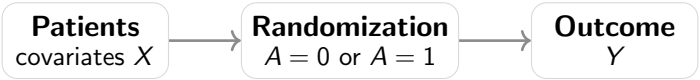
Consider a simple randomized trial



For patients with covariates $X = x$, the treatment effect is

$$\tau(x) = \mathbb{E}(Y \mid A = 1, X = x) - \mathbb{E}(Y \mid A = 0, X = x).$$

Consider a simple randomized trial



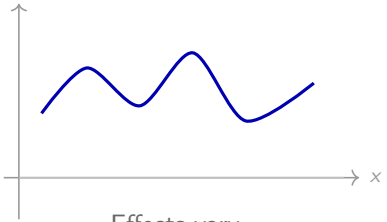
For patients with covariates $X = x$, the treatment effect is

$$\tau(x) = \mathbb{E}(Y \mid A = 1, X = x) - \mathbb{E}(Y \mid A = 0, X = x).$$

This is the conditional average treatment effect (CATE). The average (total) treatment effect is simply $\bar{\tau} = \mathbb{E}\{\tau(X)\}$.

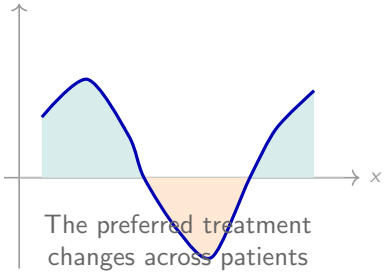
Two types of effect heterogeneity

Quantitative



Effects vary,
but have the same sign

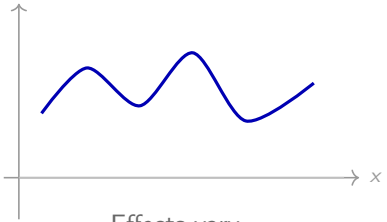
Qualitative



The preferred treatment
changes across patients

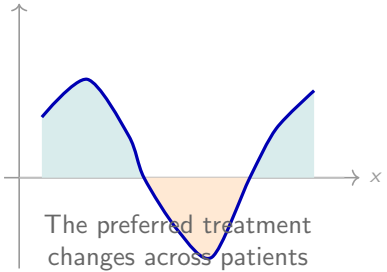
Two types of effect heterogeneity

Quantitative



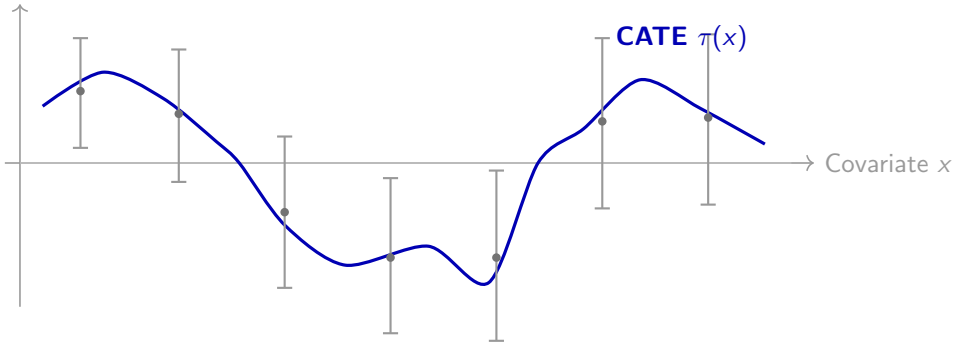
Effects vary,
but have the same sign

Qualitative



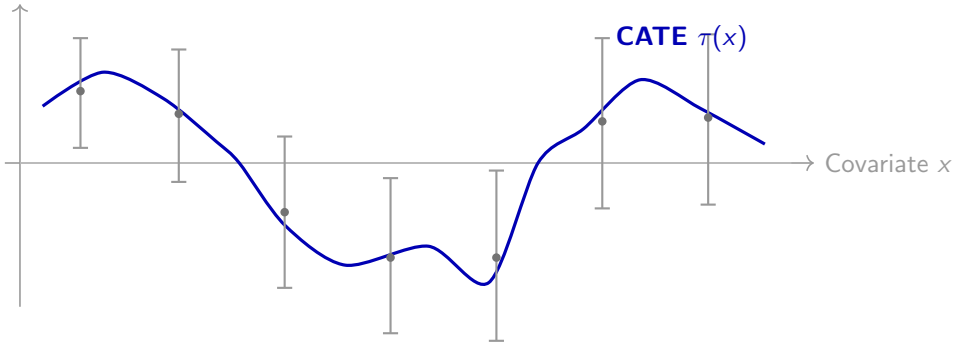
The preferred treatment
changes across patients

Why not simply estimate the entire CATE curve?



Estimating $\tau(x)$ precisely for every x is ambitious.

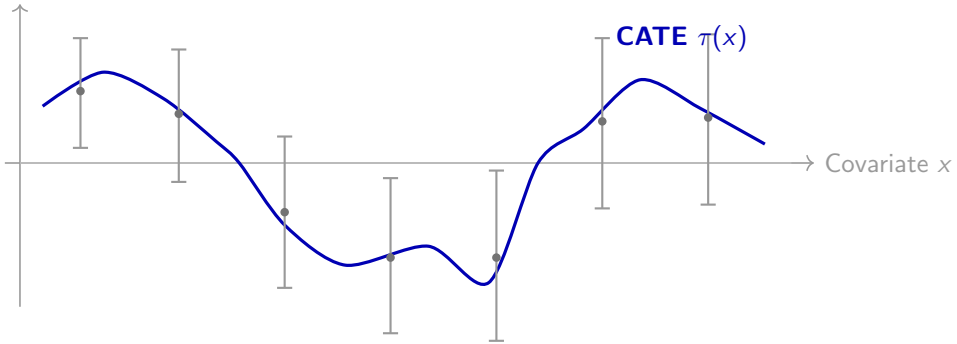
Why not simply estimate the entire CATE curve?



Estimating $\tau(x)$ precisely for every x is ambitious.

With many covariates, some regions contain little information, and uncertainty about the full curve can be substantial.

Why not simply estimate the entire CATE curve?

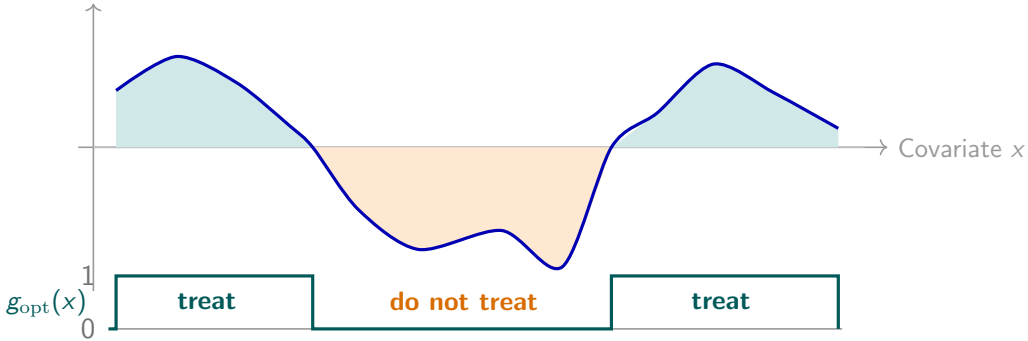


Estimating $\tau(x)$ precisely for every x is ambitious.

With many covariates, some regions contain little information, and uncertainty about the full curve can be substantial.

But our question may not require the whole curve.

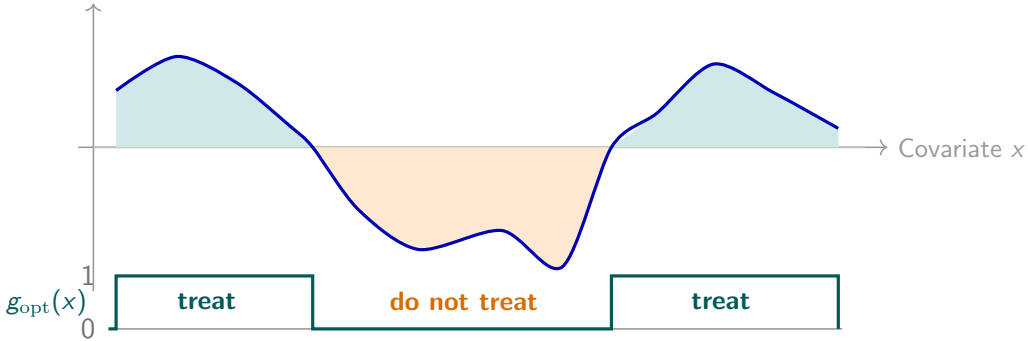
Easier to identify the sign



The optimal treatment rule, given $X = x$, is

$$g_{\text{opt}}(x) = I\{\tau(x) > 0\}.$$

Easier to identify the sign



The optimal treatment rule, given $X = x$, is

$$g_{\text{opt}}(x) = I\{\tau(x) > 0\}.$$

To choose treatment, we do not need to recover the magnitude of $\tau(x)$ everywhere.

What do we expect from a randomized experiment

Planning

- Protocol with primary and secondary endpoints
  Reasonable
- Sample size based on power calculation
  Reasonable

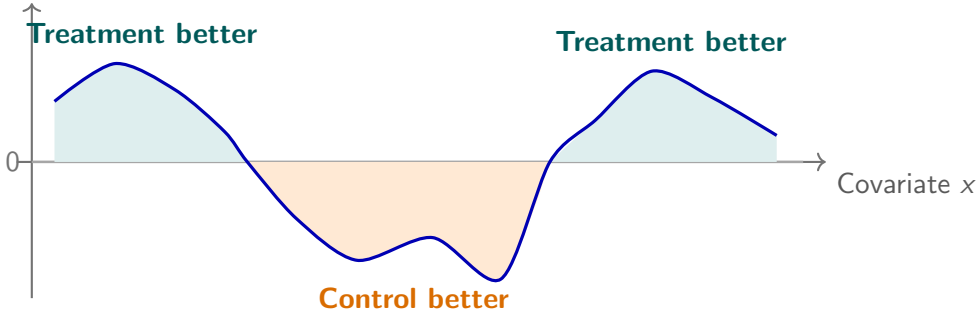
What do we expect from a randomized experiment

Planning

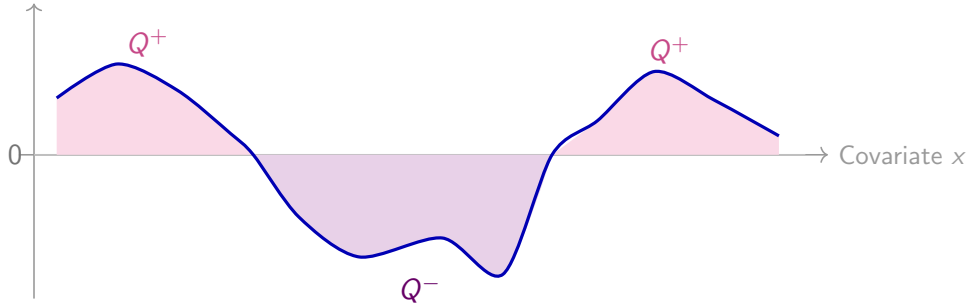
- Protocol with primary and secondary endpoints
  Reasonable
- Sample size based on power calculation
  Reasonable
- Subgroups analyses are pre-defined
  I understand the rationale, but too rigid? Can we use the data to find subgroups?
- Correction for multiple testing
  Perhaps sensible, but can we improve power compared to standard methods?

Proposal

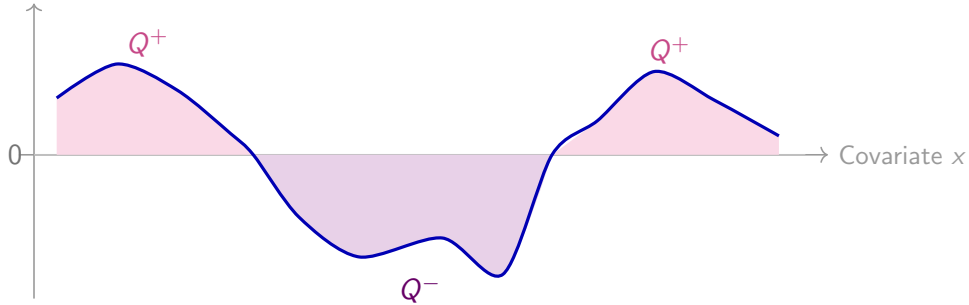
Motivating a test for qualitative heterogeneity



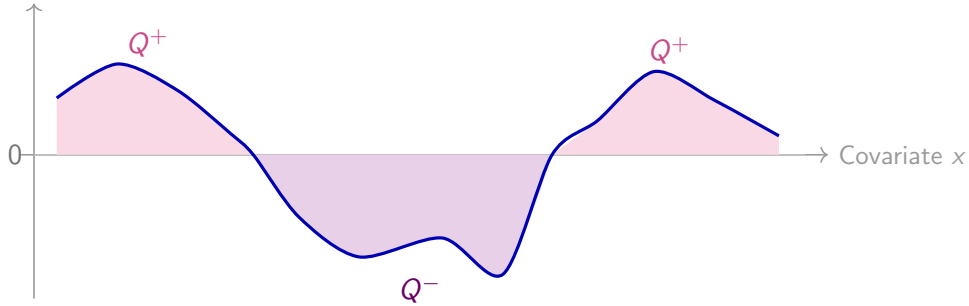
- Assess if there is heterogeneity (formally, by hypothesis testing)
- Characterize patient groups with (qualitatively) different average effects



$$Q^+ = \int \tau(x)\, I(\tau(x) > 0)\, dF(x), \quad Q^- = \int \tau(x)\, I(\tau(x) \leq 0)\, dF(x).$$



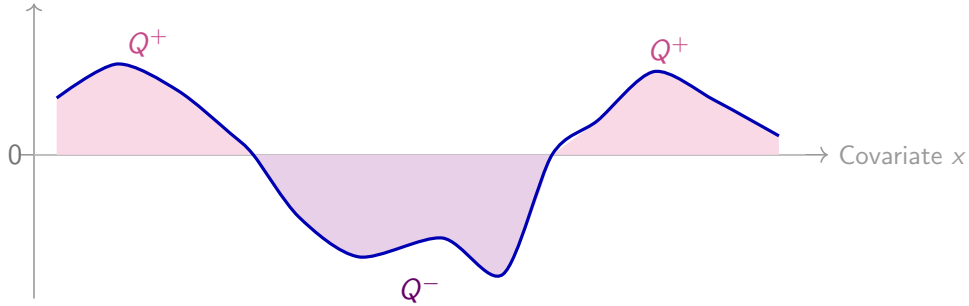
$$Q^+ = \int \tau(x)\, I(\tau(x) > 0)\, dF(x), \quad Q^- = \int \tau(x)\, I(\tau(x) \leq 0)\, dF(x).$$



$$Q^+ = \int \tau(x)\, I(\tau(x) > 0)\, dF(x), \quad Q^- = \int \tau(x)\, I(\tau(x) \leq 0)\, dF(x).$$

We can re-state the null of no qualitative heterogeneity as

$$H_0 : \; Q^+ > 0, \; Q^- = 0 \;\; \text{or} \;\; Q^+ = 0, \; Q^- < 0.$$

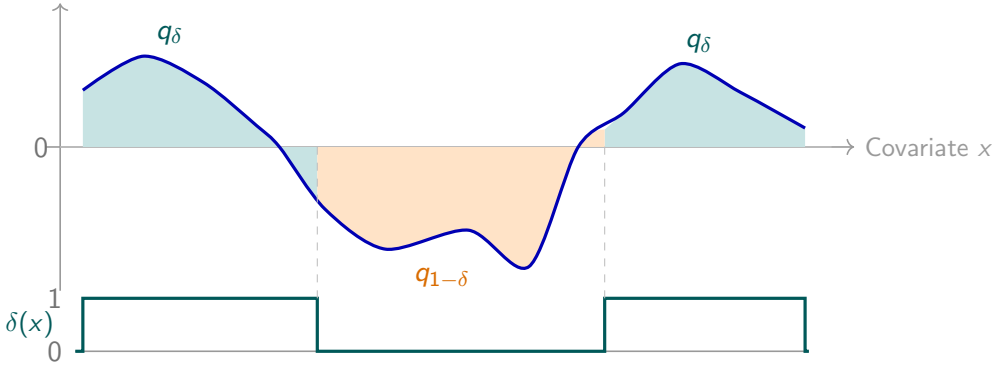


$$Q^+ = \int \tau(x)\, I(\tau(x) > 0)\, dF(x), \quad Q^- = \int \tau(x)\, I(\tau(x) \leq 0)\, dF(x).$$

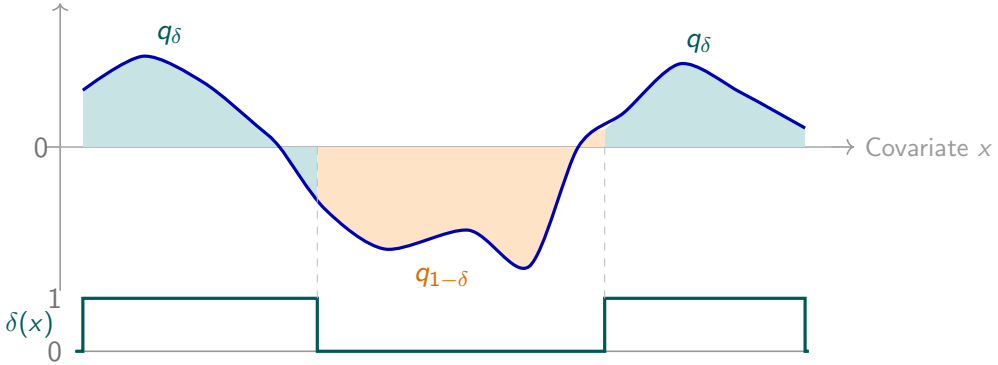
We can re-state the null of no qualitative heterogeneity as

$$H_0 : \ Q^+ > 0, \ Q^- = 0 \ \text{ or } \ Q^+ = 0, \ Q^- < 0.$$

But we don't estimate $\tau(x)$ well...



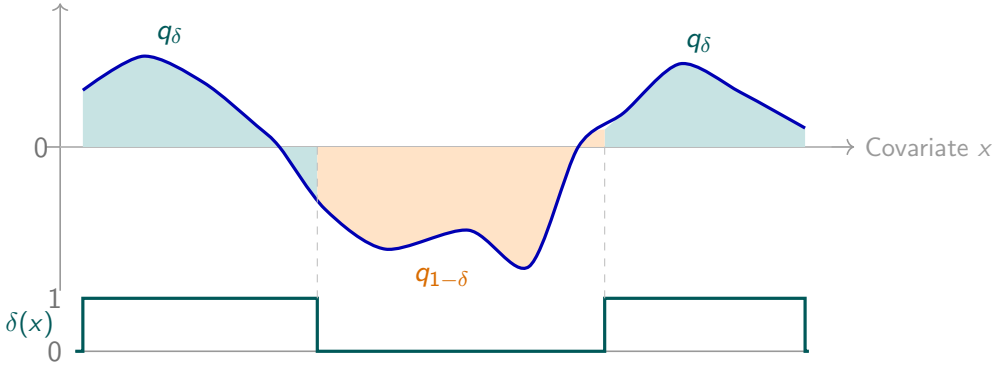
$$q_\delta = \int \tau(x)\,\delta(x)\,dF(x), \qquad q_{1-\delta} = \int \tau(x)\{1-\delta(x)\}\,dF(x).$$



$$q_\delta = \int \tau(x)\,\delta(x)\,dF(x), \qquad q_{1-\delta} = \int \tau(x)\{1-\delta(x)\}\,dF(x).$$

The oracle is $\delta(x) = I\{\tau(x) > 0\}$:

$\sup_{\delta\in\Delta} q_\delta \le Q^+$, $\inf_{\delta\in\Delta} q_{1-\delta} \ge Q^-$, for a class of functions Δ.



$$q_\delta = \int \tau(x)\,\delta(x)\,dF(x), \qquad q_{1-\delta} = \int \tau(x)\{1-\delta(x)\}\,dF(x).$$

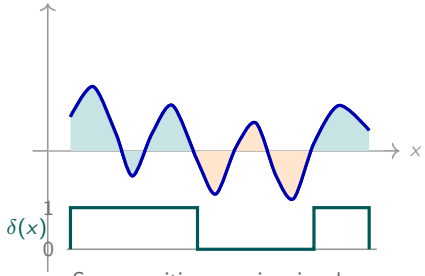
The oracle is $\delta(x) = I\{\tau(x) > 0\}$:

$\sup_{\delta\in\Delta} q_\delta \le Q^+$, $\inf_{\delta\in\Delta} q_{1-\delta} \ge Q^-$, for a class of functions Δ.

We search over a class $\delta \in \Delta$ for rules that make q_δ large and $q_{1-\delta}$ small.

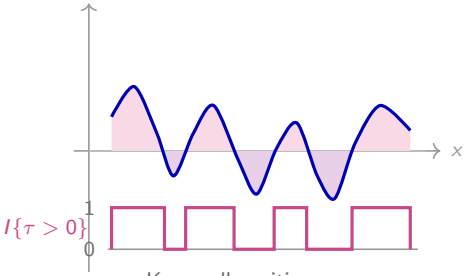
Why not choose a very small Δ?

Restricted class Δ



Some positive area is missed,
and some negative area is included.

Oracle: $I\{\tau(x) > 0\}$



Keeps all positive area
and excludes all negative area.

$$\sup_{\delta \in \Delta} q_\delta \;\leq\; Q^+, \qquad \inf_{\delta \in \Delta} q_{1-\delta} \;\geq\; Q^-.$$

The inequalities become equalities if $I\{\tau(x) > 0\} \in \Delta$.

Conceptually, why does this work?

- When Δ is not too complex, we can characterise the null distribution of $\sup_{\delta \in \Delta} |q_{\delta,n}|$ using empirical process theory.

  (van der Vaart & Wellner, 1996; Kosorok, 2008)

- Can use multiplier bootstrap to approximate the null limiting distibution.

  (Chernozhukov et al, 2013; Hsu, 2016)

- Makes testing straightforward, as we reject the null when both $\sup_{\delta \in \Delta} q_{\delta,n}$ and $\inf_{\delta \in \Delta} q_{1-\delta,n}$ are large.

In particular, no sample splitting involved

When does the test have power?

- Interpret each $\delta \in \Delta$ as a treatment strategy:
  subjects with covariates $X = x$ are treated with probability $\delta(x)$.
- Define the selected rule as

$$\delta^* \in \arg\max_{\delta \in \Delta} q_{\delta,n}.$$

When does the test have power?

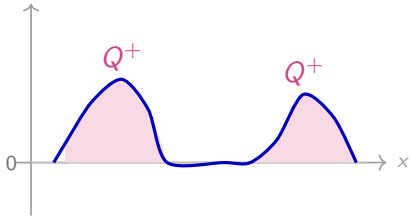
- Interpret each $\delta \in \Delta$ as a treatment strategy:
  subjects with covariates $X = x$ are treated with probability $\delta(x)$.
- Define the selected rule as

$$\delta^* \in \arg\max_{\delta \in \Delta} q_{\delta,n}.$$

- Power if there is a strategy given by $\delta^*(x) \in \Delta$ outperforms "never treat" and "always treat" (Shi et al, 2019)
- Conceptually, can lose power if
  - Δ too small: does not contain a δ^*.
  - Δ too large: $\sup_{\delta \in \Delta} |q_{\delta,n}|$ has high variance.

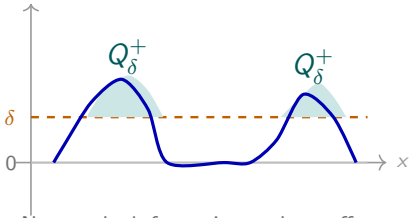
From qualitative heterogeneity to beneficial subgroups

Individuals with positive effects



No harmful region: only zero effect or beneficial effect.

Individuals with effect $> \delta$

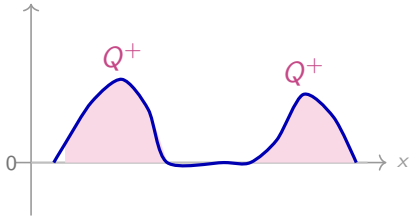


Now we look for patients whose effect exceeds a clinically relevant threshold.

$H_0 : \ Q^+ = 0$ or, more generally, $H_0 : \ Q_\delta^+ = 0.$

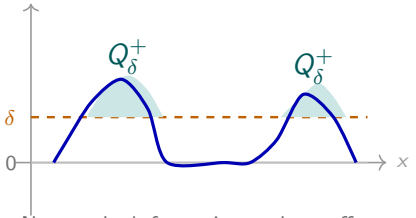
From qualitative heterogeneity to beneficial subgroups

Individuals with positive effects



No harmful region: only zero effect or beneficial effect.

Individuals with effect $> \delta$



Now we look for patients whose effect exceeds a clinically relevant threshold.

$H_0: \ Q^+ = 0$ or, more generally, $H_0: \ Q_\delta^+ = 0.$

The same testing idea adapts directly to this one-sided setting: replace "crossing zero" by "exceeding the target threshold." Because we no longer need to detect both positive and negative regions, power is often stronger in this setting.

What does our procedure guarantee?

Conceptually: over repeated iterations of the complete procedure, the probability that the test rejects and we subsequently report the selected subgroup as benefiting from treatment, when that subgroup's true average treatment effect does not exceed δ, is asymptotically at most α. I think this is relevant to pharma companies, EMA, FDA etc!

What does our procedure guarantee?

Conceptually: over repeated iterations of the complete procedure, the probability that the test rejects and we subsequently report the selected subgroup as benefiting from treatment, when that subgroup's true average treatment effect does not exceed δ, is asymptotically at most α. I think this is relevant to pharma companies, EMA, FDA etc!
Stronger and more interpretable than, e.g., Li & Brunskill, Science, 2026.
More formally:

■

$$P(\text{falsely conclude qualitative heterogeneity}) \leq \alpha + o(1).$$

What does our procedure guarantee?

Conceptually: over repeated iterations of the complete procedure, the probability that the test rejects and we subsequently report the selected subgroup as benefiting from treatment, when that subgroup's true average treatment effect does not exceed δ, is asymptotically at most α. I think this is relevant to pharma companies, EMA, FDA etc!
Stronger and more interpretable than, e.g., Li & Brunskill, Science, 2026.
More formally:

- $$P(\text{falsely conclude qualitative heterogeneity}) \leq \alpha + o(1).$$
- Guarantees for the selected subgroup:
  For any fixed data-generating law P_0 (not necessarily under the null) satisfying mild regularity conditions,

$$\limsup_{n\to\infty} P_0\left\{\text{reject } H_0,\ \theta^+_{0,\delta}(\widehat{f}_n) \leq 0\right\} \leq \alpha.$$

Practice

Like in an RCT, do power calculations

Planning

- Define the outcome Y and the covariate vector X
- Power calculation to decide on Δ

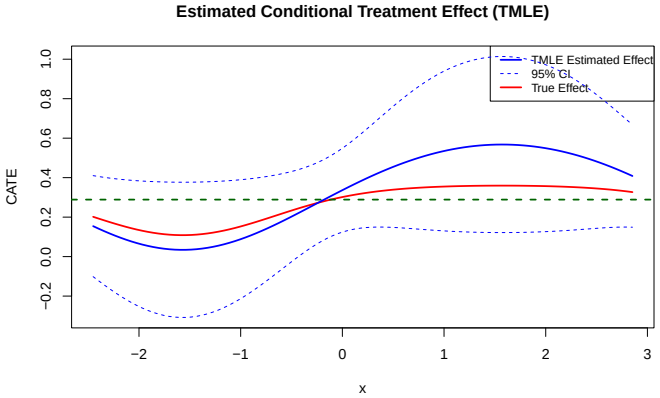
  Say, a class of classification trees (CARTs) with k leaves

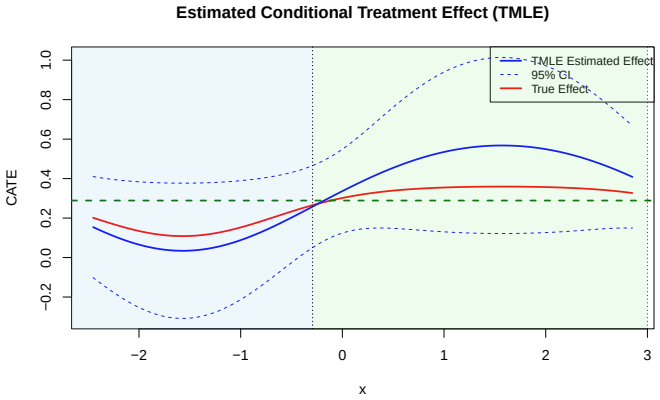
Analysis

- Compute the test statistic
- Inversion of the test statistic gives a decision tree

  with the regret guarantees from Athey and Wager, 2021, Econometrica

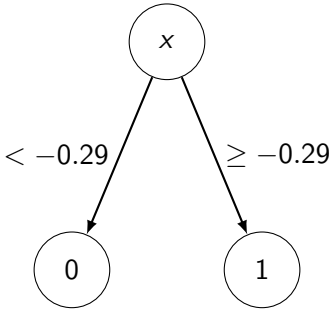
True CATE and estimated CATE with 95% CI





Inverted test statistic ($k = 2$), p-value < 0.01 .

Corresponding decision tree



nature

Explore content | About the journal | Publish with us

nature › articles › article

Article | Published: 16 August 2017

A randomized synbiotic trial to prevent sepsis among infants in rural India

Article | Published: 16 August 2017

A randomized synbiotic trial to prevent sepsis among infants in rural India

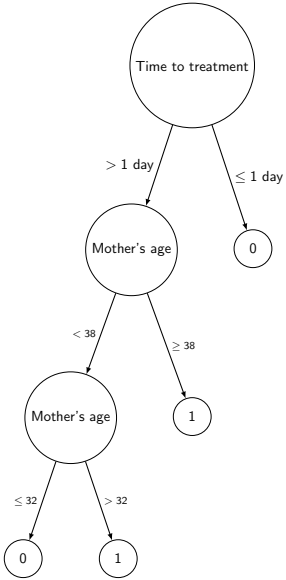


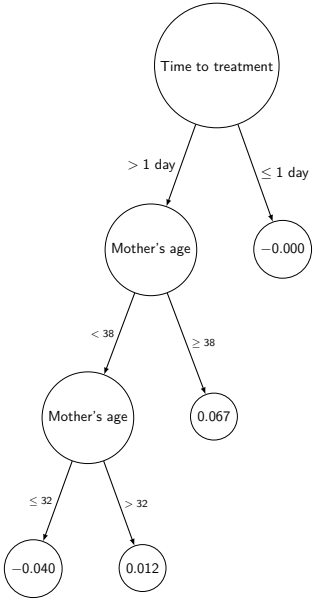
The researchers led by Pinaki Panigrahi (left) surveyed more than 4500 infants in rural Odisha. Credit: P.

Panigrahi et al. (2017)

- Synbiotic treatment vs control on sepsis or death in rural India
- Included 4,556 infants with no signs of sepsis or other morbidity
- Monitored for 60 days
- Substantial improvement with synbiotic treatment
- X included mother's age, mother parity, birth weight, gestational age, and when treatment was initiated

Here, we choose max $k = 5$ leaves, tested for a treatment effect larger than 0.01.
P-value: < 0.01 (with and without sample splitting),


Time to treatment
> 1 day
≤ 1 day
Mother's age
0
< 38
≥ 38
Mother's age
1
≤ 32
> 32
0
1


Time to treatment
> 1 day
≤ 1 day
Mother's age
−0.000
< 38
≥ 38
Mother's age
0.067
≤ 32
> 32
−0.040
0.012

Categorization

Categorizing Continuous Variables

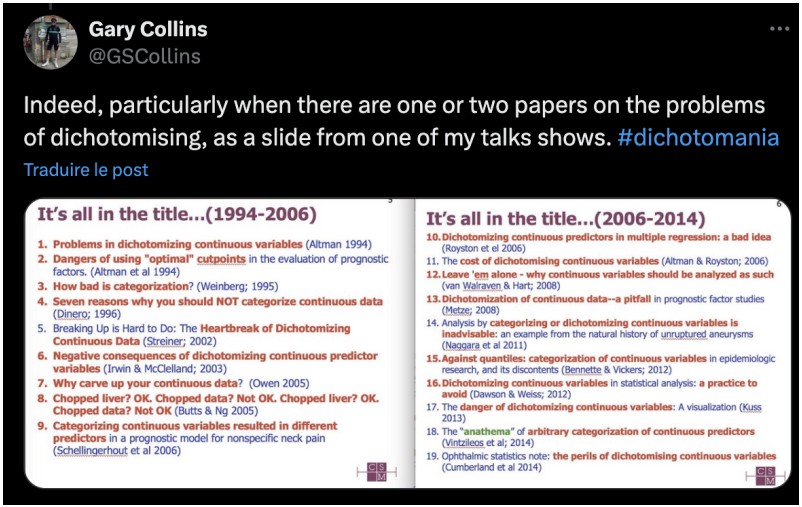
research methods | measurement | categorization

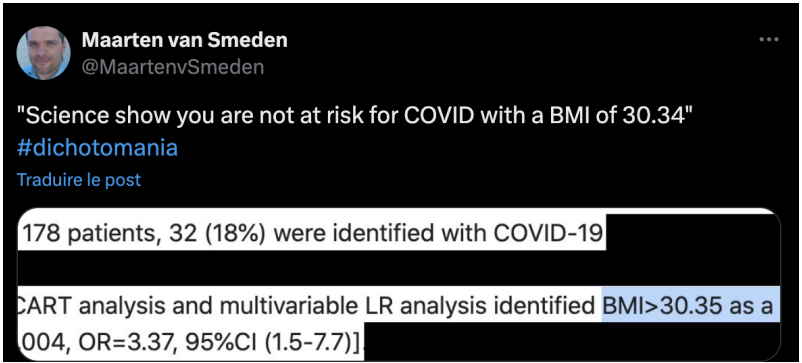


f2harrell | 6 | Jun 2022

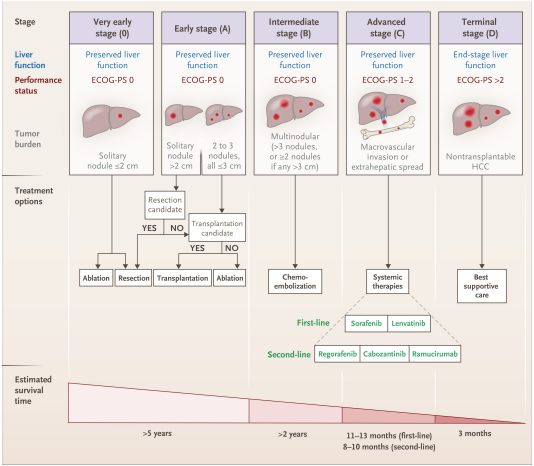
Problems Caused by Categorizing Continuous Variables

References 257 | Key reference 173 | Key Reference for Response Variables 64 | More References 40 | Myths About Risk Thresholds | Dichotomania 107


Gary Collins
@GSCollins
Indeed, particularly when there are one or two papers on the problems of dichotomising, as a slide from one of my talks shows. #dichotomania
Traduire le post
It's all in the title...(1994-2006)
1. Problems in dichotomizing continuous variables (Altman 1994)
2. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. (Altman et al 1994)
3. How bad is categorization? (Weinberg; 1995)
4. Seven reasons why you should NOT categorize continuous data (Dinero; 1996)
5. Breaking Up is Hard to Do: The Heartbreak of Dichotomizing Continuous Data (Streiner; 2002)
6. Negative consequences of dichotomizing continuous predictor variables (Irwin & McClelland; 2003)
7. Why carve up your continuous data? (Owen 2005)
8. Chopped liver? OK. Chopped data? Not OK. Chopped liver? OK. Chopped data? Not OK (Butts & Ng 2005)
9. Categorizing continuous variables resulted in different predictors in a prognostic model for nonspecific neck pain (Schellingerhout et al 2006)
It's all in the title...(2006-2014)
10. Dichotomizing continuous predictors in multiple regression: a bad idea (Royston et el 2006)
11. The cost of dichotomising continuous variables (Altman & Royston; 2006)
12. Leave 'em alone - why continuous variables should be analyzed as such (van Walraven & Hart; 2008)
13. Dichotomization of continuous data--a pitfall in prognostic factor studies (Metzel; 2008)
14. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms (Naggara et al 2011)
15. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents (Bennette & Vickers; 2012)
16. Dichotomizing continuous variables in statistical analysis: a practice to avoid (Dawson & Weiss; 2012)
17. The danger of dichotomizing continuous variables: A visualization (Kuss 2013)
18. The "anathema" of arbitrary categorization of continuous predictors (Vintzileos et al; 2014)
19. Ophthalmic statistics note: the perils of dichotomising continuous variables (Cumberland et al 2014)


Maarten van Smeden
@MaartenvSmeden
"Science show you are not at risk for COVID with a BMI of 30.34"
#dichotomania
Traduire le post
178 patients, 32 (18%) were identified with COVID-19
CART analysis and multivariable LR analysis identified BMI>30.35 as a
004, OR=3.37, 95%CI (1.5-7.7)]

Practitioners use decision trees all the time...



NEJM

Relevant work

- Several papers have considered nonparametric tests of quant. homogeneity,

  (Crump et al., 2008; Ding et al., 2019; Sánchez-Becerra, 2022)

- tests of whether the CATE is non-negative (or non-positive),

  (Chang et al., 2015; Hsu, 2017; Shi et al., 2019; Johnson et al., 2023)

- or on inference on subgroups that inform policies

  (Athey and Imbens, 2016; Athey and Wager, 2021; Chernozhukov et al., 2023; Reeve et a., 2023; Muller et al., 2024)

Discussion

- Proposed tests that assess heterogeneity
  by estimating the area above/below the treatment effect curve.
- Parametric-rate estimators of the AUC
  via its representation as a weighted treatment effect.
- Asymptotic type-I error control under very mild conditions,
  non-trivial asymptotic power
  (in directions of practical interest).
- Unifies qualitative and quantitative tests

  There are many open questions. We are working one tree implementations, survival outcomes, improved power, relaxation of assumptions in randomized experiments

Nonparametric tests of treatment effect homogeneity for policy-makers

Oliver Dukes, Mats Stensrud, Riccardo Brioschi, Aaron Hudson arXiv:2410.00985

Thanks also to my PhD student Julien Laurendeau

Appendix

A Nonparametric Test for Qual Heterogeneity (2)

- Let $\delta(x) \in [0,1]$, and define

$$q_\delta = \int \tau(x)\,\delta(x)\,dF(x),$$

- The oracle weight is $\delta(x) = I\{\tau(x) > 0\}$:
  That is, we have the relation

$$\sup_{\delta\in\Delta} q_\delta \le Q^+, \quad \inf_{\delta\in\Delta} q_{1-\delta} \ge Q^-,$$

  for a class of functions Δ, where each $\delta \in \Delta$ satisfies $\delta : \mathcal{X} \to [0,1]$,
  equalities when $I(\tau(x) > 0) \in \Delta$.

A Nonparametric Test for Qual Heterogeneity (2)

- Let $\delta(x) \in [0,1]$, and define

$$q_\delta = \int \tau(x)\,\delta(x)\,dF(x),$$

- The oracle weight is $\delta(x) = I\{\tau(x) > 0\}$:
  That is, we have the relation

$$\sup_{\delta\in\Delta} q_\delta \le Q^+, \quad \inf_{\delta\in\Delta} q_{1-\delta} \ge Q^-,$$

  for a class of functions Δ, where each $\delta \in \Delta$ satisfies $\delta : \mathcal{X} \to [0,1]$,
  equalities when $I(\tau(x) > 0) \in \Delta$.

- So we can construct an estimator $q_{\delta,n}$ of q_δ, and then assess if

$$\sup_{\delta\in\Delta} q_{\delta,n} > 0 \quad \text{and} \quad \inf_{\delta\in\Delta} q_{1-\delta,n} < 0.$$

Relation to policy learning via utilitarian regret

The test gives a treatment regime that minimizes regret.

Relation to policy learning via utilitarian regret

The test gives a treatment regime that minimizes regret. More explicitly:

- Again, interpret each $\delta \in \Delta$ as a treatment regime.

Relation to policy learning via utilitarian regret

The test gives a treatment regime that minimizes regret. More explicitly:

- Again, interpret each $\delta \in \Delta$ as a treatment regime.
- The utilitarian regret of using δ is

$$R(\delta) = \max_{\delta' \in \Delta} \mathbb{E}\{Y(\delta')\} - \mathbb{E}\{Y(\delta)\}.$$

Relation to policy learning via utilitarian regret

The test gives a treatment regime that minimizes regret. More explicitly:

- Again, interpret each $\delta \in \Delta$ as a treatment regime.
- The utilitarian regret of using δ is

$$R(\delta) = \max_{\delta' \in \Delta} \mathbb{E}\{Y(\delta')\} - \mathbb{E}\{Y(\delta)\}.$$

- Athey and Wager, Econometrica 2021[†] gave guarantees on the regret $R(\delta^*)$:
  - near-oracle performance within Δ; that is,

$$R(\delta^*) = O_P(1/\sqrt{n}).$$

  - $R(\delta^*)$ bounded on the order of

$$\sqrt{\mathrm{VC}(\Delta)/n}.$$

Relation to policy learning via utilitarian regret

The test gives a treatment regime that minimizes regret. More explicitly:

- Again, interpret each $\delta \in \Delta$ as a treatment regime.
- The utilitarian regret of using δ is

$$R(\delta) = \max_{\delta' \in \Delta} \mathbb{E}\{Y(\delta')\} - \mathbb{E}\{Y(\delta)\}.$$

- Athey and Wager, Econometrica 2021[†] gave guarantees on the regret $R(\delta^*)$:
  - near-oracle performance within Δ; that is,

$$R(\delta^*) = O_P(1/\sqrt{n}).$$

  - $R(\delta^*)$ bounded on the order of

$$\sqrt{\mathrm{VC}(\Delta)/n}.$$

To my knowledge, they did not:

- consider how to choose Δ beyond cross-fitting and expert knowledge, or
- give guarantees beyond regret bounds.

[†] Following Manski (2004, 2009), Hirano and Porter (2009), Stoye (2009, 2012), Kitagawa and Tetenov (2018), and others.

On inference guarantees

- Our test is designed to detect treatment effect heterogeneity.

On inference guarantees

- Our test is designed to detect treatment effect heterogeneity.
- If the test rejects, investigators often want conditional guarantees:
  For the subgroup selected to be treated,

$$\{x : \delta^*(x) = 1\}.$$

On inference guarantees

- Our test is designed to detect treatment effect heterogeneity.
- If the test rejects, investigators often want conditional guarantees:
  For the subgroup selected to be treated,

$$\{x : \delta^*(x) = 1\}.$$

  Can we formally test whether the selected group has beneficial effects?

On inference guarantees

- Our test is designed to detect treatment effect heterogeneity.
- If the test rejects, investigators often want conditional guarantees:
  For the subgroup selected to be treated,

$$\{x : \delta^*(x) = 1\}.$$

Can we formally test whether the selected group has beneficial effects?

A difficulty is that the subgroup is data-adaptive,
because δ^* is estimated from the data.
This is therefore a post-selection inference problem.

On inference guarantees

- Our test is designed to detect treatment effect heterogeneity.
- If the test rejects, investigators often want conditional guarantees:
  For the subgroup selected to be treated,

$$\{x : \delta^*(x) = 1\}.$$

  Can we formally test whether the selected group has beneficial effects?
  A difficulty is that the subgroup is data-adaptive,
  because δ^* is estimated from the data.
  This is therefore a post-selection inference problem.
- A simple solution is sample splitting:
  - use one part of the data to learn δ^*,
  - use the other part to evaluate effects in the selected group.

  But can we do better?

Tree procedure

Greedy algorithm

Conditional guarantees

Sample splitting
Chiseling

Simulation procedure

- Repeat the following many times (here: 20 000 times):
  - Draw a sample of 100 units of $O = (X, A, Y)$.
  - Estimate the CATE using a flexible doubly-robust estimator.
  - Estimate the optimal regime g_{opt} based on the sign of the CATE.
  - Evaluate the true value of $\mathbb{E}\big(Y_{g_{\text{opt}}}\big)$.
- In the same sample, set $\delta = 2$:
  - Learn the decision tree.
  - Invert to get a regime g_{tree}.
  - Evaluate the true value of $\mathbb{E}(Y_{g_{\text{tree}}})$.
- Compare the distributions of true values under the estimated regimes.

	$\hat{g}_{\text{bin}}$	$\hat{g}_{\text{TMLE}}$
$V(\cdot)$	**0.58**	0.55
$\sigma^2(\cdot)$	**0.012**	0.099

Table: Estimated expected performance evaluated in the sample over 20′000 runs for the estimated regimes resulting from the binning procedure with up to 4 bins versus the TMLE for $n = 100$ samples and where the CATE is given by $-\sin(2\pi x) - 0.9\tanh(25(x - 0.5))$ in $[0, 1]$. Let $\hat{g}_{\text{bin}}$ denote the estimated binned optimal regime and $\hat{g}_{\text{TMLE}}$ be the estimated optimal regime using an estimate of the CATE. Let $V(g) := \mathbb{E}[Y(g)]$, and for an estimated regime $\hat{g}$, let $\sigma^2(\hat{g}) = \text{Var}(\mathbb{E}[Y(\hat{g})])$.

Regime performance (true value distribution)

Regime (method)	Mean true value	Variance
Binning (tree)	0.58	0.012
TMLE	0.55	0.099

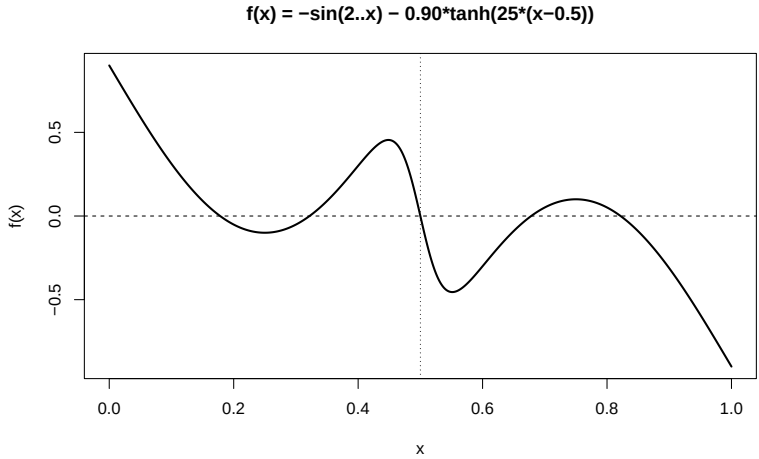
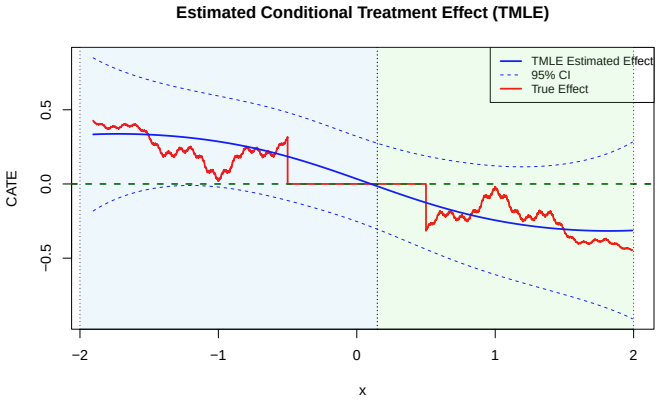


Figure: Caption


Estimated Conditional Treatment Effect (TMLE)
TMLE Estimated Effect
95% CI
True Effect
CATE
0.5
0.0
−0.5
−2
−1
0
1
2
x

A Nonparametric Test for Quant Heterogeneity (1)

- We can re-state the null of no quantitative heterogeneity as

$$H_0 : Q := \int |\tau(x) - \bar{\tau}| dF(x) = 0$$

- The alternative is $H_1 : Q > 0$.

A Nonparametric Test for Quant Heterogeneity (1)

- We can re-state the null of no quantitative heterogeneity as

$$H_0 : Q := \int |\tau(x) - \bar{\tau}| dF(x) = 0$$

- The alternative is $H_1 : Q > 0$.

- However, standard nonparametric functional estimators have an intractable limiting distribution under the null

A Nonparametric Test for Quant Heterogeneity (2)

- To get around this, we express Q as an optimal weight:

A Nonparametric Test for Quant Heterogeneity (2)

- To get around this, we express Q as an optimal weight:
  Let $\omega(x) \in [-1, 1]$ be an arbitrary function

$$q_\omega := \int \{\tau(x) - \bar{\tau}\}\omega(x) dF(x)$$

A Nonparametric Test for Quant Heterogeneity (2)

- To get around this, we express Q as an optimal weight:
  Let $\omega(x) \in [-1, 1]$ be an arbitrary function

$$q_\omega := \int \{\tau(x) - \bar{\tau}\}\omega(x) dF(x)$$

- For a class of functions Ω, we have

$$\sup_{\omega \in \Omega} |q_\omega| \leq Q$$

  with equality when Ω contains $\text{sign}(\tau(x) - \bar{\tau})$.

A Nonparametric Test for Quant Heterogeneity (2)

- To get around this, we express Q as an optimal weight:
  Let $\omega(x) \in [-1, 1]$ be an arbitrary function

$$q_\omega := \int \{\tau(x) - \bar{\tau}\}\omega(x) dF(x)$$

- For a class of functions Ω, we have

$$\sup_{\omega \in \Omega} |q_\omega| \leq Q$$

  with equality when Ω contains $\text{sign}(\tau(x) - \bar{\tau})$.

- To detect heterogeneity, suffices to show $\sup_{\omega \in \Omega} |q_\omega| > 0$.

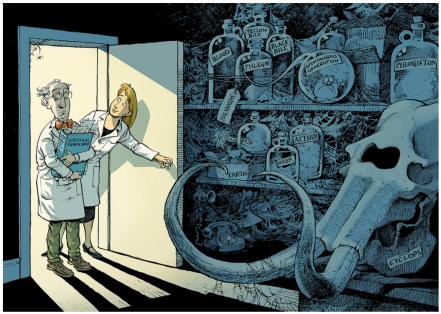
Estimation vs testing

COMMENT · 20 MARCH 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Valentin Amrhein ✉, Sander Greenland & Blake McShane


P-VALUES
P<0.05
P>0.05
STATISTICALLY SIGNIFICANT
P-VALUES
P-VALUES
P-VALUES

Estimation vs testing

Although debated, hypothesis testing serves a purpose here:

- **Guarantees.** Limit false-positive claims of effectiveness
  mirroring the logic used by FDA, EMA?
- **Feasible.** When fully estimating heterogeneous effects is too ambitious, testing *coarser* features is tractable and reliable.
- **Actionability.** Tests targeting qualitative heterogeneity can reveal groups with effects of different *sign*, directly informing clearer treatment policies.
- **Complementary, not a substitute.** Positive findings should be *supplemented* with transparent subgroup effect estimates (and confidence intervals).

Related work

- Athey and Imbens, PNAS, 2016
- Chernozhukov et al., NBER, 2023

Testing heterogeneity - easy or hard?

- If we have a finite number of subgroups:
  - To assess quantitative heterogeneity: are all subgroup effects equal?
  - To assess qualitative heterogeneity: are some effects positive whilst others are negative?

    (Gail & Simon, 1985; Piantadosi & Gail, 1993)

Testing heterogeneity - easy or hard?

- If we have a finite number of subgroups:
  - To assess quantitative heterogeneity: are all subgroup effects equal?
  - To assess qualitative heterogeneity: are some effects positive whilst others are negative?

    (Gail & Simon, 1985; Piantadosi & Gail, 1993)
- What if there are infinitely many subgroups?
  - *Does vaccine efficacy depend on a biomarker for immunogenicity?*

Testing heterogeneity - easy or hard?

- If we have a finite number of subgroups:
  - To assess quantitative heterogeneity: are all subgroup effects equal?
  - To assess qualitative heterogeneity: are some effects positive whilst others are negative?

    (Gail & Simon, 1985; Piantadosi & Gail, 1993)
- What if there are infinitely many subgroups?
  - *Does vaccine efficacy depend on a biomarker for immunogenicity?*

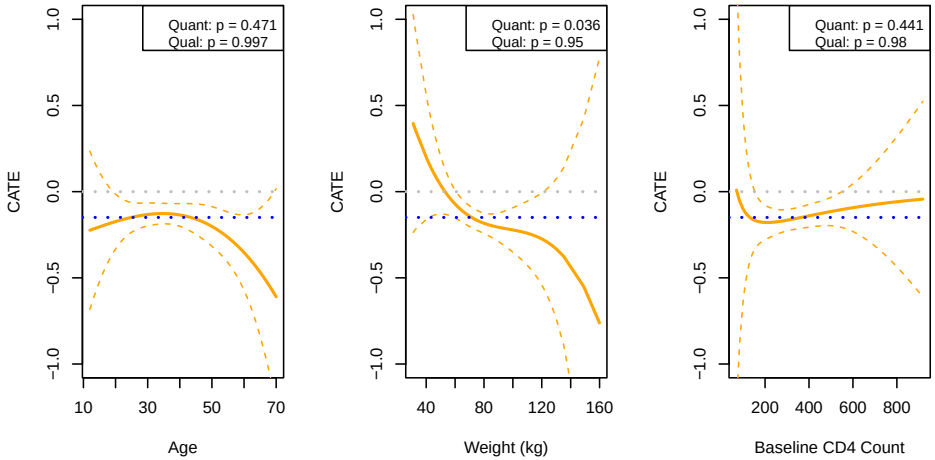
Our contribution

We provide approaches to nonparametric inference for qualitative and quantitative heterogeneity with infinitely many subgroups.

AIDS Clinical Trial Group Study 175

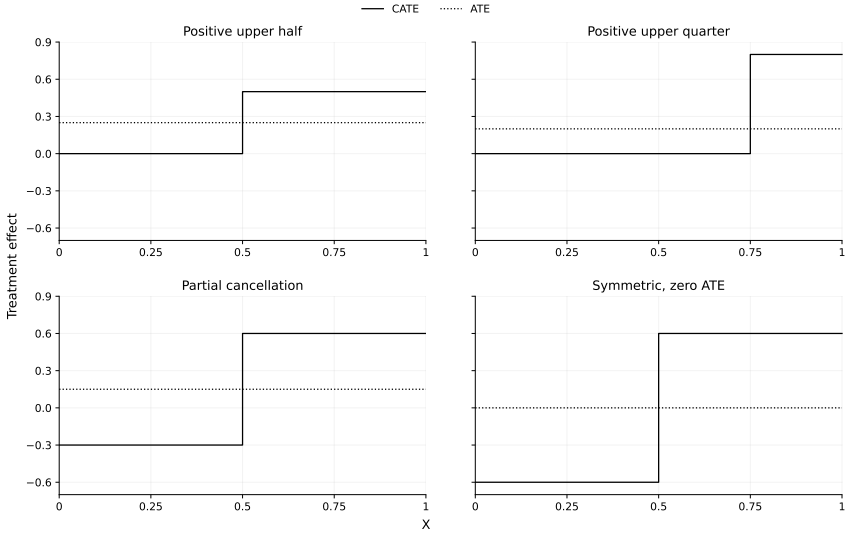
- Randomized trial comparing
  - monotherapy with zidovudine with;
  - monotherapy with didanosine, combination therapy with zidovudine and didnosine, or combination therapy with zidovudine and zalcitabine.
- Composite outcome:
  - occurrence of a fifty percent decline in the CD4 cell count,
  - development of the acquired immunodeficiency syndrome (AIDS),
  - death.
- Assess heterogeneity using
  weight, age, and baseline $\log_{10}$ CD4 count.

Results ($n = 1,938$)



Cubic spline estimates of CATE curves and p-values from tests of treatment effect heterogeneity. Dashed orange lines represent pointwise 95% confidence intervals.

Comparison to t-test - DGPs



Comparison to t-test - Rejection rates

