Comparison of recent approaches for subgroup identification from clinical and observational data

Ilya Lipkovich (Eli Lilly and Company), David Svensson (AstraZeneca)
Bohdana Ratitch (Bayer), Alex Dmitrienko (Mediana)

14th Annual Conference on Statistical Issues in Clinical Trials
April 12, 2022

Acknowledgements: Haoda Fu
Chapter 3
Data-Driven and Confirmatory Subgroup Analysis in Clinical Trials

Alex Dimitrienko, Ilya Lipkovich, Aaron Dane, and Christoph Muyers

Abstract In this chapter we provide an overview of the principles and practice of subgroup analysis in late-stage clinical trials. For convenience, we classify different subgroup analyses into two broad categories: data-driven and confirmatory. The two settings are different from each other primarily by the scope and extent of pre-specification of patient subgroups. First, we review key considerations in confirmatory subgroup analysis based on one or more pre-specified patient populations. This includes a survey of multiplicity adjustment methods recommended in multi-population Phase III clinical trials and decision-making considerations that ensure clinically meaningful inferences across the pre-defined populations. Secondly, we consider key principles for data-driven subgroup analysis and contrast it with that for a guideline-driven approach. Methods that emerged in the area of principled data-driven subgroup analysis in the last 10 years as a result of cross-pollination of machine learning, causal inference and multiple testing are reviewed. We provide examples of recommended approaches to data-driven and confirmatory subgroup analysis illustrated with data from Phase III clinical trials. We also illustrate common errors, pitfalls and misuse of subgroup analysis approaches in clinical trials often resulting from employing overly simplistic or naive methods. Overview of available statistical software and extensive bibliographical references are provided.

© Springer Nature Switzerland AG 2020
N. Ting et al. (eds.), Design and Analysis of Subgroups with Biopharmaceutical Applications, Emerging Topics in Statistics and Biostatistics, https://doi.org/10.1007/978-3-030-40105-4_3
The mythology of subgroup analysis in Pharma

<table>
<thead>
<tr>
<th>Common practices</th>
<th>“Good practices”</th>
</tr>
</thead>
<tbody>
<tr>
<td>One covariate at a time strategy, (e.g test interactions at alpha=0.1)</td>
<td>Subgroups should be “pre-specified” and “biologically plausible”</td>
</tr>
<tr>
<td>Multiplicity does not need to be controlled since “it is for internal decision making”, “not for submission”</td>
<td>The central role of covariate-by-treatment interaction test, as a “gatekeeper” (no testing in subgroups unless passing the interaction test)</td>
</tr>
<tr>
<td>Accounting for uncertainty in the very last step of a multi-stage strategy, forgetting about “preliminary data looks”</td>
<td>No testing in subgroups unless the effect in the overall population is significant (consistency)</td>
</tr>
<tr>
<td>The subgroup search involves human interaction that is rarely reported</td>
<td>“Data-driven elements should be minimized”</td>
</tr>
<tr>
<td>“Null findings” rarely reported</td>
<td>Interpreting results “with caution”</td>
</tr>
</tbody>
</table>
SA is a special case of statistical learning, rather than merely multiple testing problem

A key challenge is estimating individual treatment effects (not observable on any subject)

Intersection and cross-fertilization of different fields: causal inference, machine learning, multiple hypothesis testing.
Learning heterogeneity of TE from the data

\[ \text{CATE}(x) = \Delta(x) = E(Y(1)|X = x) - E(Y(0)|X = x) \]

CATE: Conditional Average Treatment Effect (a.k.a ITE, PTE)

\( X \) - possibly high dimensional

Post-selection inference

Machine learning

Multiple hypothesis testing

Causal inference
The set up: individual TE

- Each patient has two potential outcomes of $Y$, i.e. $Y_i(0), Y_i(1)$ corresponding to $T = 0, 1$; only one outcome is observed (SUTVA)
- Outcome function, given pre-treatment covariates
  $$m(t, x) = E(Y_i(t) | X = x), t \in \{0, 1\}$$
- Under treatment ignorability, ensured by randomization in RCT, or “no unmeasured confounder” assumption in OC
  $$m(t, x) = E(Y | T = t, X = x)$$
- Treatment contrast or conditional causal effect (CATE)
  $$\Delta(x) = m(1, x) - m(0, x)$$
- We can write the response surface as
  $$m(t, x) = h(x) + \frac{1}{2} \Delta(x)(2t - 1),$$
- $h(x)$ is the main covariate (prognostic) effect
- In studies with non-randomized treatments, we need to estimate propensity scores
  $$\pi(x) = P(T = 1 | X = x)$$
Defining subgroups based on $\Delta(x) = \text{CATE}(x)$

- Assume we managed to estimate $\hat{\Delta}(x)$
  
  Perhaps, simply as $\hat{\Delta}(x) = \hat{E}(Y|T = 1, X = x) - \hat{E}(Y|T = 0, X = x)$

Individualized treatment regimen/rule (ITR)

$\hat{D}(x) = 1$ if $\hat{\Delta}(x) > \delta$, $\hat{D}(x) = 0$ if $\hat{\Delta}(x) < -\delta$, otherwise treat randomly

$\hat{S}(x) = \{x: \hat{\Delta}(x) > \delta\}$

e.g. for $\delta = 0$

$\hat{S}(x)$ is learned from data as a biomarker signature:

e.g. $\{x : X_1 > c_1 \& X_2 = c_2\}$

May not ensure that each individual $\hat{\Delta}(x_i) > \delta$, e.g. $E\{\hat{\Delta}(x)\} > \delta$, for $x \in \hat{S}(x)$
Literature on subgroup identification is diverse

Selecting Optimal Subgroups for Treatment Using Many Covariates
Tylor J. VanderWeele,a Alex R. Luedtke,a Mark J. van der Laan,a and Ronald C. Kesslerb

Abstract. We consider the problem of selecting the optimal subgroup to treat when data on covariates are available from a randomized trial or observational study. We distinguish between four different settings including: (1) treatment selection when resources are constrained, (2) treatment selection in the presence of side effects and costs, and (3) treatment selection subject to mandatory effect homogeneity. We show that, in each of these cases, the optimal treatment selection rule involves treating those for whom the predicted mean difference in outcomes comparing those with versus without treatment, conditional on covariates, exceeds a certain threshold. The threshold varies across these four scenarios, but the form of the optimal treatment selection rule does not.

Keywords: Effect modification; Interaction; Optimal treatment selection; Precision medicine; Personalized treatment; Randomized trial; Subgroup

(Epidemiology 2019;30: 334-341)

\[ \hat{S}(x) = \{ x : \Delta(x) > \delta \} \]

CAPITAL: Optimal Subgroup Identification via Constrained Policy Tree Search
Hongzhi Cui1, Wenbin Lu1, HaoChuan Peng1, Devan V. Mehotra2, and Lingxiang Huang1

1Department of Biostatistics, Case Western Reserve University
2Department of Statistics, North Carolina State University

Abstract: Personalized medicine, a paradigm of medicine tailored to a patient's characteristics, is an increasingly attractive field in healthcare. An important goal of personalized medicine is to identify a subgroup of patients, based on baseline covariates, that benefits more from the targeted treatment than from other comparative treatments. Most of the current subgroup identification methods often form an optimal group with an undersized treatment effect without paying attention to subgroup size. Yet, a clinically meaningful subgroup learning approach should identify the maximum number of patients who can benefit from the better treatment. In this paper, we present an optimal subgroup selection rule (CAPITAL) that maximizes the number of selected patients and in the meantime, achieves the pre-specified clinically meaningful mean outcome, as well as the average treatment effect. We derive two approximation theoretical forms of the optimal MATE based on theoretical framework that describes the treatment outcome interaction in the outcome. We further propose a Constrained Policy Tree search algorithm (CAPITAL) for the optimal MATE within the computationally efficient tree class. The proposed method is flexible to handle multiple covariates that capture the features of patients with various treatment effects, and to address time to event data using the motivated mean survival time as the clinically interesting mean outcome. Extensive simulation, comparison studies, and real data applications are conducted to demonstrate the validity and utility of our method.

Optimal subgroup selection
Henry W. J. Reeve, Timothy I. Caminges and Richard J. Samworth
University of Bristol, University of Edinburgh and University of Cambridge

Abstract. In clinical trials and other applications, we often know regions of the feature space that appear to exhibit interesting behaviour, but it is unclear whether those observed phenomena are reflected at the population level. Focusing on a regression setting, we consider the subgroup selection challenge of identifying a region of the feature space on which the regression function exceeds a pre-determined threshold. We formulate the problem as one of constrained optimisation, where we seek a low-complexity data dependent selection set on which, with a guaranteed probability, the regression function is uniformly at least as large as the threshold. Subject to this constraint, we would like the region to contain as much mass under the marginal feature distribution as possible. This leads to a natural notion of regret, and our main contribution is to determine the minimax optimal rate for this regret in both the sample size and the Type I error probability. The rate involves a delicate interplay between parameters that control the smoothness of the regression function, as well as arguments that quantify the extent to which the optimal selection set at the population level can be approximated by families of well-behaved subsets. Finally, we expand the scope of our previous results by illustrating how they may be generalised to a treatment and control setting, where interest lies in the heterogeneous treatment effect.

\[ \text{the level } \tau \text{ on } B \]

The p-values are then combined via Holm’s procedure (Holm, 1979) to identify a finite union of hyper-cubes that satisfy our Type I error control property. Our final selection set \( \hat{A} \) maximises the empirical measure among all elements of \( A \) that lie within this finite union of hyper-cubes.
What to look for in papers on Subgroup Identification?
The number of predictors the procedure can handle

• $p=1$
  – focus on selecting a cutoff for a single continuous biomarker (e.g. STEPP method by Bonetti and Gelber, 2000; Han et al, 2021)

• $p \approx 10-20$

• $p \approx 100-1000$

• $p \gg n$
  – Feature space grows with sample size
• What is the complexity of the “model space” where the subgroups reside?
  – Subgroups defined based on “black box” functions of covariates,
    \( \hat{S}(x) = \{x: \hat{\Delta}(x) > c\} \)
  – Subgroups defined by simple biomarker signatures with up to 2 variables,
    \( \hat{S}(x) = \{x: X_1 \leq c_1, X_3 > c_3\} \)

• How is model complexity controlled to prevent data overfitting?
Does it apply only to RCT or to OS as well?

• For observational data, there is an interplay between confounders and modifiers of treatment effect, making model selection more challenging
  – Confounders are predictive of both treatment $T$ and outcome $Y$
  – Effect modifies are predictive of CATE, $\Delta(x)$
What output does the method produce?

- Individualized treatment contrast, $\widehat{\Delta}(x)$
- Signatures of promising subgroups, $\widehat{S}(x) = \{x: X_1 \leq c_1, X_3 > c_3\}$
- Optimal treatment assignment rule $\widehat{D}(x) = 1$ if $\widehat{\Delta}(x) > \delta$, otherwise $\widehat{D}(x) = 0$
- Predictive biomarkers (a.k.a. effect modifiers ordered) by variable importance score.
What inference is done, if at all?

- Inference on $\Delta(x)$
  - Pointwise CI for random forests (Wager and Athey, 2018), CI for $\Delta(x)$ estimated from LASSO (Ballarini et al, 2018),
  - Simultaneous bands on $\Delta(x)$ from semiparametrics (Guo et al., 2021)

- Inference on certain features of $\Delta(x)$
  - Testing for presence of treatment effect heterogeneity (via latent mixtures, Shen and He, 2015) or
  - Machine learning methods with cross-fitting (Chernozhukov, 2019)

- Controlling the probability of selecting the right subgroups, $\hat{S}(x)$ vs $S_{true}(x)$
  - Bayesian credible intervals, $\Pr(\hat{S}_{lower} \subseteq S_{true} \subseteq \hat{S}_{upper}) > 1 - \alpha$ (Schnell et al, 2018)
What inference is done, if at all (cont.)?

• Estimating “honest effect” in selected subgroup $\hat{S}(x)$
  – Using bootstrap correction for optimism bias (Foster et al, 2011; Guo and He, 2020)
  – Bayesian model averaging (Bornkamp et al, 2017)

• Inference on ITR, $D(X)$
  – Evaluating expected benefits if the regimen (rule) were applied to all patients,
  – Value = $E\{Y(\hat{D}(X))\}$ contrasted with the value of “always treat” or other strategy

• Controlling the False Discovery Rate
  – E.g., for selection of predictive biomarkers (Wei et al, 2021; Sechidis et al, 2021)
Typology of Subgroup Identification; Lipkovich et al. (2017)

Global outcome modeling: \( Y \)

- \( m(t, x) \)
- \( m(1, x) \)
- \( m(0, x) \)

Local treatment effect modeling: Subgroup search

- Enhanced treatment effect for drug A
- \( \delta = 0 \)
- \( \Delta(x) \)

Direct treatment effect modeling

- \( \delta = 0 \)
- \( \Delta(x) \)

Individual treatment regimen modeling: \( \text{sign}\{\Delta(x)\} \)

- Prescribe A
- Prescribe B
- Enhanced treatment effect for drug A

Local treatment effect modeling: Subgroup search
Global outcome modeling

A multi-stage (multi-model) process termed *meta-learning*

As a precursor, see Virtual Twins (VT) by Foster et al (2011)

- **T-(two) learning:**
  - Fit $m(t, x) = E(Y|T = t, X = x)$, separately by arms
  - Compute $\Delta(x) = \hat{m}(1, x) - \hat{m}(0, x)$

- **S-(single) learning:**
  - Fit $m(t, x) = E(Y|T = t, X = x)$, in pooled data with $X^*T$ interactions added
  - Compute $\Delta(x) = \hat{m}(1, x) - \hat{m}(0, x)$

- **X-learning** based on two version of CATE
  - $\hat{\Delta}_1(x)$ by modeling $Y(1) - \hat{m}(0, x)$, on treated subjects
  - $\hat{\Delta}_0(x)$ by modeling $\hat{m}(1, x) - Y(0)$, on control subjects
  - Compute $\Delta(x) = \hat{\Delta}_0(x)\pi(x) + \hat{\Delta}_1(x)(1 - \pi(x))$

- **Regularization challenges when modeling CATE**
  - Separate penalties for prognostic and predictive effects (Imai & Ratkovic, 2013)
  - Separate modeling of counterfactuals in X-learning (Künzel et al, 2019)
  - Separate penalties for prognostic and predictive effects in Bayesian causal forests (Hahn et al, 2020)
Direct treatment effect modeling

Directly evaluates $\Delta(x)$ obviating estimating main effects $h(x)$

- Adopt any tree-based method by modifying splitting criterion
  - Interaction trees, e.g. Su et al (2009) maximizing at every split $\left(\hat{\Delta}_{left} - \hat{\Delta}_{right}\right)^2$
  - Causal trees and causal forests (Athey and Imbens, 2016; Wager and Athey, 2018)
    - Local non-parametric estimates of $\Delta(x)$ by averaging treatment effects from terminal nodes across trees
    - “Honest trees”: divide data into two halves, use one for splitting and the second for estimating $\Delta(x)$
    - Inference for random forests (Efron, 2013 and Wager et al. 2014)
- Modified outcome and covariate (functions) methods
  - Tian et al. (2014) and Chen et al. (2017), see next
A broad framework for directly estimating $\Delta(x)$ for different types of outcomes/loss functions (R package \textit{personalized})

$$- A = 2T - 1 \in \{-1,1\}, \pi(x) = \Pr(T = 1|X = x), \pi(A|x) = A\pi(x) + \frac{1-A}{2}$$

\textbf{MOM}

$$E \left( \left( \frac{AY}{\pi(A|x)} - g(x) \right)^2 \right| X = x) \rightarrow \min$$ returns $g(x) = \Delta(x)$,

$$E \left( \frac{1}{\pi(A|x)} \left( 2AY - g(x) \right)^2 \right| X = x$$ has the same estimand $g(x) = \Delta(x)$, and so is

\textbf{MCM}

$$E \left( \frac{4}{\pi(A|x)} \left( Y - \frac{A}{2} g(x) \right)^2 \right| X = x$$, \textit{W-learning} in Chen et al. (2017)

- Choosing different loss functions allows for different outcomes types
- Options for modeling $g(x)$: linear (e.g. via penalized regression) reduces it to multiplying each covariate by $A/2$ (\textit{modified covariate}), gradient boosting, ...
Treatment effect modeling: R-learning

• R-learning for estimation of $\Delta(x)$ (Zhao et al, 2018; Nie and Wager, 2021; inspired by Robinson’s transformation and Double/Debiased Machine Learning of Chernozhukov, 2017)

• Note, $\Delta(x) = E \left( \frac{Y - m(x)}{T - \pi(x)} \right)$, where $m(x) = E(Y|X = x)$

\[
\tilde{\Delta}(\cdot) = \arg\min_{\Delta} \frac{1}{N} \sum_{i=1}^{N} [Y_i - m(x_i) - \{T_i - \pi(x_i)\} \Delta(x_i)]^2 + \Lambda_N\{\Delta(\cdot)\}
\]

– Prognostic effects and propensity (for non-randomized trials) need to be estimated at first step, but the focus is placed on the target $\Delta(x)$

– $m(x_i)$ and $\pi(x_i)$ are estimated from off-the-shelf ML methods and their cross-fitted versions are plugged-in $\tilde{m}^{-i}(x_i)$ and $\tilde{\pi}^{-i}(x_i)$
Software for subgroup identification

SIDES method
R package SIDES: implementing the regular SIDES method (Subgroup Identification Based on Different Effects Scores), by Ilyas S. Koppikar et al. (2017). The package is hosted at https://github.com/sides-tutorial.

The SIDES package includes an Excel add-in which implements the regular SIDES and SIDES-0 mixture methods (last update: March 25, 2018). The package is maintained by Ilyas S. Koppikar (ilyas.koppikar@gmail.com).

FindIt method
FindIt package (for finding treatment effects) by Ilyas S. Koppikar et al. (2017). The package is hosted at https://github.com/sides-tutorial.

Blasso method
Download the R functions and examples for the Blasso Trees method (last update: December 30, 2014). The code and examples are provided by Peng Jiang (pengjiang@umich.edu).

Virtual Trees method
Download the R code for the Virtual Trees method (last update: December 30, 2014). The code is provided by Jared Huling (jared.huling@umich.edu).

GUIDE method
GUIDE package for classification and regression trees now includes methods for subgroup identification. The GUIDE package is maintained by Wei-Yin Loh (Wei-Yin.Loh@ucdavis.edu). For more information on the subgroup identification features, see Section 3.13 of the GUIDE User Manual (last update: September 25, 2018) and papers by Wei-Yin Loh, Xia He, and Michael Tan.

In addition, MSGUIDE package implements the GUIDE method for randomized trials and observational studies.

QUINT method
QUINT package for Qualitative Interaction Trees. The package is maintained by Elke Dusseldorp (Elke.Dusseldorp@live.com) and colleagues. Reference: Dusseldorp and Machmuller (2014).

FindIt method

Blasso method
Download the R functions and examples for the Blasso Trees method (last update: December 30, 2014). The code and examples are provided by Peng Jiang (pengjiang@umich.edu).

Virtual Trees method
Download the R code for the Virtual Trees method (last update: December 30, 2014). The code is provided by Jared Huling (jared.huling@umich.edu).

GUIDE method
GUIDE package for classification and regression trees now includes methods for subgroup identification. The GUIDE package is maintained by Wei-Yin Loh (Wei-Yin.Loh@ucdavis.edu). For more information on the subgroup identification features, see Section 3.13 of the GUIDE User Manual (last update: September 25, 2018) and papers by Wei-Yin Loh, Xia He, and Michael Tan.

In addition, MSGUIDE package implements the GUIDE method for randomized trials and observational studies.

QUINT method
QUINT package for Qualitative Interaction Trees. The package is maintained by Elke Dusseldorp (Elke.Dusseldorp@live.com) and colleagues. Reference: Dusseldorp and Machmuller (2014).

FindIt method

Blasso method
Download the R functions and examples for the Blasso Trees method (last update: December 30, 2014). The code is provided by Peng Jiang (pengjiang@umich.edu). Reference: Gu, Yin, and Wei (2013).

ROWSI method
Download the R code for the ROWSI method (Regularized Outcome Weighted Subgroup Identification) by Chen et al. (2015).

Model-based Recursive Partitioning
R package party: A Toolbox for Recursive Partitioning, which can perform subgroup analyses using the functions intpart, glmtree (or more generally, mob) and treebag.

Recently a new package modelR has been created that specializes on stratified and personalized treatment effect estimation. The package is maintained by Heidi Seibold (heidiseibold@gmail.com).

See examples of subgroup analysis in Seibold et al. (2015) and Seibold et al. (2016).

Other packages
R package personalr (maintained by Jared Huling) for subgroup identification and estimation of heterogeneous treatment effects. It is a general framework that encompasses a wide range of methods including ROWSI, outcome-weighted learning, and many other new methods. See documentation and article explaining the underlying methodology.

R package SubGID implements several algorithms for developing threshold-based multivariate (predictive/prognostic) biomarker signatures via bootstrapping and aggregating of thresholds from trees (BART, Monte-Carlo variations of the Adaptive Indexing Method (AIM)) by Huang X. et al. (2017) and of adaptation of Patient Rule Induction Method (PRIM) for subgroup identification by Chen G. et al. (2015).

Fu, Zhou and Faries (2016) developed a search approach that provides simple and interpretable rules defining subgroup of patients with maximizes average patients' benefit for different treatments within a general framework of outcome weighted learning (OWL). Here you can find the C++ implementation.

R package DynaRegime implements methods to estimate dynamic treatment regimes using interactive Q-Learning, Q-Learning, weighted learning, and value-search methods based on Augmented Inverse Probability Weighted Estimators and Augmented Inverse Probability Weighted Estimators.

R package histconv constructs list-based rules (lists of if-then clauses) to estimate the optimal dynamic treatment regime based on the approach by Zhang et al. (2016).

The submit R package implements method for bootstrap-correction estimation after subgroup selection described in RosvallKrainsk (2016) and a model averaging approach from Bork Kempf et al. (2016).

TIDT: Treatment-Specific Subgroup Detection ToolBox by Chakib Batiou, Brian Denton and Li Shen (2010).

StratifiedMedicine by Thomas Jernvall is a broad toolkit for subgroup identification and stratified precision medicine. The package also includes a novel algorithm PRISMA (Patient Response Identifiers for Stratified Medicine) by Jernvall and Hiekkonen (to appear).

Generalized Random Forests (grf) is a package for forest-based statistical estimation and inference. The package currently provides methods for non-parametric least-squares regression, quantile regression, survival regression and treatment effect estimation (optionally using instrumental variables), with support for missing values.

Policy learning via doubly robust empirical welfare maximization over trees (gprk) extends policy learning via doubly robust empirical welfare maximization over trees. This package implements the multi-action doubly robust approach of Zhou, Athey and Wager (2018).

R package (delsidewegroup) implements bootstrap-assisted descriptivalized Lasso and bootstrap-assisted R-sq split estimators on selected subgroup's treatment effect estimation. The implemented estimators remove the subgroup selection bias and the regularization bias induced by high-dimensional covariates. For more information, see Guo, Wei, Wu, and Wang (2017).

R package (learner) supports quasi-convex estimation of heterogeneous treatment effects based on Nie and Wager (2021).

R package (causalTree) is available to estimate causal interactions using generalized treatment effects using machine learning based on Kuhnel, Sekhon, Bickel and Yu (2019).

R code (CAVARisk) for the implementation of optimal subgroup identification via constrained policy tree search based on Cal. Li, West, Mehmetra and Huang (2021).

R package (Sebb) supports causal inference for a binary treatment and continuous outcome using Bayesian causal forests based on Hahn, Murray and Carvalho (2019).
Simulation example: A single data set

- \( N=1000 \), randomization 3:1, \( T \in \{0,1\} \)
- \( X_1, X_3, X_4 \sim N(0,1), X_2 \in \{1,2,3\} \) with \( p = 1/3 \)
- \( Y = 100 - (X_1 + 5X_2) + T\{g_1(X_3) + g_2(X_4)\} + \varepsilon, \varepsilon \sim N(0,1) \)
- \( g_1(x) = a - b(x - 0.5)^2, 0 \leq x \leq 1 \)
- \( g_2(x) = \frac{c}{1 + e^{-(x-0.5)}} \), \( 0 \leq x \leq 1 \); else \( g_2(X) = 0 \)
- Add 16 noise biomarkers \( X_5, \ldots, X_{20} \sim N(0,1) \) to the analysis data set
Treatment effect: $\Delta(x) = E(Y(1) - Y(0)|x)$

Overall ATE: $\Delta = 0.21$

TE within subgroup:
$E(\Delta(x)|\Delta(x) > 0) = 0.45$

Subgroup size:
$E\{I(\Delta(x) > 0)\} = 0.69$
Methods

- **T, S, X - learning**
- **Causal forest (CF)**
- **Modified Outcome method (MOM) with RF and Xgboost**
- **R - learning**

- For each method, identify *subgroup signature* on fitting training data: 
  \[ \hat{S}(X) = \{ x : I(\Delta_{\text{train}}(x) > 0) \} \]

- Evaluate average treatment effect in identified subgroup \( \hat{S} \) on independent test data (n=10,000),
  \[ TE(\hat{S}) = E_x\{\Delta(X)|\Delta_{\text{train}}(X_{\text{test}}) > 0\} \]

- Compute subgroup utility index: Treatment effect per subject in overall population
  \[ \eta = TE(\hat{S}) \times \frac{n(\hat{S})}{n} \]
Estimating CATE with NO noise variables in the analysis set

- Regularization bias towards zero: largest in causal forest and smallest in T-learning
- Large variability: MOM (not shown) and T learning

\[
\eta = 0.28, 0.32, 0.27, 0.23, 0.27
\]

ATE = 0.21

Red line is smoothed predicted CATE, black line 45 degrees
Estimating CATE, with noise variables in the analysis set

- Regularization bias towards zero: largest in causal forest and smallest in T-learning
- Large variability: MOM (not shown) and T learning

\[ \eta \]

<table>
<thead>
<tr>
<th>( \eta )</th>
<th>0.24</th>
<th>0.25</th>
<th>0.28</th>
<th>0.24</th>
<th>0.26</th>
</tr>
</thead>
</table>

ATE=0.21

Red line is smoothed predicted CATE, black line 45 degrees
Summary

• A shift from ad-hoc “subgroup chasing” methods towards principled methods of personalized/precision medicine utilizing ideas from causal inference, machine learning and multiple testing emerged in last 10 years producing a vast number of diverse approaches.

• For naïve multistage methods (requiring fitting the response surface $m(t, x)$) regularization bias can be large, as each step is optimized for prediction, rather than for the final estimation target (Künzel et al, 2019; Chernozhukov, 2019; Nie and Wager, 2021).

• While methods that estimate $\Delta(x)$ obviating fitting main effects $h(x)$ are attractive, substantial efficiency can be gained by using doubly-robust methods, such as utilizing augmented inverse propensity weighted scores, even in the context of RCT where propensities are known (Athey and Wager, 2021; Kennedy, 2021).

• There is increasing interest in developing ITRs respecting constraints on costs, adverse events, sample size (Wang et al, 2018; Athey and Wager, 2021; Cai et al, 2021).

• There is a need in interpretable personalized solutions (ITR’s) within a pre-defined policy class, e.g tree-structured or boxes (Laber and Zhao, 2015; Cai et al, 2021; Doubleday et al., 2021).


Thank you!

Q & A

Ilya.Lipkovich@lilly.com
Direct estimation of CATE: A-learning

• It is easy to see that the following modified covariate function method has as population minimizer, $g(x) = \Delta(x)$

$$E \left( \left( Y - (T - \pi(x))g(x) \right)^2 \bigg| X = x \right) \rightarrow \text{min}$$

• Like with W-method, this generalizes to different types of outcomes by replacing squared loss with appropriate loss functions (See Chen et al., 2017)
  
Doubly robust estimators of CATE

• Combines estimators of expected PO’s with IPW (Athey and Wager, 2019; Kennedy, 2021)
• Consistent if at least PO or propensity model is correct
• Reduces variability compared with direct methods
• Let \( \hat{\Delta}(x) = \hat{m}(1, x) - \hat{m}(0, x) \) be an estimator of CATE from any metalearner, causal forest, etc

\[
\hat{\Delta}_{DR}(x) = \hat{\Delta}(x) + \frac{T - \hat{\pi}(x)}{\hat{\pi}(x)(1 - \hat{\pi}(x))} \left[ Y - \hat{m}(x) - (T - \hat{\pi}(x))\hat{\Delta}(x) \right]
\]

where, \( m(x) = E(Y|X = x) \)
Modeling ITRs (outcome weighted learning)

While ITR can be estimated based on methods of outcome modeling (1) or treatment effect modeling (2), some methods estimate directly the sign of $\Delta(x)$ by casting it as a classification problem (Zhao et al, 2012)

- One approach is to write the expected value of ITR 
  \[ E\{Y(D(X))\} = E\left[\frac{I(D(X)\neq T)Y}{Pr(T|x)}\right] \rightarrow \text{max} \]

- This is equivalent to minimizing weighted classification loss 
  \[ E\left[\frac{I(D(X)\neq T)Y}{Pr(T|x)}\right] \rightarrow \text{min} \]

- Minimizing 0-1 loss is an NP problem so typically we modify it using a smooth convex surrogate loss function. E.g hinge, or exponential loss: 
  \[ E[L_w(T, f(x))] \]

- This allows using off-the-shelf packages to identify ITRs, e.g. logistic regression with lasso penalty and weights 
  \[ w_i = \frac{Y_i}{Pr(T = t_i|X = x_i)} \]
Modeling ITRs: Recent advances

- Treatment allocation based on simultaneous confidence band estimated from semiparametric modeling of $\Delta(x)$ (Guo et al, 2021)
- Multi armed angle-based direct learning for ITR (Qi et al, 2020)
- Learning optimal ITR adopting risk/costs constraints (Wang et al, 2018)
- Risk controlled decision trees and random forests for precision medicine (Doubleday et al, 2021)
- Searching treatment policies within a restricted class of fixed depth trees. Uses doubly robust estimator of treatment effect function. Athey and Wager (2021), policytree R package (by Sverdrup et al.)
  - Extending work on maximizing empirical welfare (value) of policies within restricted classes from randomized studies by Kitagawa and Tetenev (2018).
  - Recent application/extension: CAPITAL: Optimal subgroup identification via constrained policy tree search (Cai et al, 2021)
Direct subgroup search (local treatment effect modeling)

• Instead of estimating the response function $\Delta(x)$ on the entire covariate space and then carving out segments, search directly for such regions

• Recent methods
  – SIDEScreen (Lipkovich and Dmitrienko, 2014)
  – Adaptation of PRIM method in Chen et al, 2015
  – Sequential-BATTing (Huang et al, 2017) implemented in R package SubgrlD