Global Drug Development
Analytics

A systematic framework to compare methods for subgroup identification in realistic scenarios

Björn Bornkamp
PSI Subgroup SIG Webinar
November 17, 2021

NOVARTIS | Reimagining Medicine

Contributors

- Sophie Sun
- Kostas Sechidis
- Yao Chen
- Jiarui Lu
- Chong Ma
- Ardalan Mirshani
- David Ohlssen
- Marc Vandemeulebroecke

Personalized medicine on the rise

- **39%** of FDA-approved therapies in **2020** are personalized medicines[1]
  - stable over past few years, up from <10% in 2010
- Most (all) biomarkers close to drug mechanism (highly biologically plausible)
- What about situation with lower biological understanding & more covariates/biomarkers?
  - Statistical learning approaches to rescue?
    - How well do these methods work?
    - Which methods work better?

1. Personalized Medicine Coalition (PMC). Personalized medicine at FDA the scope and significance of progress in 2020, https://personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/PM_at_FDA_The_Scope_Significance_of_Progress_in_2020.pdf
PMC categorizes personalized medicines as those therapeutic products for which the label includes reference to specific biological markers, often identified by diagnostic tools, that help guide decisions and/or procedures for their use in individual patients.

Overview

- Data generation
- Metrics
- Compared methods
- Results

Data generation

- Simulate data from (A – treatment, $\boldsymbol{X}$ – baseline covariates)

$$Y \sim N(f(\boldsymbol{X}, A), \sigma^2)$$

$$f(\boldsymbol{X}, A) = f_{prog}(\boldsymbol{X}) + A\left(\beta_0 + \beta_1 f_{pred}(\boldsymbol{X})\right)$$

- β_1 measures amount of treatment effect heterogeneity
  - $\beta_1 = 0$ all patients have the same treatment effect
- Open question: How to choose
  - Sample size, σ, overall treatment effect (determined by β_0, β_1)
  - Magnitude of prognostic effects $f_{prog}(\boldsymbol{X})$
  - Amount of treatment effect heterogeneity (determined β_1)
  - Distribution of $\boldsymbol{X}$
  - Functional form of $f_{pred}(\boldsymbol{X})$

Simulation scenarios

- Sample size, σ, overall treatment effect (determined by β_0, β_1)
  - Choose these parameters so that power of trial is 50%
  - Nuisance parameter when it comes to detection of differential treatment effects
- Magnitude of prognostic effects $f_{prog}(\boldsymbol{X})$
  - Use real trial data; develop model for control arm and determine R^2.
    Iterate size of prognostic effects such that the specified R^2 is obtained
  - Use two prognostic covariates (linear effects)
  - Use scenarios so that one prognostic covariate is predictive (the other not)

Simulation scenarios (cont)

- Amount of treatment effect heterogeneity (determined β_1)?
  - Choose different scenarios based on the underlying true simulation model
    - Calculate β_1 that can be detected with 80% at a one-sided type 1 error of 10%
    - Vary β_1 in a 2-fold range
- Distribution of $\boldsymbol{X}$?
  - Use synthpop R package (Nowok et al 2016) fitted to real trial covariate data to preserve correlation structure
  - Generate $\boldsymbol{X}$ using synthetic data
  - Here use 30 candidate covariates
    - 8 categorical (7 of them binary)
    - 22 continuous (standardized to [0,1])

Scenarios investigated $f_{pred}(X)$

Scenario	$f_{pred}(X)$
1d step	$\Phi(20(X_{11} - 0.5))$
1d linear	X_{14}
2d step AND	$(X_{14} > 0.25)\ AND\ (X_1 = 'N')$
2d step OR	$(X_{14} > 0.3)\ OR\ (X_4 = 'Y')$

Metrics
1) Ability to detect treatment effect heterogeneity

- What is the evidence for treatment effect heterogeneity?
  - Or: How likely is it to see the observed evidence for treatment effect heterogeneity in case there is no treatment effect heterogeneity?
  - Schandelmaier et al. (2019) review 150 publications on assessing subgroup findings
    - Top recommendation: Significant test for subgroup by treatment interaction
  - Natural extension to multiple covariates: Global interaction test (appropriately adjusts for multiplicity)
  - Global interaction test/joint likelihood ratio test in a regression model

2) Ability to identify covariates/biomarkers that modify the treatment effect

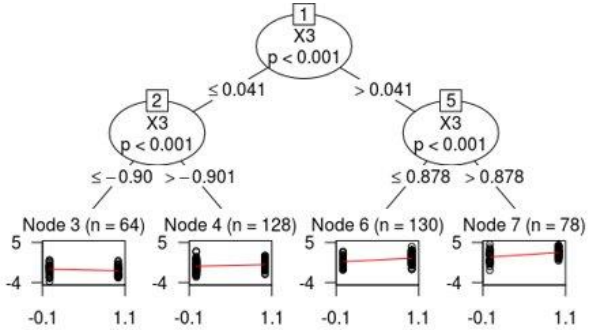
- Two sub-questions
  - In the situation of no treatment effect heterogeneity, do methods select specific variable types (variable selection bias)
  - In the situation of treatment effect heterogeneity: What is the probability that the top identified variable is actually predictive?

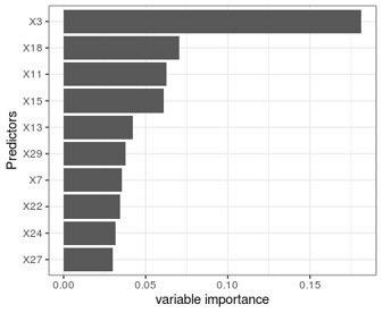
3) Ability to identify patients with increased treatment effect & provide a reliable treatment effect estimate

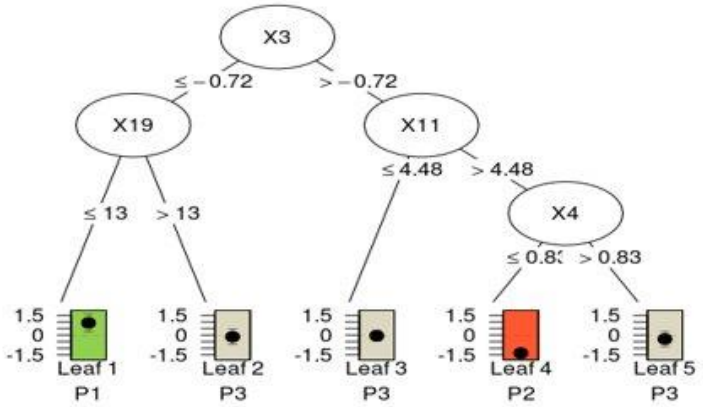
- Difficult to find a metric for subgroup detection
  - Trade-off between size of subgroup and treatment effect within subgroup
  - Solution of trade-off often context-specific
- Use predicted individual treatment difference declare patients with top 25% (50%) predicted treatment effect as subgroup
  - Metric 1: Assess true treatment effect in this subgroup (should be as large as possible)
  - Metric 2: Estimated treatment effect in subgroup (as returned by method) versus true treatment effect in the subgroup

Methodologies

- Tree-based methods
  - Model-based partitioning (MOB)
  - GUIDE
- Forest based methods (tree ensembles)
  - Causal forest
  - MOB forest
- Shrinkage-based regression methods
  - LASSO (separately fitted by treatment arm)
- Standard methods
  - Univariate analysis
  - Multivariate regression
- *Not all methods applied to all metrics*

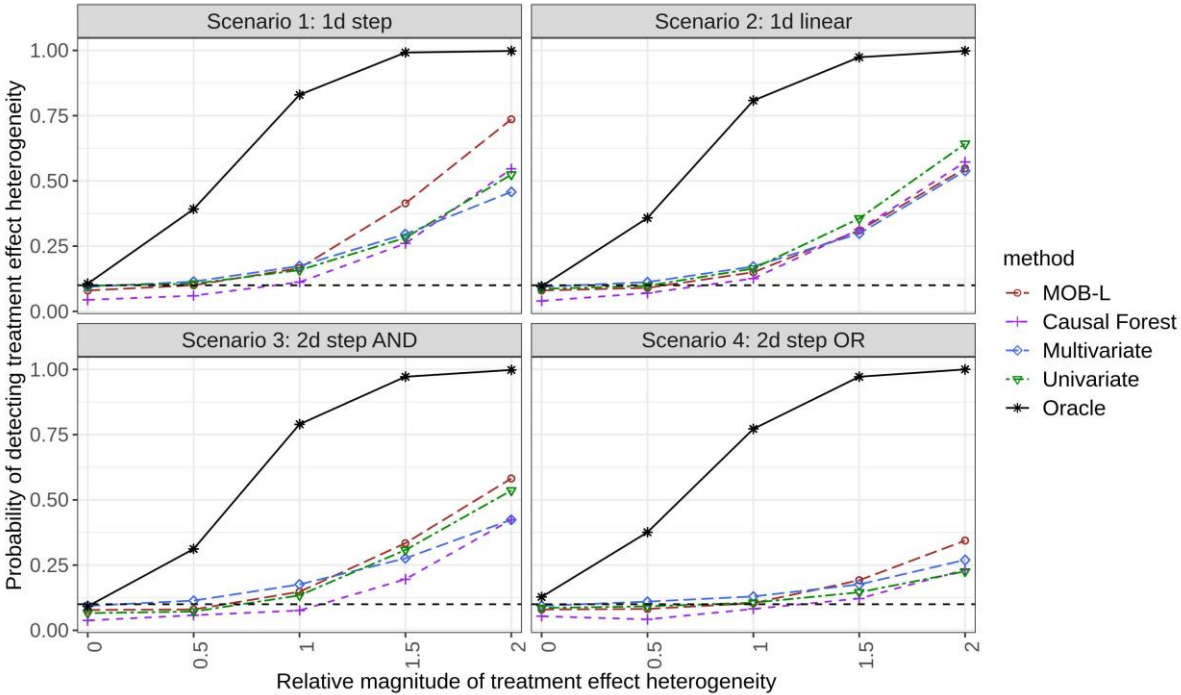






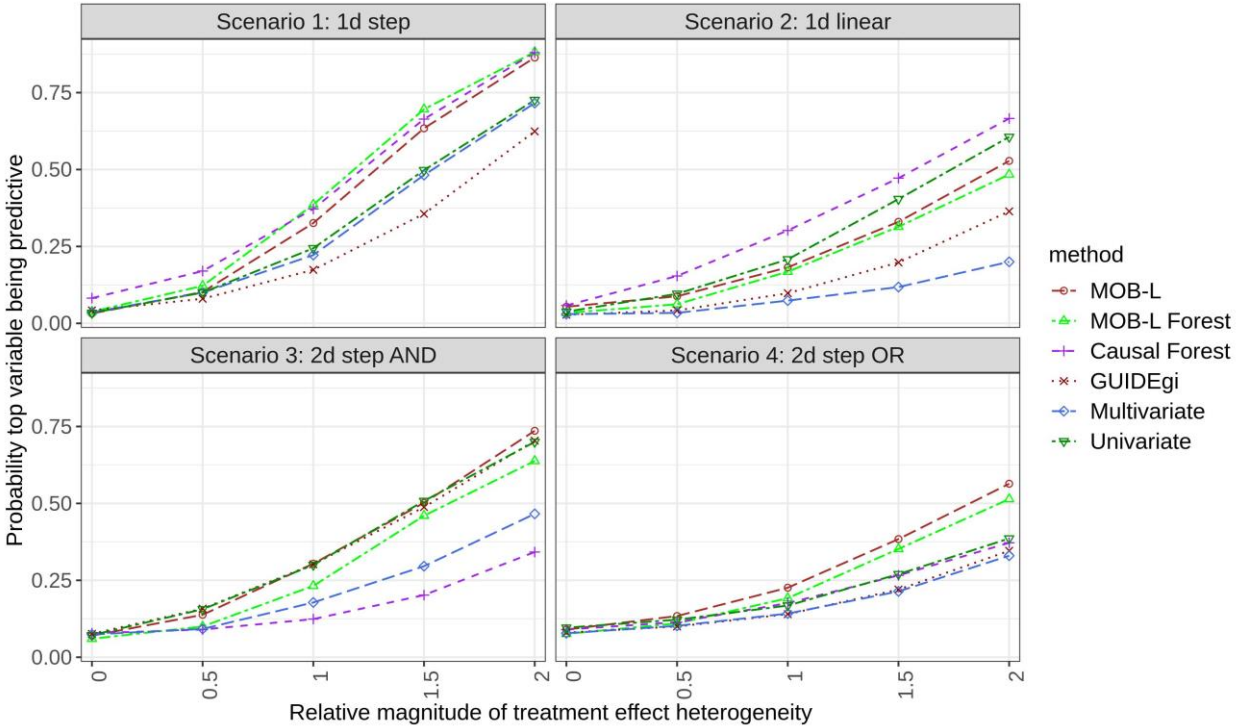
NOVARTIS | Reimagining Medicine

Probability to detect heterogeneity



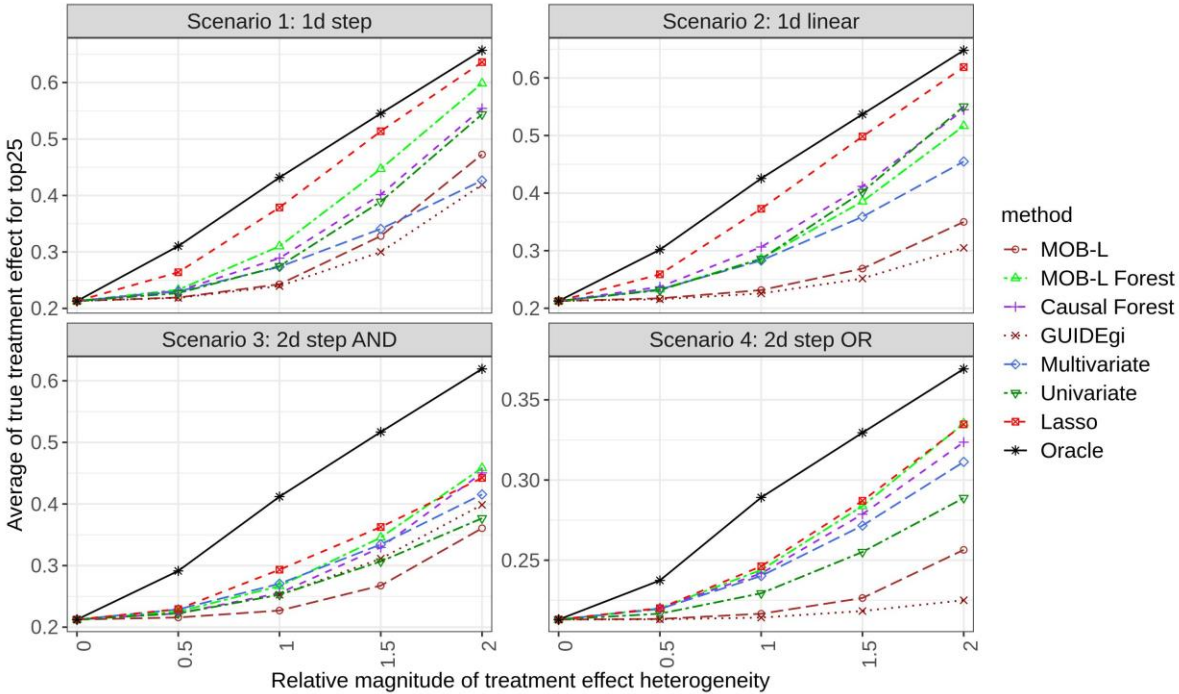
- All methods control false positive rate; not overly conservative despite use of Bonferroni within MOB-L or Univariate
- Even for considerable treatment effect heterogeneity power small correct positive rate → Inclusion of 30 covariates creates natural „false“ signals...
- MOB-L most consistent overall performance

Probability that top variable is predictive



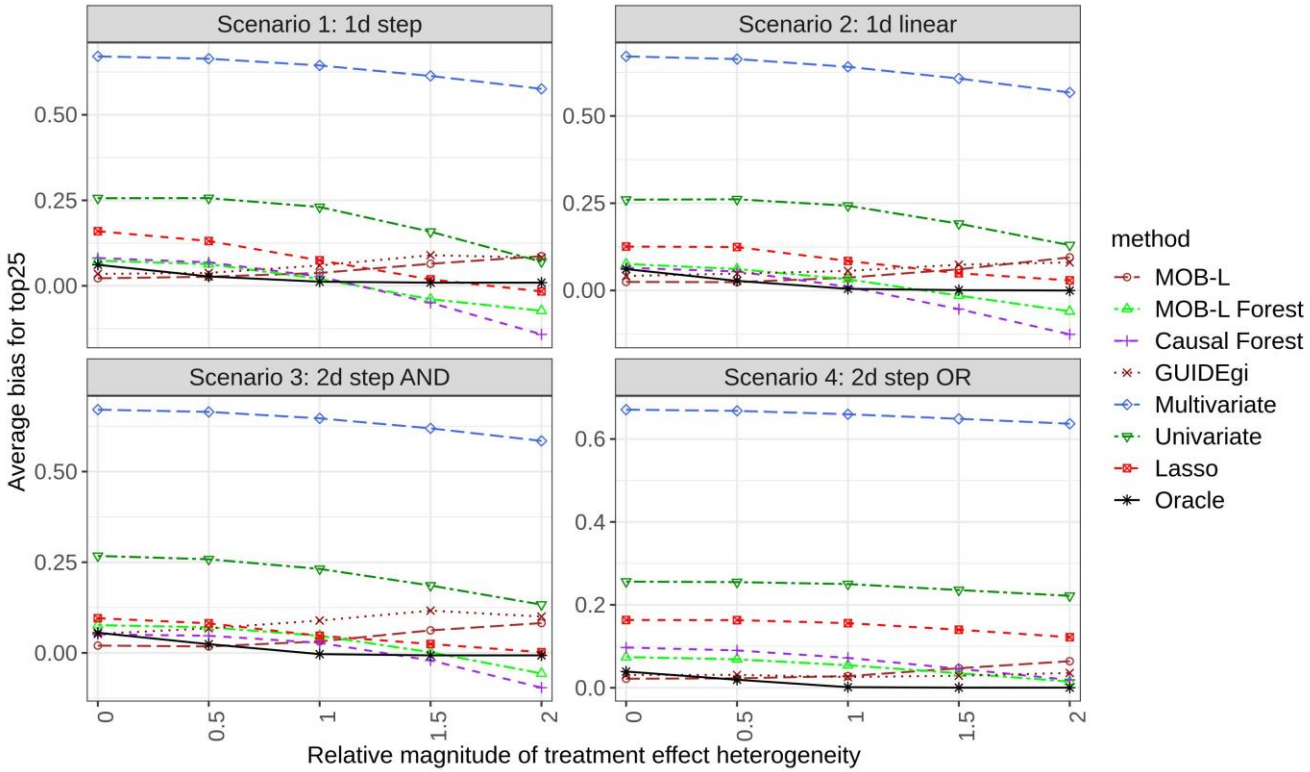
- Probability > 50% only for strong heterogeneity
- MOB-L and MOB-L Forest most consistent overall performance

True treatment effect for 25% of patients with largest *predicted* treatment effect



- LASSO and MOB-L Forest overall best performance

Compare true treatment effect in top 25% patients against the model prediction



- Naive standard methods (multivariate and univariate regression)
  → strongly overestimate treatment effect in subgroup; Need for adjustment
- Causal forest, MOB-L Forest and LASSO with better performance

Insights

- Reliable signal detection is challenging (even for 30 variables/biomarkers)
  - *Data alone* often cannot provide definite evidence in most scenarios
    → Take external data into account for subgroup assessment (mechanistic plausibility, external replication for similar drug or same drug in different indication)
    → Recommendations in the EMA guideline on subgroup analyses
- Methodology comparison
  - No strong separation of most methods (depends on metric & scenario): MOB-L, GUIDE, LASSO and MOB-L Forest provide good results
  - Standard univariate and multivariate regression: Surprisingly good, but:
    Do not use unadjusted treatment effect estimates in subgroups based on univariate or multivariate regression models
- Plan to make simulation scenarios & data available as R package!

References

- Nowok B, Raab GM, Dibben C (2016). “synthpop: Bespoke Creation of Synthetic Data in R.” Journal of Statistical Software, 74(11), 1–26. doi: 10.18637/jss.v074.i11
- Schandelmaier, S., et al. (2019). A systematic survey identified 36 criteria for assessing effect modification claims in randomized trials or meta-analyses. *Journal of clinical epidemiology*, *113*, 159-167.

Thank you

NOVARTIS | Reimagining Medicine