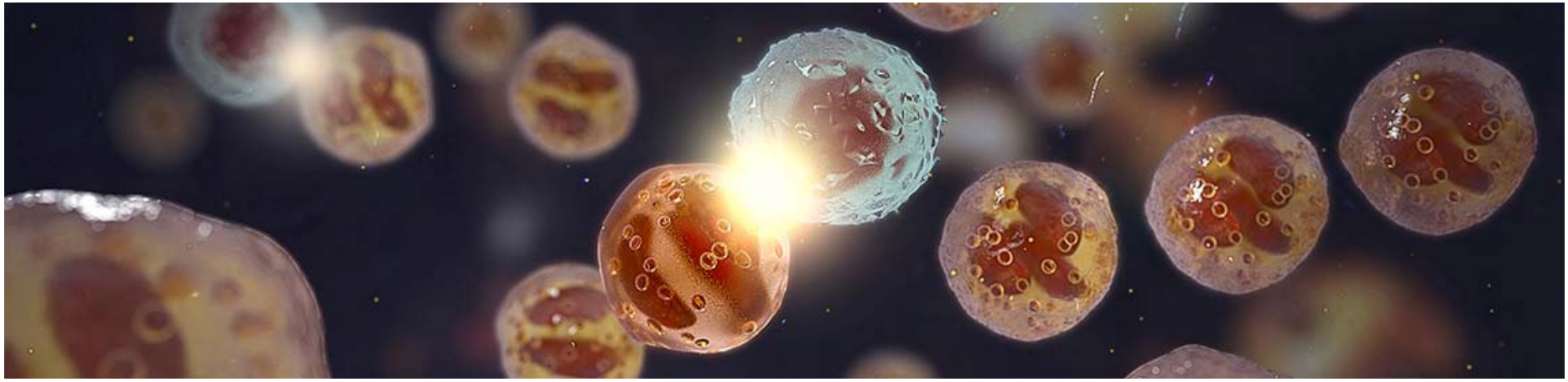
AstraZeneca

Recent Developments in Biomarkers and Subgroups in Drug Development - WORKSHOP 20 MARCH 2019

Overview of some Recent Methodologies for Biomarker Subgroup Identification

David Svensson[1] & Ilya Lipkovich[2]

[1]AstraZeneca (speaker). [2]Eli Lilly, [1,2]EFSPI Subgroups Special Interest Group.



Acknowledgements

Much of this has came out from discussions/collaborations with others, e.g., under the EFSPI Subgroups SIG umbrella.

In particular, with **Ilya Lipkovich** (Eli Lilly) – special thanks for lots of input.

Also, thanks to **Mattis Gottlow** (AZ, Advanced Analytics Centre).

(Any errors are due to David Svensson).

DISCLAIMER: the **opinions** expressed in this presentation are those of the **authors,** and do **not** necessarily reflect the official policy of AstraZeneca/Eli Lilly.

OVERVIEW

HTE: **H**etereogeneous **T**reatment **E**ffects= a reality …

Characterizing patients **responding better** to treatment:

- A **complex task** & a strong trend in the industry.

Obviously, it goes beyond mere computational aspects.

- But: the **computational aspects** are important too

Plenty of methods proposed in recent literature.

Is any method '**best'…** !? *A deceivingly simple-looking Q. And the answer is … [`wait-for-it`]*.

Most powerful method?

RCTs are seldomly sized for **D**ata-driven **S**ubgroup **D**etection (**DSD**). Given this underpower …

- *… it does make sense to think about good methods*…

But: we will look at the **plethora of recent methods**, and see why it is non-trivial to arrive at a robust answer…

Also, we will see that **DSD** isn't what we typically mean with '**M**achine **L**earning' (**ML**)... (but some similarities exist).

(A standard chapter on **supervised learning** won't help you).

It has to be mentioned: subgroups are tricky!

Well-known issues when interpreting subgroups, even if pre-specified.

– High **false positive risk**, **low power** of interaction tests [4a]

– **Biased** estimate in best-looking selected subgroup. [3a], [18a]

Guidelines & papers warn for this – for good reasons.

- E.g., not enough with *inference*, also biol. *plausibility*? Etc. [7a], [9a]

Other end of the spectrum:

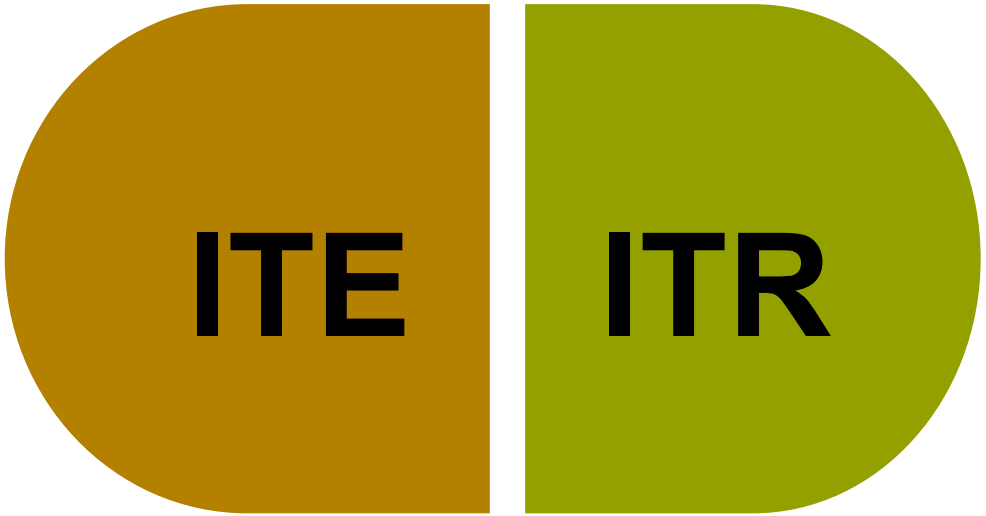
– the scientific spirit – full usage of the data – '*let it speak*'

– heterogeneuos diseases (e.g., oncology, diabetes)

The Right Patient, The Right Treatment, The Right Time

Two major HTE frameworks:

The modern literature can broadly be divided into



Individual Treatment Effect:
mapping a treatment to patients
(=*'find a subgroup'*).

Individual Treatment Rules:
mapping a patient to a treatment
(=*'find a treatment'*).

ITE (Individ.Treat. Est): setting

Parallel two-arm trial (Active vs Control).
y=endpoint, trt=randomisation, x=(x_1, …, x_p) baseline biomarkers, covariates.

Could be:

- Few baseline candidates? (say, 5?).
- Many baseline candidates? (e.g., 100+).
- Strict requirements: Subg.size>m and subg.effect>δ (some pre-set values).

Already here, somewhat of a cross-road:

"SUBGROUP DETECTION" (hence) can mean slightly different things:

- *identify key **treatment-interactions** ('narrowing it down')*
- *strict pre-defined scheme for **defining a subgroup** & estimating effect.*

Key feature: pre-specification & controlling size of search

- **Past: unstructured** (~ adhoc dreadging, unknown performance)
  - Low power with e.g., interaction testing
  - Can't easily detect multiple-biomarker signatures

- **Now: structured** (~ systematic, special case of model selection)
  - Stronger for discovering multiple-biomarker signatures
  - built-in complexity control + cross-validation.

Modern Subgroup Detection = **special case of Model Selection**.
(≠ data dreadging) [8a], [26b], [3a], [18a].

  - I.e., pre-specified of entire search scheme/strategy - structured approach!
  - Idea: Limiting search-space (in a trackable way).
  - In principle, allowing NULL assessments (weak t1e) – and reproducibility!

Two approaches to ITE:

The modern approaches for ITE comes **in two flavours**:

- **LOCAL**: zoom in on sub-area in covariate space, ignoring the rest.
  - ”*Give me a subgroup*” (E.g., {x_2>10 and smoker=yes}).

- **GLOBAL**: modelling y over the entire covariate space: $\delta(\mathbf{x}) = E(Y_{(1)} - Y_{(0)} | \mathbf{x})$.
  - ”*Give me a model for how Treat.Eff varies with x*”

This ”is” a specific patient

***Potential* outcome** on active (1), control (0)

Can’t observe both for a given patient!

I.e., Note the Causal Inference connection:
- for each patient, we only can observe either $Y_{(1)}$ or $Y_{(0)}$

The Modern Methods are often:

Typically:

- **TREE**-based (*CART-style, but with a twist*), or
- Regularized/penalized **REGRESSION** (*'lasso' style*).

TREE-based: e.g.,

- VirtualTwin [11b] (L)
- SIDES [24b] (L)
- GUIDE [27b] (G)
- QUINT [10b] (G)
- IT [39b] (G)
- MOB [44b] (GLMtree) (G)
- STIMA [7b] (hybrid) (G)
- BATTING/AIM/PRIM [19b] (L)
- mCART [34b], (L)
- RFIT [37b], MOBFOREST [20b] (G)
- CFOREST [42b] (G)
- mBART [6b] (L), bartMachine [21b] (G)

REGRESSION-based: e.g.,

- Lasso & Ridge [15b], GLMnet [14b] (G)
- Boosting [30b] (G)
- 'FindIT' (SVT+Lasso) [22b] (G)
- STIMA (hybrid) [7b] (G)

Tree-based ones ….

For better or worse,
correspond well to medical practice:
The **dichotomize** biomarkers

Therefore, explicitly suggest subgroups
(at least, single trees; ensambles more tricky)

TREE-based: e.g.,

- VirtualTwin [11b] (L)
- SIDES [24b] (L)
- GUIDE [27b] (G)
- QUINT [10b] (G)
- IT [39b] (G)
- MOB [44b] (GLMtree) (G)
- STIMA [7b] (hybrid) (G)
- BATTING/AIM/PRIM [19b] (L)
- mCART [34b], (L)
- RFIT [37b], MOBFOREST [20b] (G)
- CFOREST [42b] (G)
- mBART [6b] (L), bartMachine [21b] (G)

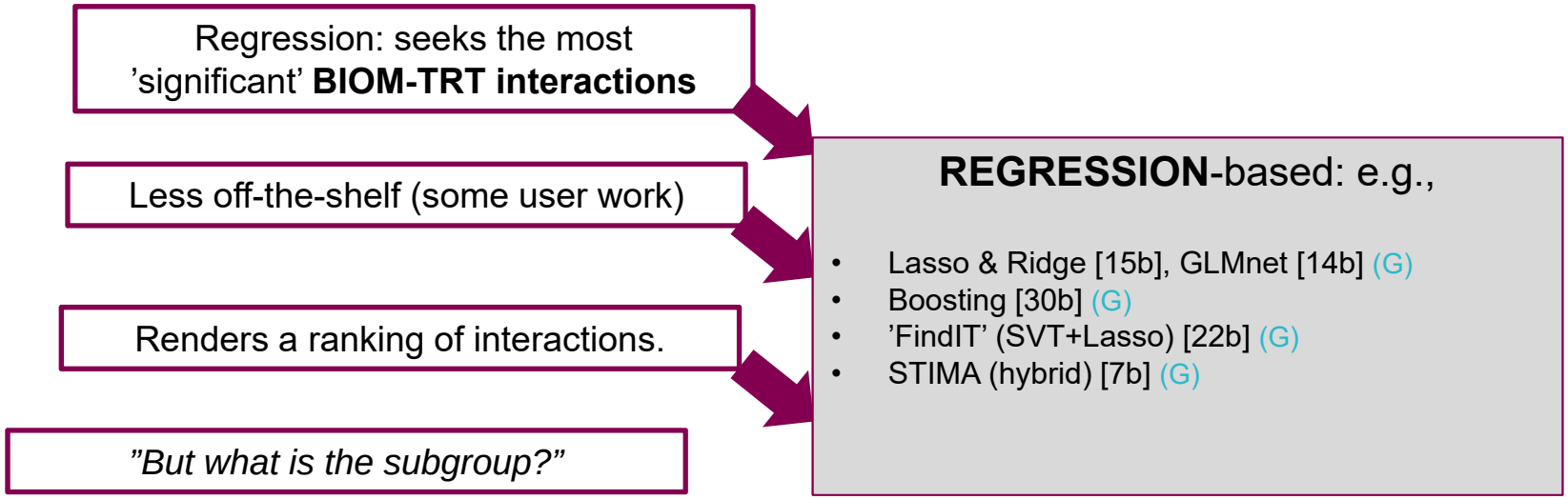
Typically 'off-the-shelf':
almost no data pre-processing

Might look misleadingly simple?

"Often it is still very hard to tell from a given tree structure whether interactions really exist and how variables interact with each other." [39b].

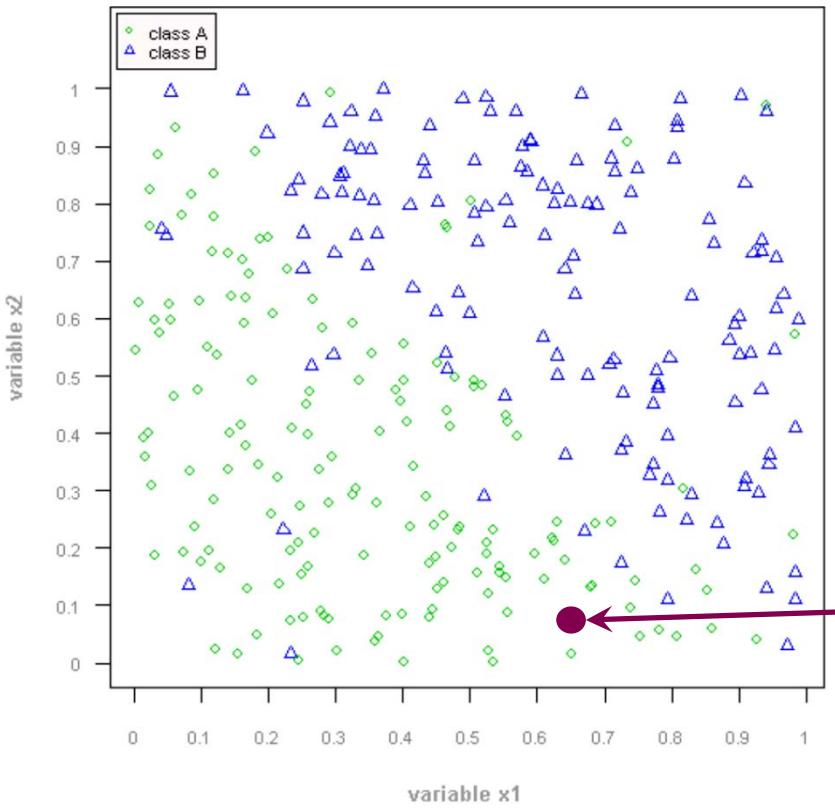
Regression-based ones …

The user must add the biomarker-treatment interactions as terms:
(p>>n no issue if regularized/penalized regr)



*'A natural question to ask at this point is, **how should one define subgroups** of patients who are likely to experience a beneficial treatment effect **based on penalized regression models**? One possible solution is to plot the estimated treatment contrasts against the covariates […] to identify reasonable cut-offs"* [26b]

Tree Crash Course (1)



Toy example:

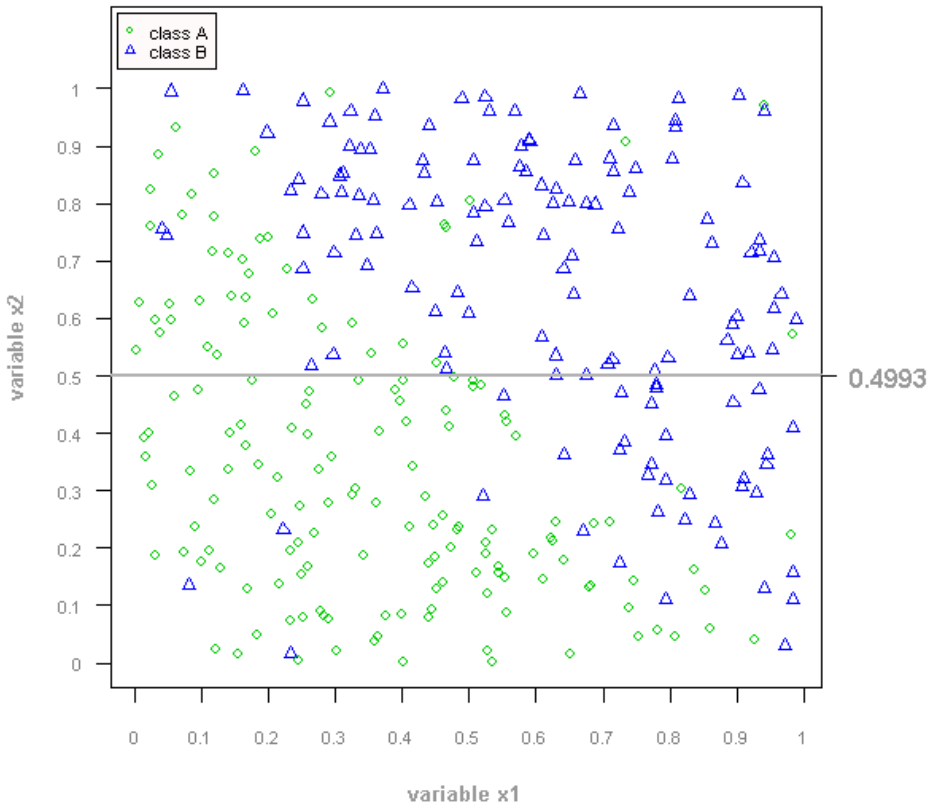
Two variables, x1 and x2

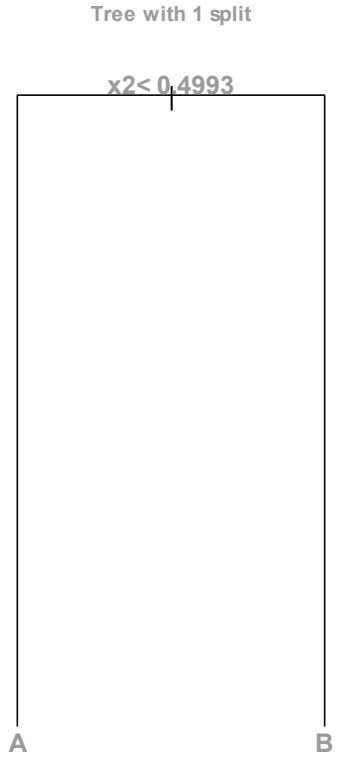
Each observation: 'A' or 'B'.

Can we fit a predictive model, and predict the class of a new observation?

Tree Crash Course (2)

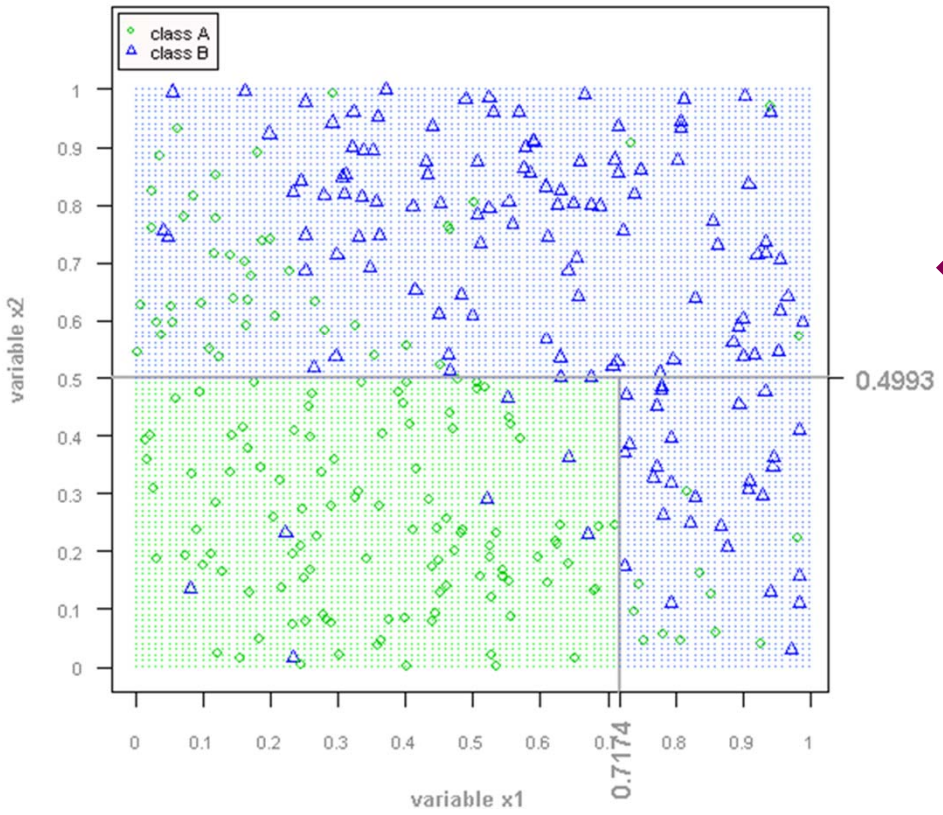
Recursive Partitioning: seeking homogeneous ’boxes’

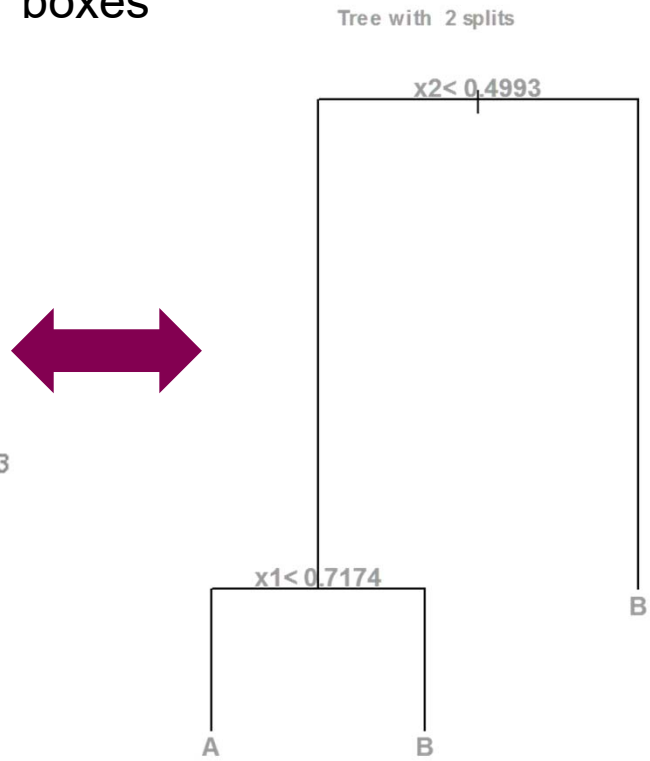




Tree Crash Course (3)

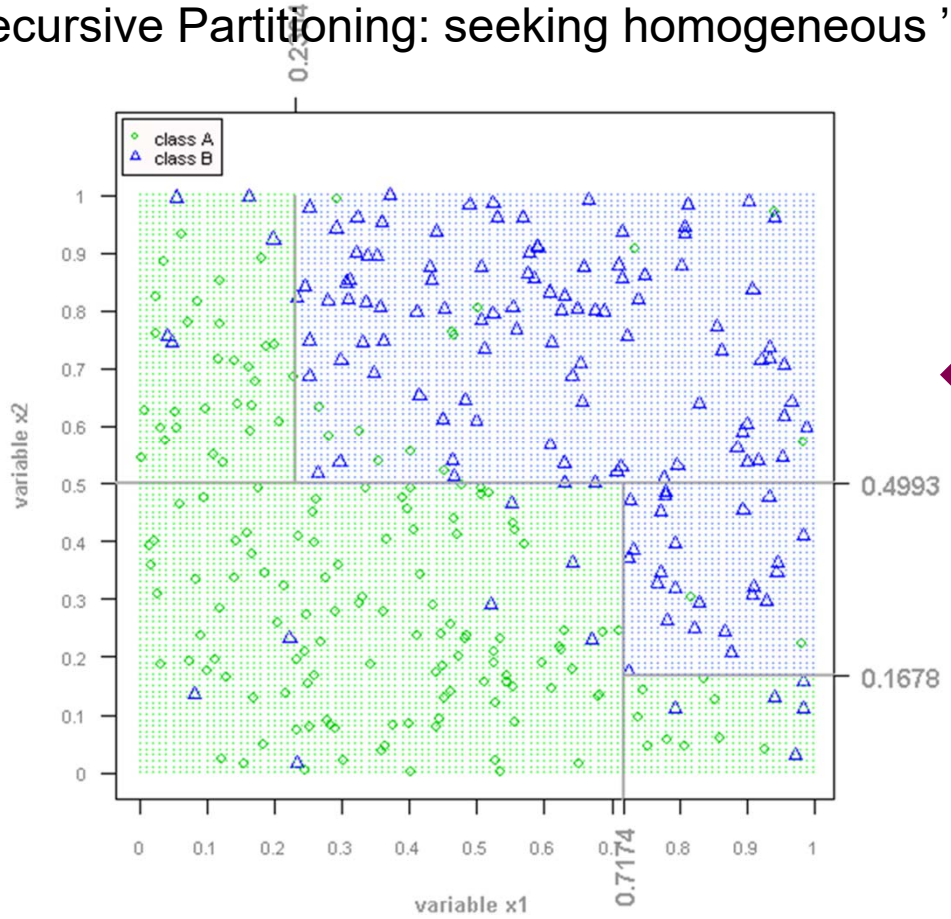
Recursive Partitioning: seeking homogeneous ’boxes’

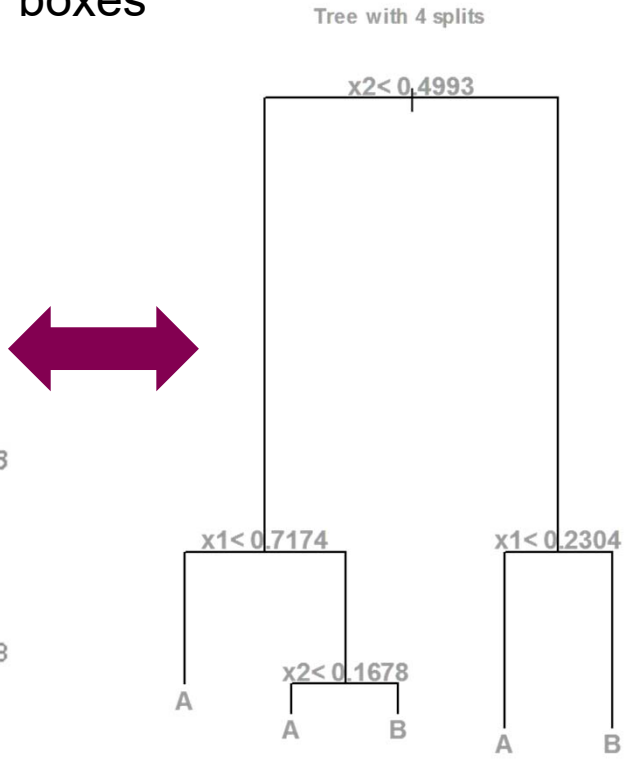




Tree Crash Course (4)

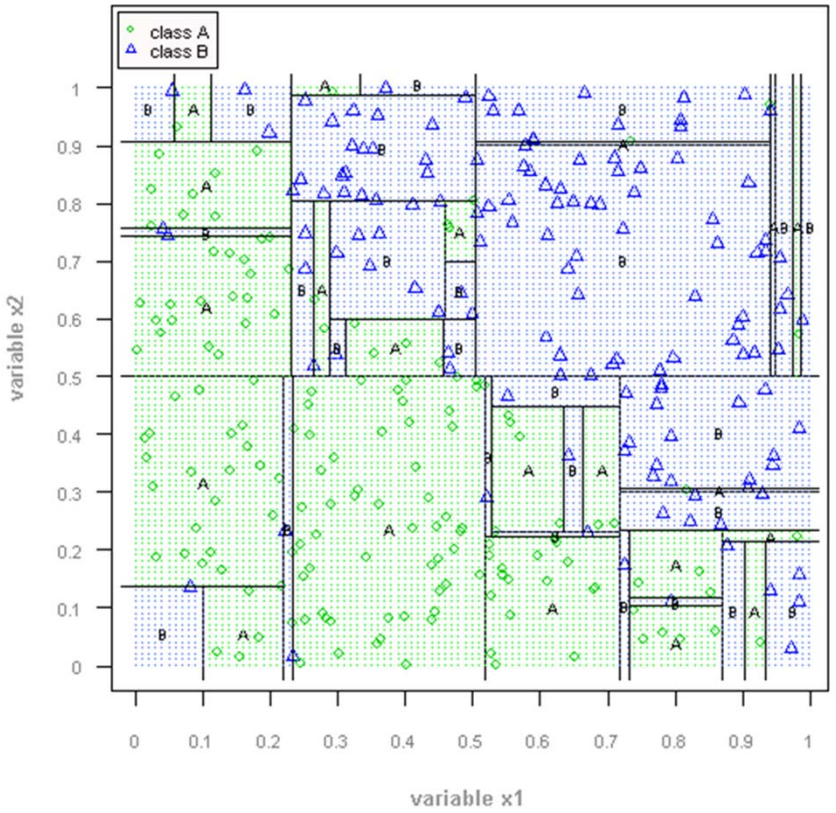
Recursive Partitioning: seeking homogeneous ’boxes’

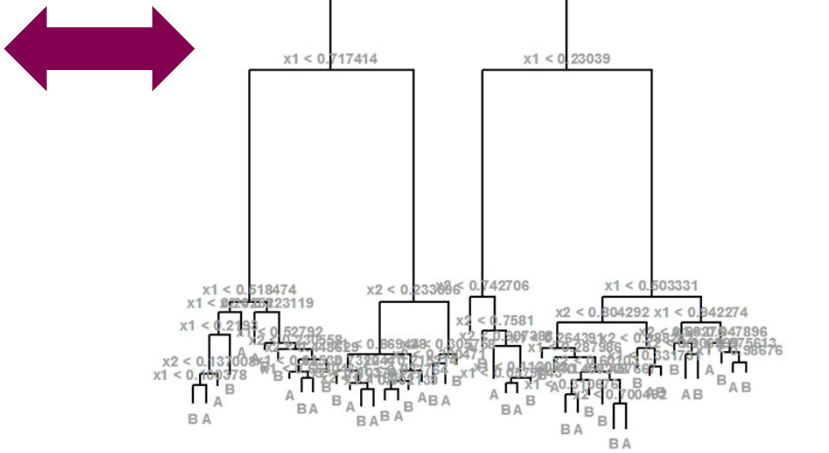




Tree Crash Course (5) (overfitted…)

Recursive Partitioning: seeking homogeneous ’boxes’





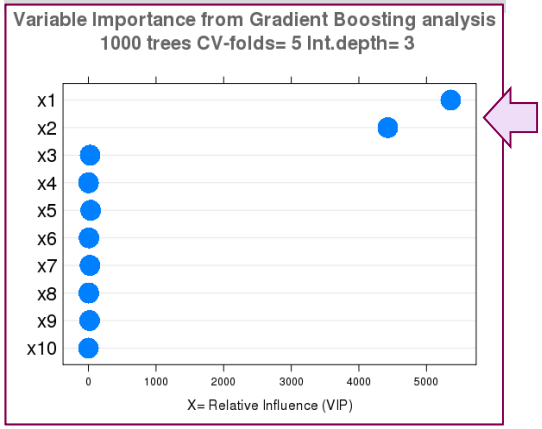
Super-quick: Machine Learning is strong .. (often trees)

Logistic Regression:

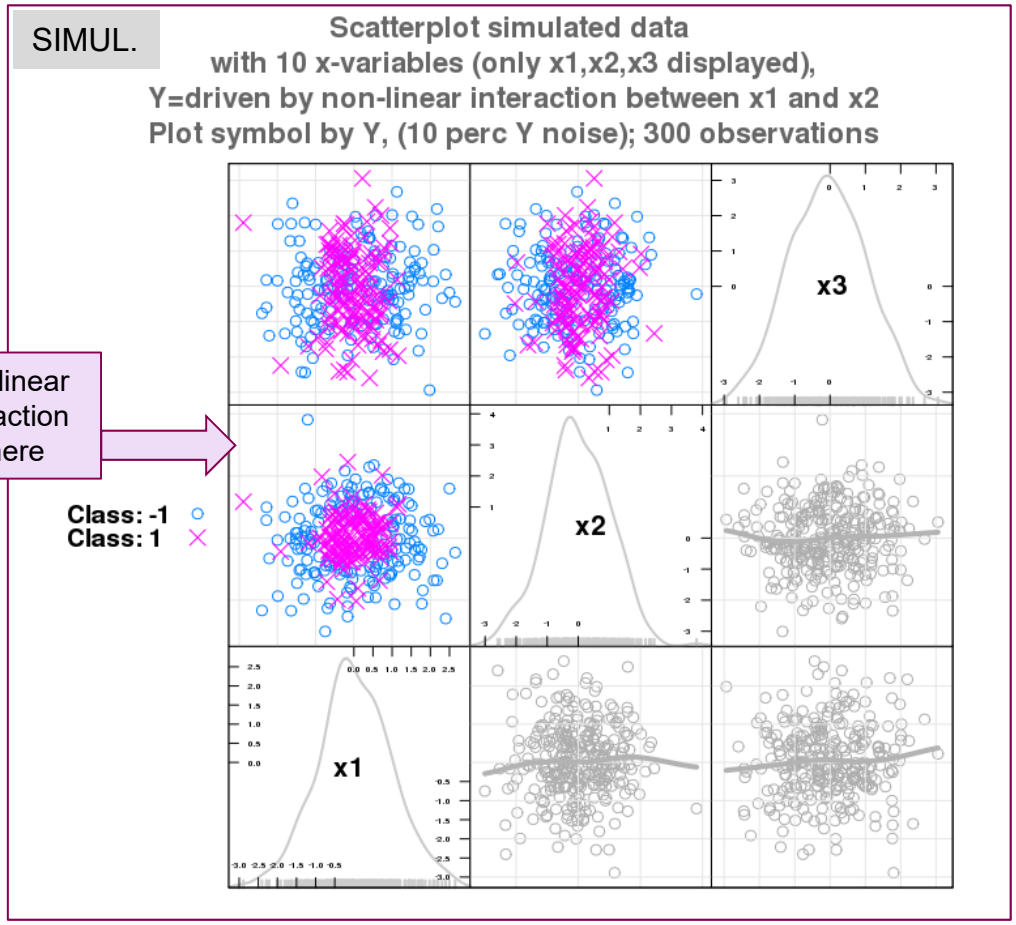
	Estimate	Pr(>|z|)
(Intercept)	-0.25	0.03
x1	-0.16	0.20
x3	-0.03	0.76
x2	0.10	0.38
x1:x2	-0.05	0.68
x1:x3	-0.12	0.28
x3:x2	0.03	0.81

Table 1: Exploratory GLM model fit

GRAD. BOOST



SIMUL.



But General Mining isn't what we are dealing with here:

The previous example illustrated the **general strength** of modern tree-based algorithms (e.g., Gradient Boosting, Random Forest) …

… but was *unrealistic and atypical*: there was no treatment involved.

KEY POINT:

Supervised Learning (Machine Learning)

– predicts y ~ **x,** ranks the x

Subgroup Detection:

– seeks **individual treat.contrasts ~ x**

– semi-supervised? Causal-inference element (i.e., incomplete data)

Can't we just run ML anyway … ?

Classical Supervised Learning (**Machine Learning** - ML) predicts y ~ **x**

- *Gradient Boosting, Random Forest, Elastic Nets, etc*
- *Support Vector Machines, Neural Nets, etc*

Of course, ML could be fitted as y ~ (**x, trt**) [i.e., trt as another column] but prognostic variables will then typically dominate!

Why? - *Because that's what prognostic variables do*: useful for predicting y!
(Easy to see via simulations – see later slides).

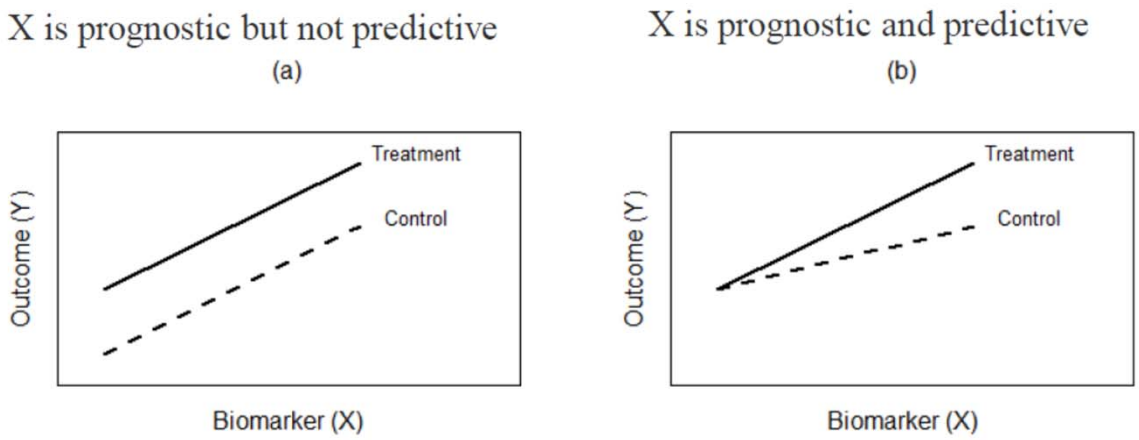
Subgroup Detection: surely has ML similarities, but has the *counterfactual* component/ focus treatment *contrasts*

- Explains *why so many novel methods (*despite decades ML research!).

(Prognostic vs Predictive terminology)

Just a short recap:

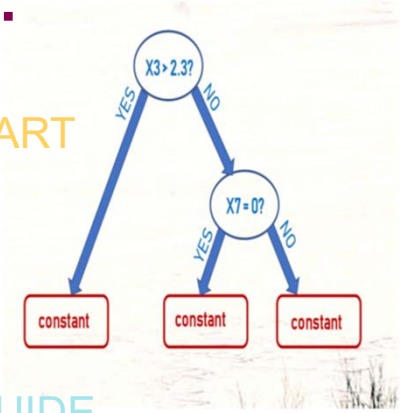
- X is “prognostic” if it predicts the outcome Y
- X is “predictive” if it predicts differential treatment effect



‘Classical Trees’ Vs ‘Subgroup-Detection-Trees’:

Classical Trees:
- Recursively splits the covariate space into rectangles
- And fits a **constant** within each

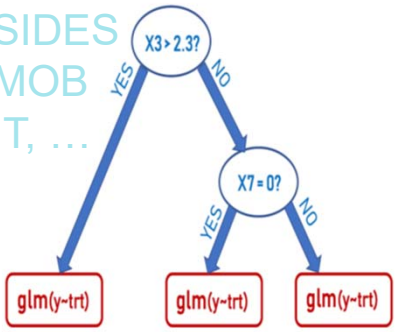
CART



Modern Subgroup-Detection-trees:
- Recursively splits the covariate space into rectangles
- And fits a **model*** within each

*Typically, fits some **GLM**(y ~ trt) - splitting process different across different methods

GUIDE
QUINT
SIDES
MOB
IT, …



I.e., fitting contrasts, ’heterogeneous style’ due to the different model fits across covariate space …

‘Subgroup-Detection-Trees’: how do they differ?

If we look at e.g., the tree-based methods: **how do they differ**?

“*Sometimes seemingly different methods developed by different groups of authors turn out to be **almost equivalent** to each other*” [26b]

“*it is important to realize that popular approaches to subgroup identification […]**come from such diverse fields** of research as machine/statistical learning, multiple testing, and causal inference*“. [26b]

‘Subgroup-Detection-Trees’: how do they differ? E.g.,

- - **QUINT** always aims at dividing the covariate region into three regions: treatment is ‘HARMFUL’, ‘NO DIFFERENCE’, or ‘BETTER’. (Qualitative interactions).
- - **SIDES:** modelling a local part of the covariate region under certain side-conditions, and lets the operator pre-limit the complexity of the resulting subgroup.
- - **IT** greedily seeks cut-offs *c* via likelihood ratio tests between a simpler model and a model with an indicator I(x>c) as term in the model. (Interactions)
- - **GUIDE** avoids seeking c directly; assesses globally most promising x to split on using lack-of-fit test, avoiding selection bias if covariates have differently no. unique values.
- - **MOB (GLM-tree)** uses a instability test regarding each biomarker x, before attempting to split - arguably the most complex of all these methods (see [43b])
- **- mCART:** uses propensity scoring for ‘straightening up’ imbalance of covariates.

New Trend: Ensembles of trees – (again!)

Classical Trees:
- Known to have high variance
- Research established: ensembles more powerful. [15b]

e.g., **RANDOM FOREST**, **GRADIENT BOOSTING** [4b], [31b]

Modern Subgroup-Detection-trees: Similar recent ensemble-trend:
- **CausalFOREST**, **RFIT**, **MOBFOREST**,
- **BARTMACHINE**, **VTGUIDE**
- Also, **VIRTUAL TWIN** based on ensemble-of-trees

Comes with a price: interpretability?
(Open reseach question: can improved model for $\delta(\mathbf{x}) = E(Y_1 - Y_0|\mathbf{x})$ be projected down to useful low-dimensional summeries (graphs)?

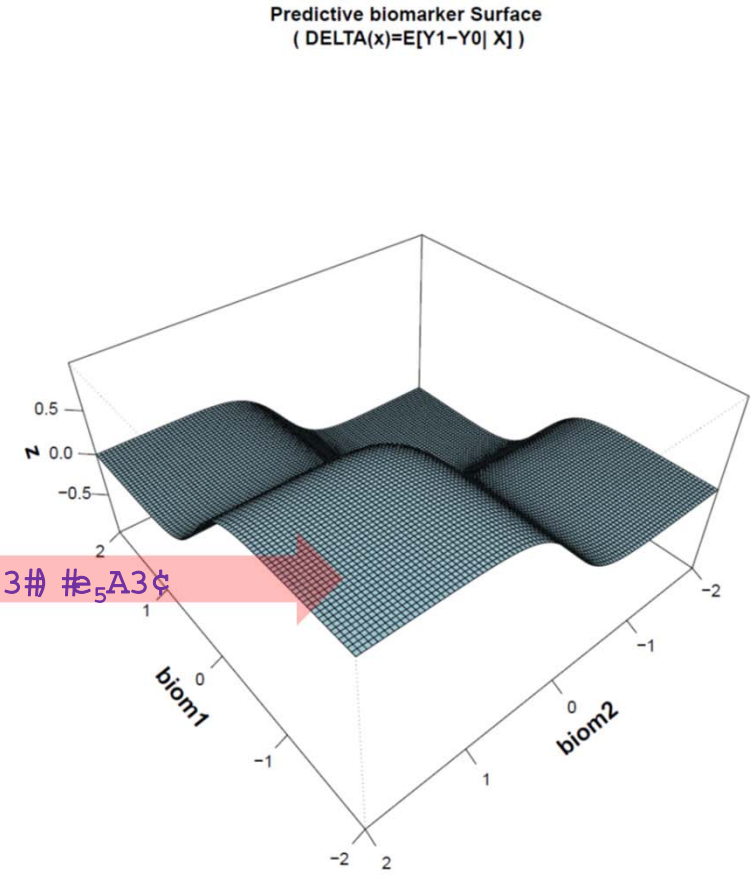
Let's see some action: simulated toy example

For illustration, sim RCT 1:1 rand,

- n.per.arm = 1000,
- Biomarkers: b_1, … b_4 ~ N(0,1), i.i.c
- **AGE** ~ unif(15,90).

Prognostic: **AGE**, b_3 and b_4

Predictive: b_1 and b_2




```
## Highlights from R sim. function:
## First generating X-over Trial, but then
## subsampling to a parallell RCT […]
progn <-  2 + 0.4*age + b3 + b4
delta.x <-  fun.logistic(b2, 5) - fun.logistic(b1, 5)
D0$y <-progn
D1$y <-progn + delta.x
D0$true.delta <- delta.x
D1$true.delta <- delta.x
DD.xo <- data.frame(rbind(D0, D1))
DD.xo$y <- DD.xo$y + rnorm(nrow(DD.xo), 0, Sd)
# each patient now factually & cfactually simulated,
## further code then assigns a patient to only a single trt…
```

Let's see some action: simulated toy example [2]

Strictly speaking a **non-linear biom~trt interaction** ('XOR' style).

And not obvious 'for the eye' due to the impact of prognostic variables.

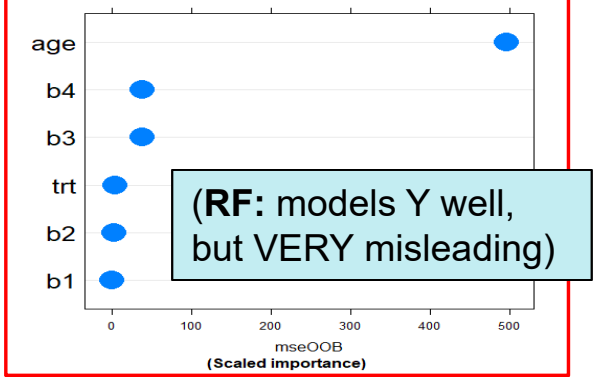
Let's run some methods:

- Udqgrp Iruhvw (pretending it is a standard ML problem)
- VWLP D, TXLQW, VLGHV, & PRE
- FdxvdoIruhvw& YlwkdoWzlq

(We stress that these runs are only for *illustrational* purposes and we admit that it is not obvious how to set up the various tuning parameters in an entirely fair fashion; details omitted here).

Results on Toy Data [1]

RandomForest:
Simply: Y~ {biomarkers,age and trt}.
(RF 98.85% Var.Explained).
Default Settings.
Variable Ranking below



(**RF:** models Y well, but VERY misleading)

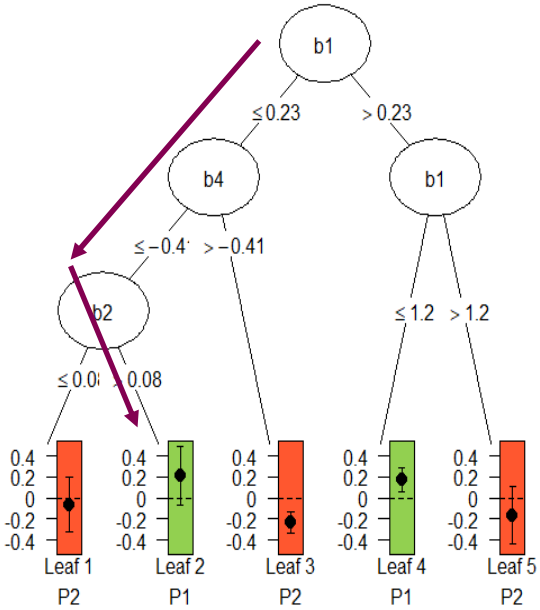
SIDES:

```
[…] Subgroup 2:
b1 <= 0.32945 b2 > -1.22852
Treatment effect: 1.21 (95%CI: 0.2-2.22)
 P-value 0.0096 (unadj[…])
```

- SIDES gave a shortlist of qualifying subgroups, most based on at least b1, all roughly relevant (we didn't impose any permutation-based NULL adj).

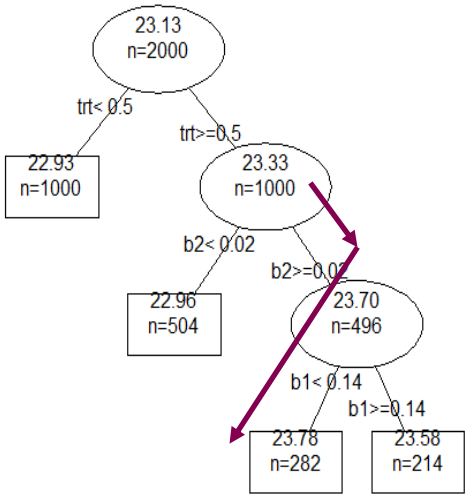
QUINT.

QUINT Default Setting GIVES:
'no clear qualtitative interaction present in the data.',
Relaxing tuning par dmin to zero gives these splits



STIMA.

STIMA tree splitting process:
(deciding which interaction terms to enter in the regr.model)



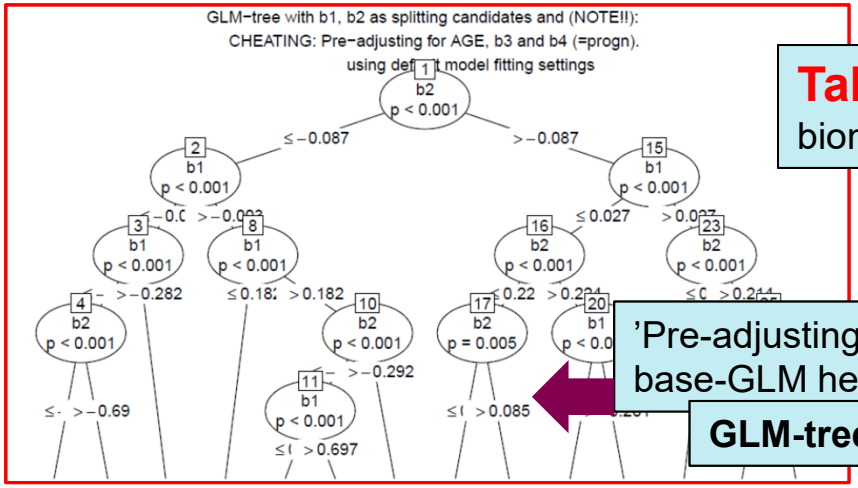
STIMA REGR adjusted for:
b3 (0.99), b4 (0.95) and age (0.40)
Correct signs on b1*trt (-0.59) and b2*trt (0.52)

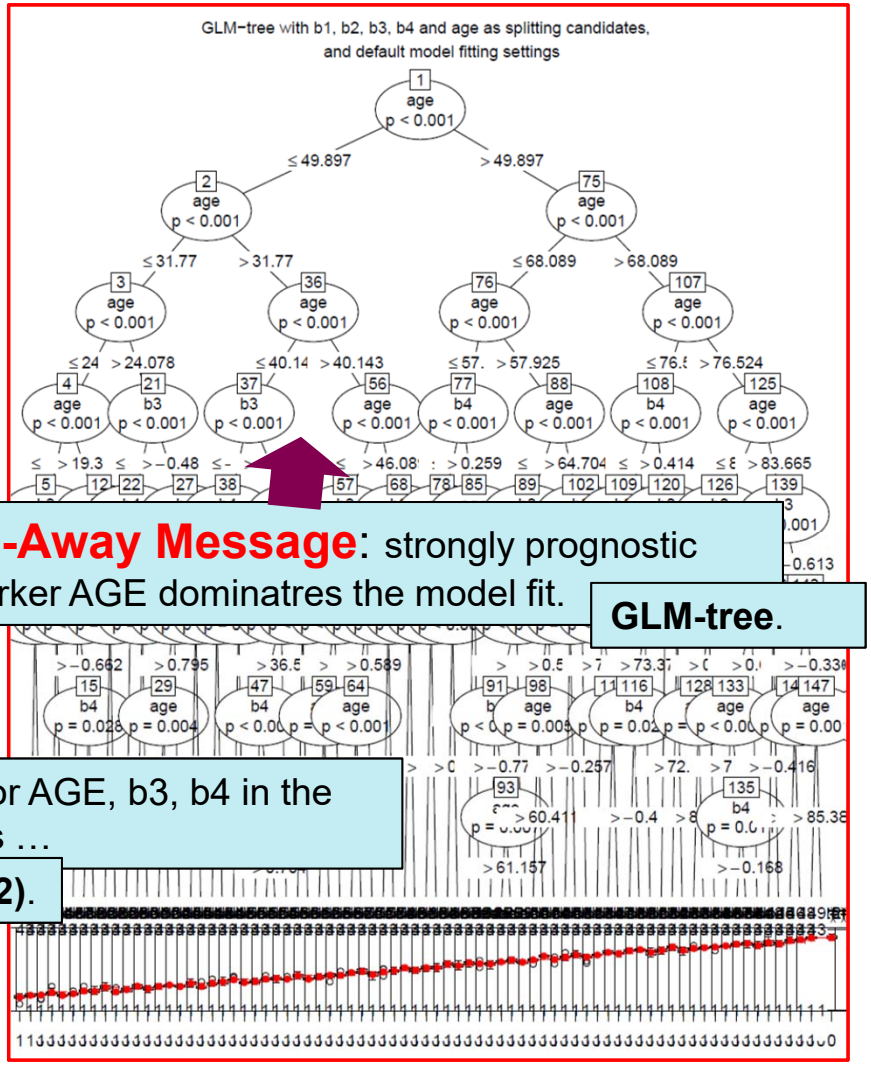
(*) This is not a manual and for illustrational purposes it suffices to say that all methods have some tuning parameters to set. This makes it non-trivial when comparing methods. The above runs are only inlcuded to reflect the kind of output the methods give, and give a flavour of how they picks up the signal in the data. (Beyond scope: t1e & bias assessments using NULL simulations)

Results on Toy Data [2]

- **RIGHT**: Biomarkers & Age = splitting candidates

- **BELOW**: Only b1 & b2 = splitting candidates,
- (cheating)





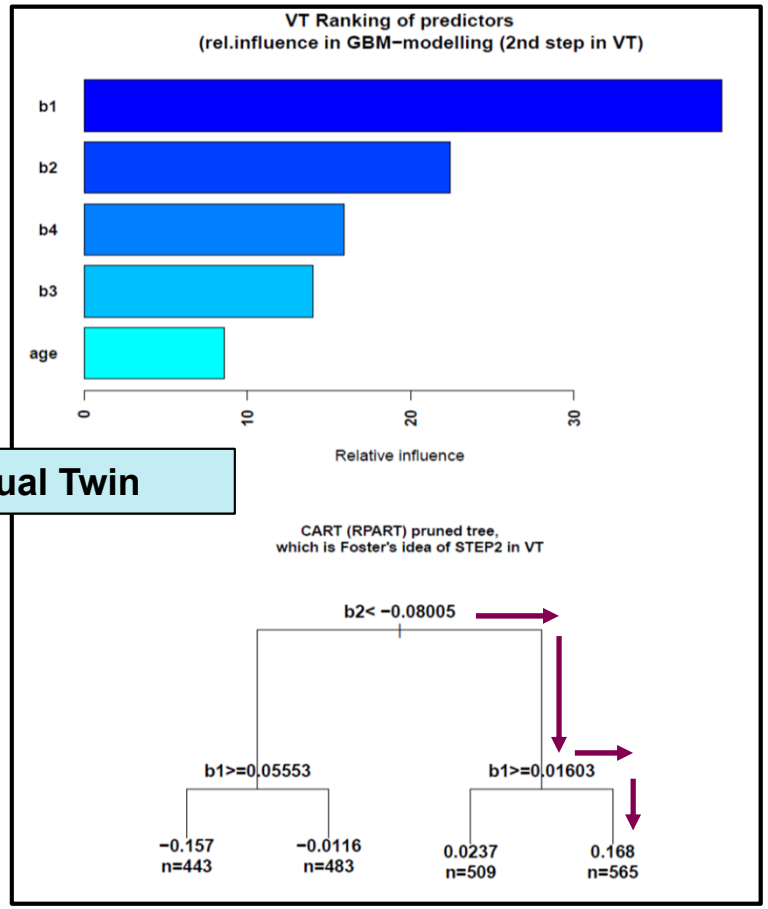
Take-Away Message: strongly prognostic biomarker AGE dominatres the model fit.

GLM-tree.

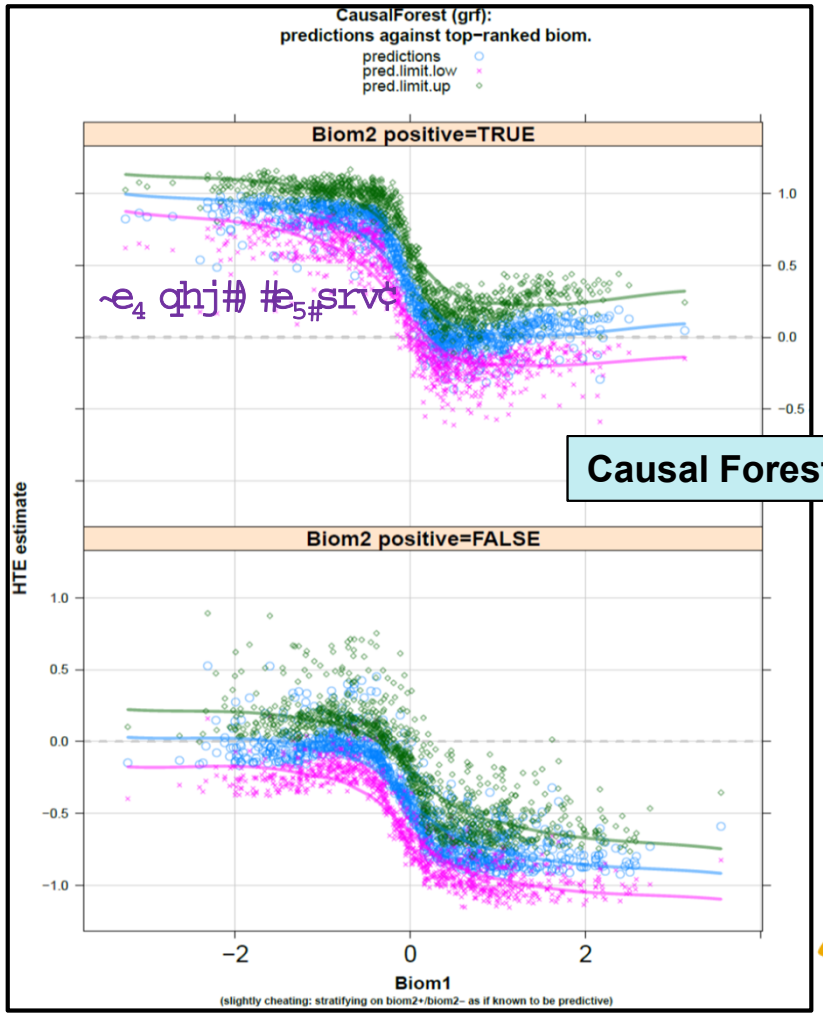
’Pre-adjusting’ for AGE, b3, b4 in the base-GLM helps …

GLM-tree (2).

Results on Toy Data [3]



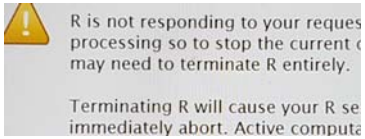
Virtual Twin



Causal Forest.

Some aspects to consider when comparing methods …

- For which **types of endpoints** can the method be applied? (continuous, binary, counts, censured).
- Does the method **suggest subgroups**, or does it merely seek interactions?
- **Stringency/Complexity control?** (E.g., re. subgroup definition).
- Not to underestimate: coding difficulties (wildly different output objects from R packages – and some methods doesn't have R code ..)



Method A could outperform method B in some simulation setting?
But *A only applicable if Y continuous,* whereas B can handle all endpoints?

Some literature comparisons:

In [5b] the following methods were compared: IT, SIDES, MOB, STIMA, L2-SVM (FindIt).

In [19b] simulation comparisons were conducted regarding BATTING, AIM, AIM-r, PRIM and SIDES.

In [10b] many simulations are presented, but mostly regarding certain performance measures for QUINT (e.g., type-1-errors, recovery probabilities, tree complexities, split points, etc) under certain assumptions. They compared STIMA, IT and QUINT on a cancer trial.

In [35b] eight simulation models, comparing ability to top-rank predictive biomarkers in the presence of prognostic; INFO+, VT, SIDES.

In [26b] the method L2-SVM (FindIt) is compared against several other regression-based methods (GLMNET, MOB, BART, Boosting, Bayesian GLM, Conditional Inference Trees).

in [28b] simulation comparisons were presented for GUIDE, VT, QUINT, SIDES, MOB, and IT in terms of various metrics (such as selection bias) and estimation accuracy.

Based on the evidence presented here, it is obvious that every method exhibits strengths and weaknesses. […] ***further research is needed*** *to better understand the performance characteristics of these […]"* [1b]

An aspect that might need further some research*?

Most papers deal with the problem schematically described as

$y \sim \mathbf{x}, \mathbf{trt}$ and aiming to model $\delta(\mathbf{x}) = E(Y_1 - Y_0 | \mathbf{x})$

But not so much on <u>known baseline covariate to adjust for</u>? (y, or else).

$y \sim y_b, \mathbf{x}, \mathbf{trt}$

(E.g., Y=**hbA1C**, y_b=base.**hbA1C**, or Y=no.events, y_b=prev.no.events, or y_b= AgeGr).

- IGNORE Y_b and run the black-box on y~**x?**
- Let Y_b be another SPLITTING CANDIDATE i.e. as another **x**

} Problems! Power drop, prognostic domination

(* We stress that simulation comparisons are generally highly demanding in the practice, and it is very understandable that there still are some scope for further investigations; work in progress).

Connection: prognostic / pre-adjustment of covariate

Assume methods A and B are compared (via simulations):

Then you can have

- **A > B** on setups (y, **x, trt)** – (no baseline y)
- **B > A** on setups (y, y_b, **x, trt)** – (with baseline y present)

  – Because **B** can adjust for it statistically (e.g., SIDES, GLMtree, GUIDE)
  – Whereas **A** can't – must attempt to split on it (e.g., VT, QUINT)

Conclusions:

HTE/ITE: rapidly developing, complex area. A bit different from standard Machine Learning.

Many recent fine methodological contributions.

Not straightforward to assess performance characteristics.

Still room for further research. (E.g., impact of adjust. for baseline-cov).

[ZRUN#LQ#SURJUHVV]

If interested, check out: **Biopharmnet Subgroups repository** [2b]

References. (a=Pre-specified Subgroups, b=Data-Driven) – [1]

[1a] **Alosh, M.** *Statistical Considerations on Subgroup Analysis: Interpretation of clinical trial findings and study design for targeted subgroup.* Conference paper, FDA/DIA Statistics Forum, At North Bethesda, Maryland, US, April 2014**.**

[2a] **Altman, D., Royston P**. *The cost of dichotomising continuous variables***.** BMJ, 332(7549), 1080. 2006 May 6.

[3a] **Bornkamp, B., Ohlssen D., Magnusson B., Schmidli, H**.; Model Averaging for Treatment effect estimation in subgroups. Phar. Statistics, 2017, vol 16.

[4a] **Brookes ST., Whitely E., Egger M., Smith GD., Mulheran PA., Peters TJ.;** *Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test.* J Clin Epidemiol. 2004, Vol 57(3).

[5a] **Byar P. D.** Assessing apparent treatment-covariate interactions in randomized clinical trials. Statistics In Medicine 1985 (Vol 4).

[6a] **Cui L, Hung JHM, Wang SJ, Tsong Y**. *Issues related to subgroup analysis in clinical trials*. *Journal of Biopharmaceutical Statistics* 2002, Vol 12.

[7a] **Dane, A., Spencer, A., Rosenkranz, G., Lipkovich, I., Parke, Tom**. *Subgroup Analysis and Interpretation for Phase 3 Confirmatory T7ials: White paper of the EFSPI/PSI Working Group on Subgroup Analysis.* [accepted 2018], Pharmaceutical Statistics.

[8a] **Dmitrienko, A., Millen, B., Lipkovich, I.**; *Multiplicity considerations in Subgroup A5alysis*. Stat. In Medicine, 2017.

[9a] **EMA** *Guideline on the investigation of subgroups in confirmatory clinical trials* http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf).

[10a] **FDA and Johns Hopkins University Center of Excellence in Regulatory Science and Innovation (JHU-CERSI).** *Assessing and Communicating Heterogeneity of Treatment Effects (HTE) for Patient Subpopulations: Challenges and Opportunities.* Public Symposium November 28, 2018.

[11a] **Foster J., Nan B., Shen L., Kaciroti N., Taylor J.**; *Permutation Testing for Treatment-Covariate Interactions and Subgroup Identification*. Stat. Biosci. 2016 Jun;8(1).

[12a] **ICH Efficacy Guidelines** http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html#16

[13a] **Marius T., Borncamp, B.**; *Comparing Approaches to Treatment Effects Estimation for Subgroups In Clinical Trials*. Stat. in Bioph. Research, 9:2, 2016.

[14a] **Mayer C, Lipkovich I and Dmitrienko A**. *Survey results on industry practices and challenges in subgroup analysis in clinical trials*. Stat Biopharm Res 2015; 7.

[15a] **Paget M., Chuang-Stein C., Fletcher C., Reid C.;** *Subgroup analyses of clinical effectiveness to support health technology assessments.* Pharm. Stat., Vol 10 (6) 2011.

[16a] **Pocock, S. J., Assmann, S. E., Enos, L. E. and Kasten, L. E.** *Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems*. Statistics in Medicine 21, 2917–2930, 2002.

[17a] **Rosenkranz G.;** *Exploratory Subgroup Analyses in Clinical Trials.* ISCB Pre-Conference Training Course, Birmingham, August 21, 2016

[18a] **Rosenkranz G.;** *Bootstrap corrections of treatment effect estimates following selection.* Computational Statistics and Data Analysis, Vol 69, 2014.

[19a] **Russek-Cohen., E**.; *Comments from the FDA working group on SUBGROUPS.* http://www.ema.europa.eu/docs/en_GB/document_library/Presentation/2015/03/WC500183614.pdf. 2015.

[20a] **Svensson, D**. *Consistency of treatment effect across pre-specified subgroups – should we (and, if so, how?) adjust for biases?* PSI Conference, Amsterdam, June 2018.

References. (a=Pre-specified Subgroups, b=Data-Driven) – [2]

[1b] **Alemayehu D., Chen Y., Markatou M.** *A Comparative Study of Subgroup Identification Methods for Differential Treatment Effect: Performance metrics and recommendations*. Statistical Methods In Medical Research, june 2017.
[2b] **Biopharmnet Subgroups repository**: Biopharmaceutical Network: Collaboration, Knowledge Sharing, Statistical Software. http://biopharmnet.com/subgroup-analysis/
[3b] **Breiman, L.;Friedman, JH.;Olshen, RA.; Stone, CJ**. *Classification and Regression Trees.* New York: Chapman and Hall; 1984.
[4b] **Breiman L**. *Random Forests*. Machine Learning. 2001;45:5–32.
[5b] **Chen, S., Cai, L., Yu, M**. *A general statistical framework for subgroup identification and comparative treatment scoring*. 2017. Biometrics, doi:10.1111/biom.12676
[6b] **Chipman HA, George EI, McCulloch RE** *BART: Bayesian Additive Regressive Trees.* (2010).*The Annals of Applied Statistics*, **4**(1), 266–298. doi:10.1214/09-aoas285.
[7b] **Dusseldorp E., Conversano C., Van Os B.** *Combining an Additive and Tree-Based Regression Model Simultaneously: STIMA.* Journal of Computational and Graphical Statistics, Vol 19, 2010.
[8b] **Dusseldorp E, Meulman, J**. *The Regression Trunk Approach to Discover Treatment Covariate Interactions*. Phsykometrika 69, Sept 2004.
[9b] **Dusseldorp E., Doove L., Mechelen I.**; *Quint: An R package for the identification of subgroups of clients who differ in which treatment alternative is best for them*. Behav Res, 2016, 48:660-663.
[10b] **Dusseldorp E., Mechelen I.** *Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions*. *Statistics in Medicine, 2014, 33(2)*
[11b] **Foster J., Taylor J., Ruberg S**.; *Subgroup identification from randomized clinical trial data*. Stat. In Medicine, 2011.
[12b] **Foster J., Nan B., Shen L., Kaciroti N., Taylor J.**; *Permutation Testing for Treatment-Covariate Interactions and Subgroup Identification*. Stat. Biosci. 2016 Jun;8(1):
[13b] **Gottlow, M.; Svensson D.**; Application of analyses to identify and characterize predictive properties of biomarkers and support the definition of a biomarker positive population, PSI Poster, London 2017.
[14b] **Hastie T., Qian J.** *Glmnet Vignette.* CRAN, 2016.
[15b] **Hastie, T; Tibshirani, R; Friedman, J**. *The Elements of Statistical Learning; Data Mining, Inference and Prediction.* Springer series in statistics, 2001.
[16b] **Hauck, W.W., Anderson, S., Marcus SM**. *Should we adjust for covariates in nonlinear regression analyses of randomized trials?* Control Clin. Trials 1998, 19(3), 249-56.
[17b] **Hothorn T, Zeileis**. A (2015). partykit*: A Toolkit for Recursive Partytioning*. R package version 1.0-3, URL http://CRAN.R-project.org/package=partykit.
[18b] **Huling, J, Menggang, Y**. *Subgroup Identification Using the Personalized Package.* Webarchive (link) submitted, 2018 Sept.
[19b] **Huang X., Sun Y., Trow P., Chatterjee S., Chakravartty A., Tian L., Devanarayan V.** *Patient Subgroup Identification for Clinical Drug Development*. Stat. In Medicine January 2017.
[20b] **Kasey Jones K**. mobForest R CRAN package. https://cran.r-project.org/web/packages/mobForest/index.html (2018).
[21b] **Kapelner A.**, **Bleich J**. *bartMachine: Machine Learning with Bayesian Additive Regression Trees*. CRAN.
[22b] **Imai L., Ratkovic, M.** *Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation.* The Annals of Applied Statistics, 2013, vol 7.
[23b] **Liaw A.**: RandomForest R CRAN package. https://cran.r-project.org/web/packages/randomForest/index.html (2018).

References. (a=Pre-specified Subgroups, b=Data-Driven) – [3]

[24b] **Lipkovich I., Dmitrienko A., Denne J., Enas G.**; *Subgroup Identification based on Differential Effect Search (SIDES) – a recursive partitioning method for establishing response to treatment in patient subpopulations*. Stat. In. Medicine, 2011.
[25b] **Lipkovich I., Dmitrienko A, Patra K., Ratitch B., Pulkstenis E**; *Subgroup Identification in Clinical Trials by Stochastic SIDEScreen Methods*. Statistics In Biopharmaceutical Research, 2017, Vol 9, (4) 368-378.
[26b] **Lipkovich I., Dmitrienko A., D'Agostino Sr R.**; *Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials*, Stat. In Medicine. July 2016.
[27b] **Loh W.** *GUIDE Approach to Subgroup Identification and Analysis for Precision Medicine*. Presentation 30 October 2017.
[28b] **Loh W., He X., Man M.** *A Regression Tree Approach to Identifying Subgroups with Differential Treatment Effects*. Statistics in Medicine, January 2014.
[29b] **Loh W., Man M., Wang S.** *Subgroups from Regression Trees with adjustment for Prognostic effects and post-selection inference*. Statistics in Medicine, 2018.
[30b] **Natekin N., Knoll A.**; *Gradient boosting machines, a tutorial*. Front Neurorobot. 2013; 7: 21.
[31b] **Riviere, M-K**. R CRAN Sides package. https://cran.r-project.org/web/packages/SIDES/index.html (2017).
[32b] **Ridgeway G., Southworth H**: Github Gradient Boosting R package https://github.com/harrysouthworth/gbm (2015).
[33b] **Ridgeway G**., CRAN Gradient Boosting R package https://cran.r-project.org/web/packages/gbm/index.html (2015).
[34b] **Ridgon J., Baiocchi M., Basu S**.: Preventing False Discovery of Heterogeous Treatment Effect subgroups in Randomized Trials. (2018) 19:382. https://doi.org/10.1186/s13063-018-2774-5
[35b] **Sechidis K., Papangelou A., Metcalfe P., Svensson D., Weatherall J., Brown G**.; *Distinguishing prognostic and predictive biomarkers: An information theoretic approach*. Bioinformatics (accepted May 2018).
[36b] **Svensson D., Gottlow M.** *To what extent do biomarker subgroup detection algorithms top-rank prognostic variables instead of predictive ones? A simulation comparison of SIDES and Virtual Twin.* PSI Conference, Amsterdam, June 2018.
[37b] **Su X., Peña A., Liu L., Levine R**. *Random forests of interaction trees for estimating individualized treatment effects in randomized trials*. Stat. in Med. Dec 2017.
[38b] **Su X., Tsai C., Wang H.** *Subgroup analysis via recursive partitioning*. J Mach Learn Res 2009; 10: 141–158.
[39b] **Su X., Zhou T., Yan X., Fan J., Yang S.** *Interaction Trees with Censured Survival Data*. The International Journal of Biostatistics, 2008; vol 4, issue 1.
[40b] **Vieille F., Foster J.**: CRAN Virtual Twins R package https://github.com/prise6/aVirtualTwins (2018)
[41b] **Vieille F., Foster J.**: Vignette aVirtualTwin on CRAN: https://cran.r-project.org/web/packages/aVirtualTwins/aVirtualTwins.pdf (2018).
[42b] **Wager S., Athey S.:** Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of Am. Stat Association, (2018), 113:523-
[43b] **Zeileis A., Hornik K.** Generalized M-Fluctuation Tests for Parameter Instability. Preprint of article published in Statistica Neerlandica, 61(4), 488-508, Blackwell Publishing, doi:10.1111/j.1467-9574.2007.00371.x.
[44b] **Zeileis A, Hothorn T.** Parties, Models, Mobsters: A New Implementation of Model-Based Recursive Partitioning in R, CRAN Vignette.

BACKUP SLIDES

Quote

- “*Another shortcoming of these methods is that the accompanying software lacks instructions about how to use it to identify treatment-subgroup interactions. Some of these methods merely provide software code without a manual […] and some of them provide only general instructions that are not adapted to treatment-subgroup interactions (e.g., STIMA). As a solution, recently a new tree-based method […]*” [9b]

Prognostic or predictive?

Modern Approaches are design to focus in on the ***predictive*** biomarkers and to avoid 'being tricked' by the **prognostic** variables.

Great improvement over classical ML. E.g., Virtual Twin scheme:

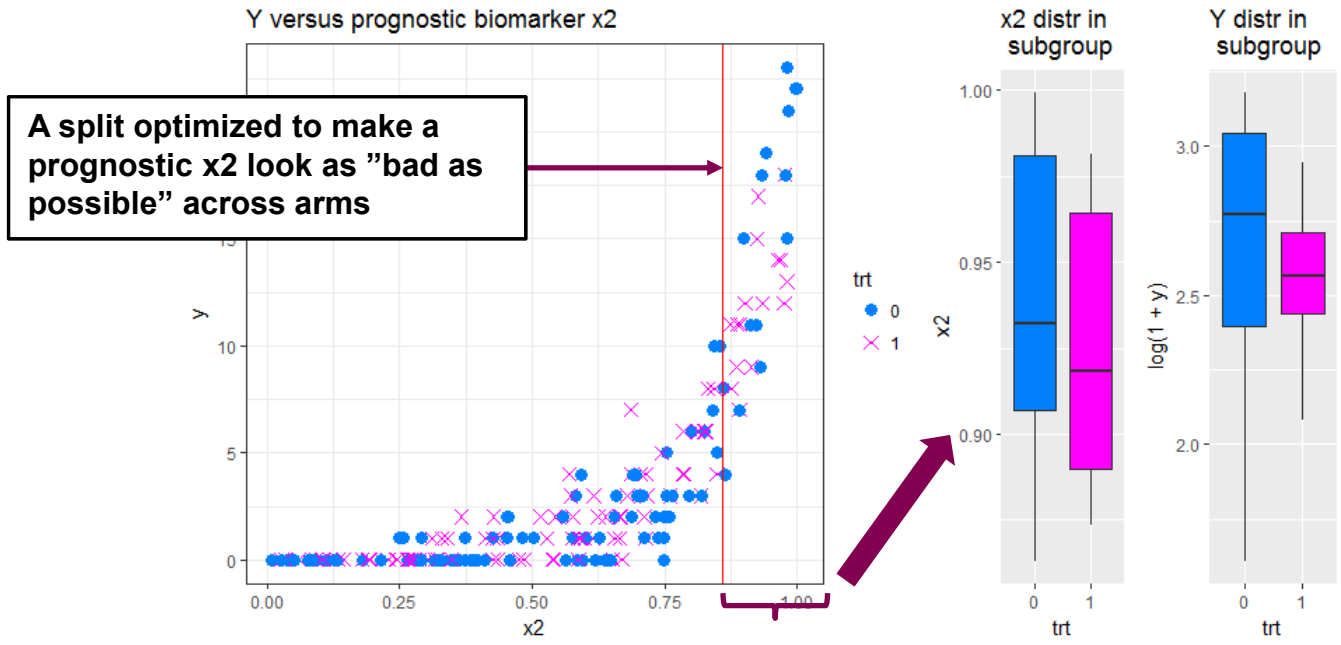
- Fits predictive models to active arm & control arm,
- Then, predict each patient **factually** & **counterfactually**
- Hence, VT 'knows' patient's responses to both treatments $Y_{(1)}$ & $Y_{(0)}$
- Then explores driving biomarkers behind differences $z=Y_{(1)} - Y_{(0)}$

(*Just consider the RF vs VirtualTwin in the previous toy example section* !).

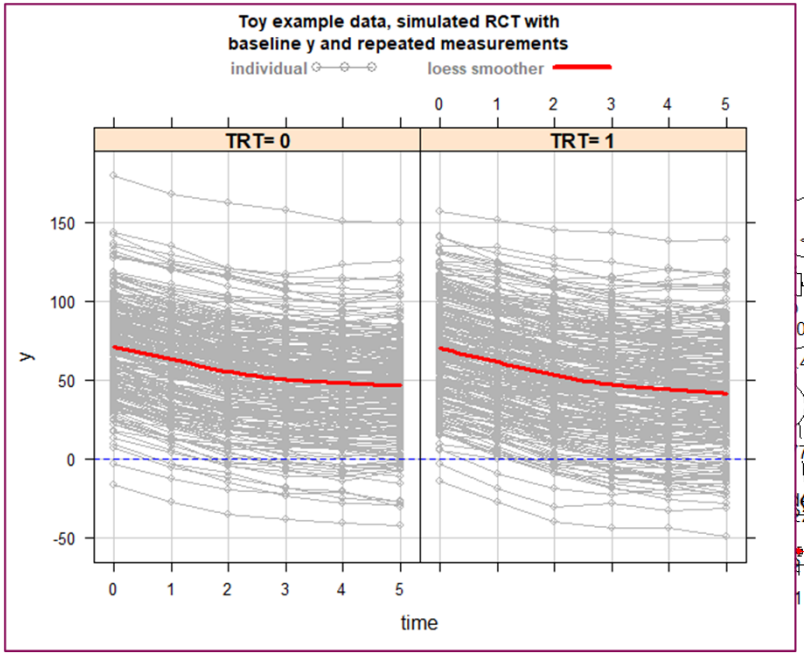
However, still some **tendency** to top-rank prognostic. (e.g., [35b], [36b], [37b]).
(probably largely unknown to what extent, across all methods).

Why could <u>prognostic</u> biomarkers appear predictive?

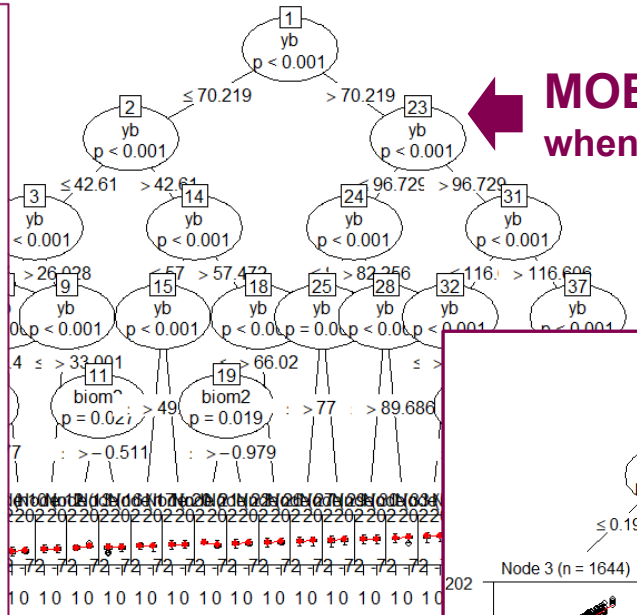
Small-n, and "*more can happen with a prognostic*":
RCT full population: bsl balance.
Subgroup node: bsl can be imbalanced:



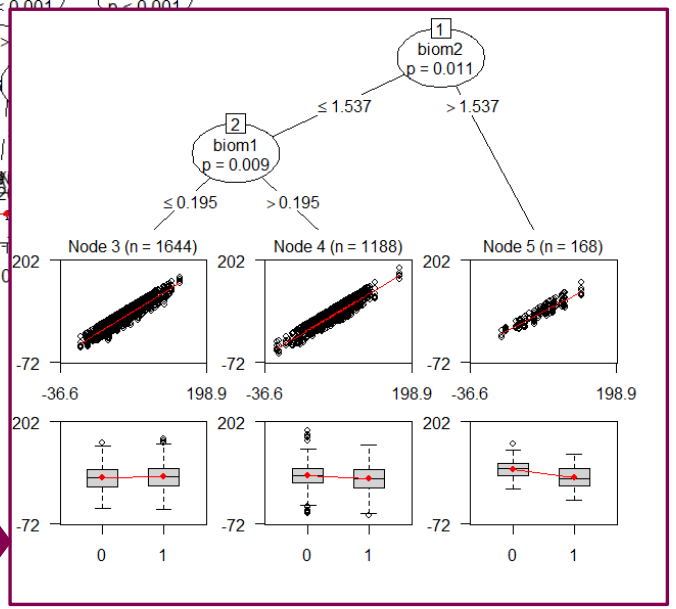
Toy example with y and y_b (and true subgroup b1>0.25)



RCT n=500/arm, ”A1c” style,
Dection run on bsl+week5



MOB (GLMtree)
when yb splitting candidate



MOB (GLMtree)
adjustment for yb

VT: a crash course: (blue=highlights)

- Let $f(\mathbf{x}, t) = E[Y|\mathbf{X}, T = t]$ denote the expected response of a patient, as a function of biomarkers $\mathbf{x}$ and treatment assignment t.

  – STEP 1: Fit a predictive model, and estimate **each patient's expected response** $\hat{f}(\mathbf{x}, 1)$ and $\hat{f}(\mathbf{x}, 0)$ to active (1) & control (0).
  – One is counterfactual!
  – Subgroup? Patients with large **differences** $z_j = \hat{f}(\mathbf{x}_j, 1) - \hat{f}(\mathbf{x}_j, 0)$

  – STEP 2: Explore if **differences can be predicted by biomarkers**? I.e., fit new model (tree). [4] used a CV-pruned CART tree to get to subgroup. (Subgroup = Union{high-response leaves}).

Foster et. al [4]: Random Forest =STEP 1.
(But single RF model or 2 RF models? Both suggested).

SIDES: a crash course: (blue=highlights)

- **SIDES** [10] also recursive partitioning, but based on a suitable GLM model. More complexity control, e.g., max-depth *L* of subgroup (e.g., *L*=2).

  – **All possible biomarker splits** c are considered, for all biomarkers (Low, High): $L_i(c) = \{X_i \leq c\}$ and $H_i(c) = \{X_i > c\}$

  – Splitting criterion $\mathtt{D}(c)$ used to assess each such candidates.
  – **Optimal cut** exist: $c_i^* = \mathtt{argmin}_{c \in \mathcal{C}} \mathtt{D}(X_i, c)$

  – **One biomarker & split must win; then recursive repeat**.
  – (Some other parameters; e.g., keep M most promising splits in each step).

  – Stop search if too small size (or not good enough effect).
  – Final subgroup: only split once per biomarker.

From []: Y binary. Biom1=prognostic, Biom2=predictive.

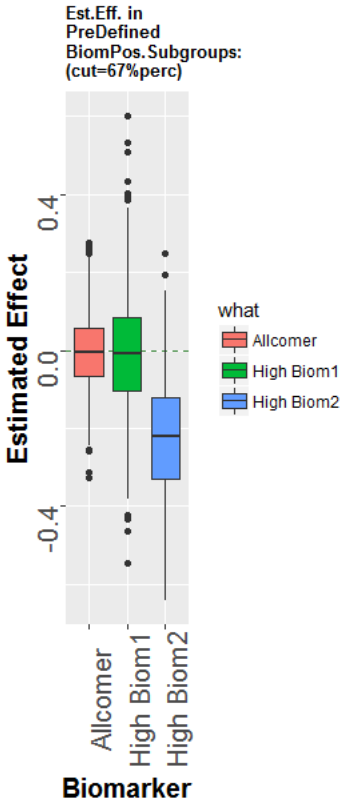
- Mixture case:
  real subgroups exist
  (based on biom2),
  biom1 is prognostic.

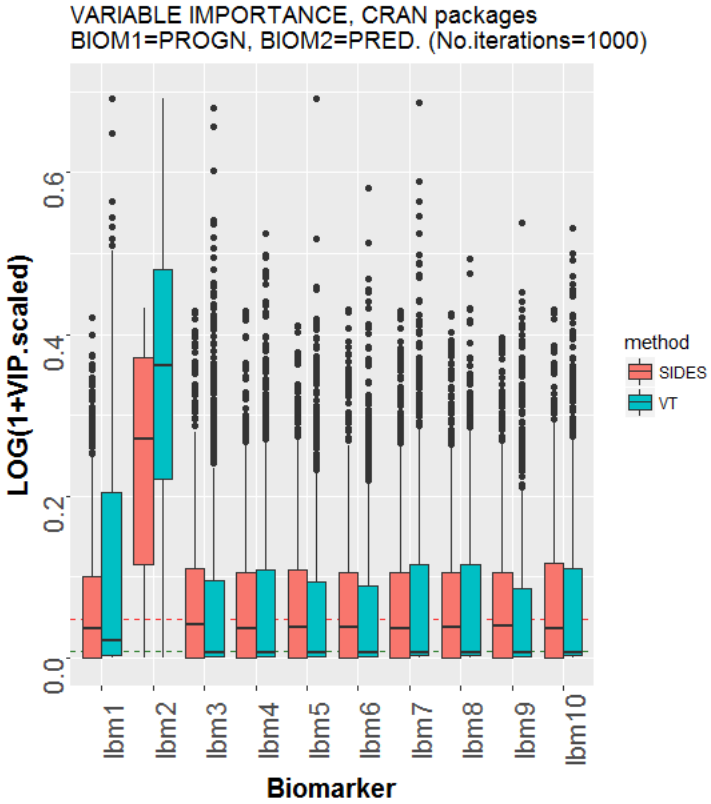
	biom	prob.split.SIDES	prob.split.VT
1	lbm1	6.60	8.70
3	lbm2	49.10	59.00
4	lbm3	6.90	3.60
5	lbm4	5.70	4.30
6	lbm5	6.10	4.80
7	lbm6	5.10	4.40
8	lbm7	5.40	4.70
9	lbm8	4.90	3.30
10	lbm9	4.50	3.20
2	lbm10	5.70	4.00

ANTINULL2: Probability (percentage) 1st split. (1000 iterations)

	corr.SIDES	corr.VT
1	0.645	0.768

ANTINULL2: Correlation(1stSplit, TopRanked) (1000 iterations)





From []. Biom1=prognostic, Biom2=predictive.

Additional aspect:

VT based on one and two GBM models.

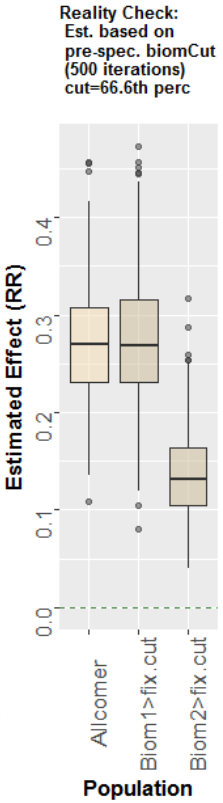
Tracking VIP and loss-of-effect.

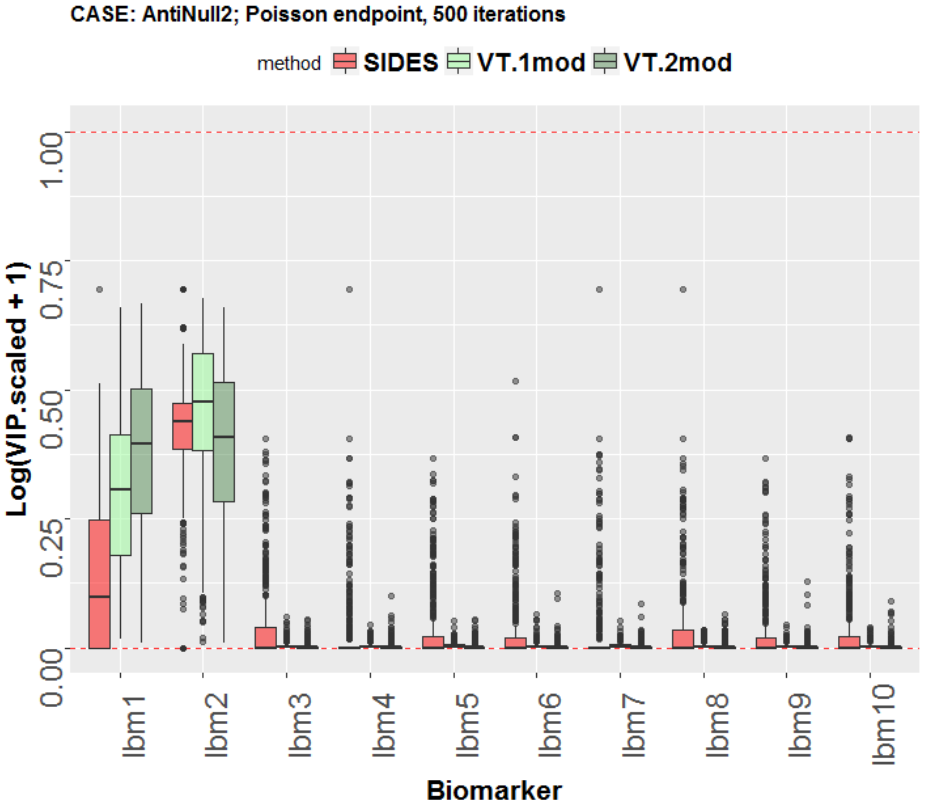
	corr.VT1	corr.VT2
1	0.758	0.708

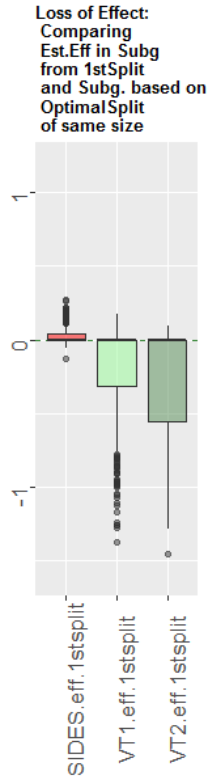
VT Corr(1stSplit, TopRank) (500 iterations)

	biom	sides.prob	vt1.prob	vt2.prob
1	lbm1	0.052	0.286	0.444
2	lbm2	0.716	0.714	0.556
3	other	0.232	0.000	0.000

Table 3: Prob.1st split (500 iterations)







Confidentiality Notice