Data mining methods for subgroup identification

Ilya Lipkovich and Alex Dmitrienko, Quintiles

TICTS, April 22, 2014

Outline

- Introduction
  - Principles and standards for “Subgroup Analysis” in clinical research
  - Data-driven approaches to subgroup identification
  - Taxonomy of data mining methods for subgroup identification
- SIDES approach
  - Basic SIDES algorithm and SIDES parameters
  - Clinical trial example
  - SIDEScreen, an extension of basic SIDES method
- Summary and conclusions
- References

Principles and Standards for Subgroup Analysis in Clinical Research

- Subgroup analyses are often (rightfully) viewed by regulatory agencies as a way for the trial sponsor to make unsubstantiated claims about efficacy in subpopulations by *data dredging*
- Many authors came up with various "checklists" of principles for Subgroup Analyses
  - NHS R&D HTA Programme (Brookes et al. 2001) provides a list of 25 recommendations
  - Rothwell (2005) proposed a guideline with 21 rules
  - Sun et al (2009) listed the existing 7 plus 4 additional criteria for assessing credibility of subgroup analysis
- General theme
  - Subgroups need to be pre-specified, biologically plausible, significance tests multiplicity adjusted, "*no testing in subgroup unless interaction test is significant*"; sometimes stating that "*no testing in subgroup if the overall effect is not significant*", …, and finally "*interpreted with caution*"

Data-driven versus Guidance –Driven Approach for Subgroup Identification

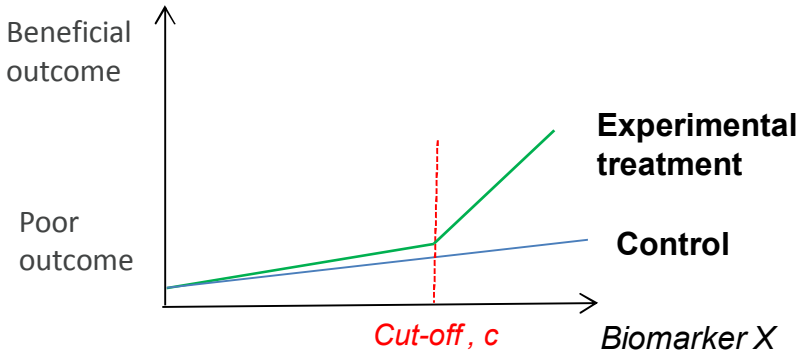
- Subgroup Identification is framed as a special case of Model Selection
- Pre-specified is the entire biomarker/subgroup selection strategy, not the specific subgroup(s) which need to be found!

Data Mining Methods for Subgroup Identification. Some Notation

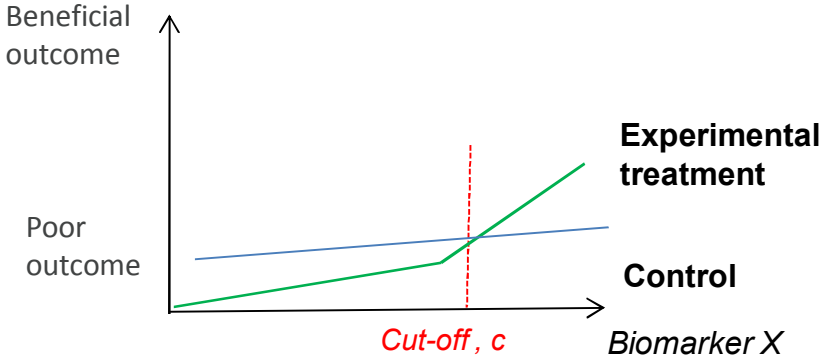
- Outcome (Y), treatment (T)
  - t_i=1 if subject i is assigned to "Drug" and t_i=0 if assigned to "Placebo")
- subject characteristics measured prior to treatment, $X_1,..,X_p$ combined in $\mathbf{x}=\{x_{i1},..,x_{ip}\}$
- Expected response for i-th subject: $f(\mathbf{x}_i,t_i)$
  - Expected individual outcomes if treated and untreated: $f(\mathbf{x}_i,1)$ and $f(\mathbf{x}_i,0)$
- Treatment contrast $z(\mathbf{x}_i) = g(f(\mathbf{x}_i,1), f(\mathbf{x}_i,0))$
  - $z(\mathbf{x}_i) = f(\mathbf{x}_i,1) - f(\mathbf{x}_i,0)$
- Subgroup effect (excess of the overall treatment effect)
  - $E(z(\mathbf{x}) \mid \mathbf{x} \in S)) - E(z(\mathbf{x}))$

Two Frameworks of Personalized Medicine

- Identifying the right patient for a given treatment
  - therapy provides minimal or no benefit in the overall population
  - find subpopulations that will have enhanced benefit from the treatment vs. control
- Identifying the right treatment for a patient
  - finding optimal treatment regime or policy for a given subpopulation.



The case of quantitative interaction



The case of qualitative interaction

Taxonomy of Data Mining Methods for Subgroup Identification

- Global outcome modeling
  - modeling underlying outcome function $f(\mathbf{x},t)$
- Global treatment effect modeling
  - modeling underlying treatment effect, $z(\mathbf{x})$
  - optimal treatment regime modeling: $\text{sign}(z(\mathbf{x}))$
- Local treatment effect modeling (subgroup search)
  - identifying subgroups $\{\mathbf{x} \in S\}$ with higher values of $z(\mathbf{x})$

Global outcome modeling $f(\mathbf{x},t)$

- Non-parametric "black box" model for f(**x**,t)
  - Virtual Twins method of Foster et al.(2011): estimating f(**x**,t) via Random Forest at the first stage, then using CART to predict estimated treatment contrast z(**x**)= f(**x**,1)-f(**x**,0) at the second stage.
- Parametric modeling with main effects and interactions (shrinking coefficients associated with T*X interaction terms)
  - Bayesian hierarchical models, e.g. Jones et al.(2011)
  - Frequentist analysis via penalized regression (LASSO, elastic net, etc.)
    - Need different penalty for main effects and T*X interactions, Imai & Ratcovic (2013)

Global treatment effect modeling, z(**x**)

- Advantage: obviates the need to model prognostic (main) effects and focuses on predictive (T- by-X interaction) effects
- Non-parametric modeling:
  - Interaction tree method of Su et al. (2009).
  - Adding new splitting criterion for detecting differential treatment effect in GUIDE recursive partitioning platform (Loh et al. 2014)
- Parametric approaches (penalized regression)
  - "Modified covariates" method of Tian et al.(2012).

Optimal treatment regime modeling, *sign*($z(\mathbf{x})$)

- Advantage: directly identifies optimal treatment for a given subject
- Examples:
  - Zhao et al. (2012) showed that minimizing weighted misclassification loss for predicting treatment labels (T=0/1) leads to OTR
    - Outcome based weighting: $W=Y/\pi$ for treated subjects and $W=Y/(1-\pi)$ for control subjects
    - π = Prob(T =1)=1/2 for RCT with1:1 randomization
    - Y is continuous outcome with larger values indicating beneficial outcome
  - Zhang et al. (2012) proposed a more general framework with doubly robust estimator of treatment contrast $z(\mathbf{x})$

Local treatment effect modeling

- Identifying subgroups {$\mathbf{x} \in S$} with higher values of $z(\mathbf{x})$
  - Advantage: obviates the need to estimate the response function over the entire covariate space and focuses on identifying specific regions with a large differential treatment effect.
  - Examples:
    - Bump hunting approach proposed by Kehl and Ulm (2006) (extending the PRIM methodology by Friedman and Fisher (1999).
    - SIDES method by Lipkovich et al. (2011).
    - Bayesian subgroup analysis via model averaging, Berger et al, (2014).

SIDES - **S**ubgroup **I**dentification based on **D**ifferential **E**ffect **S**earch

- All (p) candidate covariates are ordered from "best" to "worst" in terms of a treatment effect-based splitting criterion
- All candidate splits per covariate are evaluated and the optimal split into two child subgroups identified, resulting in one "**promising subgroup**" and one "**non-promising**"
- Retained are first k (e.g. k=2) covariates (rather than only the top one)

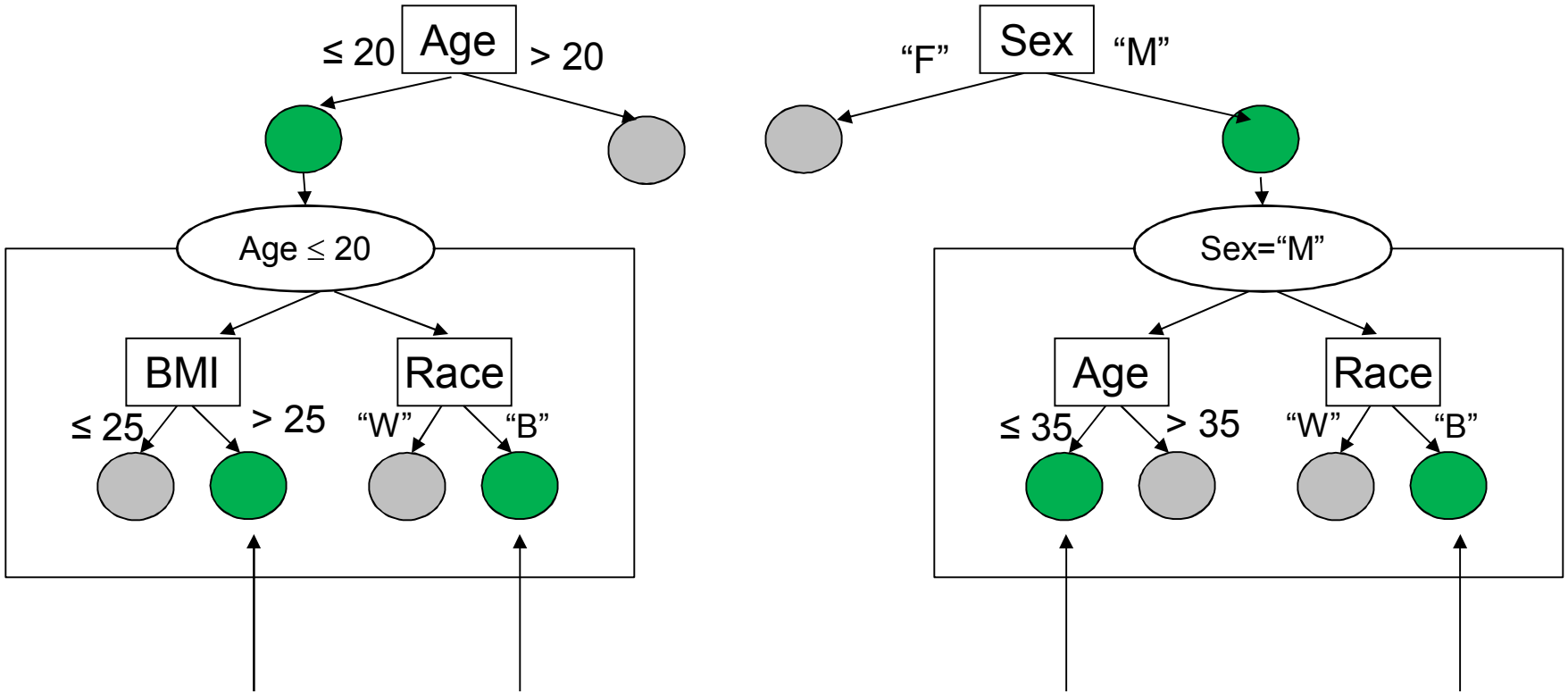
Z_1 and Z_2 are test statistics for testing
H_0: treatment effect=0 in child subgroups

Differential splitting criterion:

$$1 - \Phi\left(|Z_1 - Z_2| / \sqrt{2}\right)$$

Recursive Partitioning Algorithm

- Apply the same procedure recursively **up to L** times



SIDES Parameters

- Number of best candidate covariates pursued, "width" (e.g. k=3)
- Splitting criterion evaluated when forming a split (e.g. treatment by split interaction)
- The "depth", or number of levels in a subgroup (e.g. subgroup X1=0 & X2=0 has 2 levels), L=2
  - Note a subgroup of depth 3 corresponds to a 4-level treatment by covariate interaction!
- Complexity (tuning) parameter: relative improvement in treatment effect in child versus parent node, P(child)<γ P(parent) required to make a split. γ can be calibrated via cross-validation
- Type I error control
  - Construct reference distribution by permuting treatment labels and computing the proportion of null sets with smallest $Pval_{perm} \leq Pval_{obs}$
  - Multiplicity adjustment is not the end goal but is used in the process of subgroup selection in conjunction with tuning parameters to control the size of the search space and increase chances of subgroup replication in future trials

Running Example. Severe Sepsis Trial

- Testing for the difference in proportions in subjects who died within the first 28 days (assuming lower proportion is preferable)

Subgroup	Subgroup size	Test statistic (z)	1-tailed p value	treatment event rate	control event rate	treatment sample size	control sample size	effect size
All subjects (Analysis data)	470	-1.4001	0.9193	0.407	0.340	317	153	-0.138

Potential Covariates	Description
TIMFIRST	TIME FROM FIRST SEPSIS-ORGAN FAIL TO START DRUG
AGE	
BLLPLAT	BASELINE LOCAL PLATELETS
blSOFA	SUM OF BASELINE SOFA (CARDIOVASCULAR, HEMATOLOGY, HEPATICRENAL, RESPIRATION SCORES)
BLLCREAT	BASELINE CREATININE
ORGANNUM	NUMBER OF BASELINE ORGAN FAILURES
PRAPACHE	PRE-INFUSION APACHE-II SCORE
BLGCS	BASELINE GLASGOW COMA SCALE SCORE
BLIL6	BASELINE SERUM IL-6 CONCENTRATION
BLADL	BASELINE ACTIVITY OF DAILY LIVING SCORE

Applying SIDES to Data Example

- Apply SIDES with width (k=3), depth=up to 2 levels (L=2), and child-to-parent ratio (γ)=0.2
- Note the last column contains multiplicity-adjusted p-values

Subgroup	Subgroup size	Test statistic (z)	1-tailed p value	treatment event rate	control event rate	treatment sample size	control sample size	effect size	multiplicity -adjusted 1-sided p-value
AGE <=63.471 & PRAPACHE <=26	215	2.78	0.0028	0.157	0.320	140	75	0.40	0.30
PRAPACHE <=26	313	2.06	0.0198	0.236	0.347	212	101	0.25	0.55

Evaluating Variable Importance

- Apply SIDES with loose constrains to generate a large number of candidate subgroups
- Variable importance is computed for each biomarker as the average of its "contributions" over all subgroups where X was involved as the splitter
  - "X-Contributions"
    =–log(split-by-trt pval)
    =0, if X is not involved

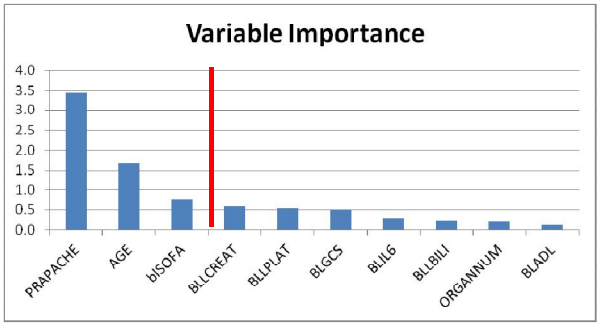
Subgroup	Subgroup size	Test statistic (z)	1-tailed p value (unadjusted)
PRAPACHE <=26 blSOFA>3 AGE <=73.138	239	3.2331	0.0006
AGE <=63.471 PRAPACHE <=26 blSOFA>3	195	3.1433	0.0008
AGE <=63.471 PRAPACHE <=26 BLIL6>92.8	160	3.1073	0.0009
PRAPACHE <=26 AGE <=73.138 BLIL6>60.3	212	3.0552	0.0011
BLLCREAT<=1.1 AGE <=59.871 blSOFA>3	80	3.0158	0.0013
AGE <=63.471 BLLCREAT<=3.8 PRAPACHE <=27	201	2.9580	0.0015
AGE <=63.471 blSOFA <=12 PRAPACHE <=27	205	2.9500	0.0016
AGE <=63.471 PRAPACHE <=26 BLLBILI<=4.8	192	2.9315	0.0017
BLGCS>11 AGE <=59.871 BLIL6>60.3	128	2.8948	0.0019
BLGCS>11 PRAPACHE <=26 BLIL6>162.65	139	2.8905	0.0019
PRAPACHE <=26 AGE <=73.138 BLGCS>10	210	2.8866	0.0019
BLGCS>11 BLLPLAT<=284 PRAPACHE <=27	208	2.8767	0.0020
PRAPACHE <=26 BLLPLAT<=284 blSOFA>3	260	2.8663	0.0021
PRAPACHE <=26 AGE <=73.138 BLLPLAT<=284	232	2.8526	0.0022
…….			

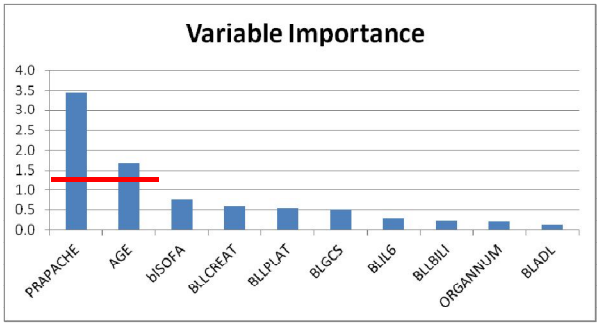
First 14 of 108 subgroups

Data Example. Screening Noise Variables Using Variable Importance

- Interpretation →Look at sharp decline in VI
- A more formal procedure for selecting variables
  - Fixed number of top covariates, e.g. 3
  - Select only those that are above a benchmark (from the dummy data sets)

Variable	Importance Based on Splitting Criterion
PRAPACHE	3.434
AGE	1.670
blSOFA	0.748
BLLCREAT	0.571
BLLPLAT	0.533
BLGCS	0.502
BLIL6	0.300
BLLBILI	0.217
ORGANNUM	0.197
BLADL	0.115
TIMFIRST	0.084
Permutation-based Mean(Max VI)	0.929
Permutation-based Std(Max VI)	0.422
Variable Screening Rule: Choose covariates above Mean(MaxVI)+1Std(MaxVI)	1.351





SIDEScreen: Screening Noise Variables Using Variable Importance

- A more formal procedure for selecting variables
  - Choose a fixed number (e.g. 3) of top covariates
  - Select only those covariates that are above a data-dependent benchmark (calibrated from generated null data sets)
- SIDEScreen – an extension of basic SIDES procedure, Lipkovich and Dmitrienko (2014)
  - Step 1: apply basic SIDES with loose complexity control, generating a large number of subgroups
  - Step 2: apply SIDES again to only selected covariates with largest VI exceeding a cut-off.
    - Calibrate the significance cut-offs for the final subgroups using data resampling, so as to ensure desired type I error rate of the entire procedure
    - This involves replicating entire 2-stage procedure on additional null data sets

SIDEScreen vs. SIDES. Data Example

- **Strategy I.** Apply SIDES with width=3 (pursuing 3 top candidates from each parent node), depth=up to 2 levels, and child-two-parent ratio (γ) =0.2

Subgroup	Subgroup size	Test statistic (z)	1-tailed p value	treatment event rate	control event rate	treatment sample size	control sample size	effect size	multiplicity -adjusted 1-sided p-value
AGE <=63.471 PRAPACHE <=26	215	2.78	0.0028	0.157	0.320	140	75	0.40	0.30
PRAPACHE <=26	313	2.06	0.0198	0.236	0.347	212	101	0.25	0.55

- **Strategy II.** Apply SIDES with a more loose "search space", with less restrictions (width=5, depth=3 and child-two-parent ratio =1) and apply SIDES again only to those covariates whose VI is above benchmark

Subgroup	Subgroup size	Test statistic (z)	1-tailed p value	treatment event rate	control event rate	treatment sample size	control sample size	effect size	multiplicity -adjusted 1-sided p-value
AGE <=63.471 PRAPACHE <=26	215	2.7752	0.0028	0.157	0.320	140	75	0.397	0.03
PRAPACHE <=26 AGE<=73.138	261	2.7179	0.0033	0.186	0.337	172	89	0.355	0.03

The same subgroup has been selected but its adjusted p-value is smaller because it passed through a more stringent two-stage screening process

Conclusions and Discussion

- SIDES is a novel method for subgroup identification incorporating 3 elements
  - Flexible algorithm for direct subgroup search/identification
  - Complexity control
  - Multiplicity control.
- Using the SIDEScreen – a multistage subgroup identification strategies with preliminary screening of noise variables by variable importance – may substantially boost the probability of indentifying true predictors, especially for data sets with larger number of covariates
- Using data-dependent reference cutoffs for VI substantially improves detection of true subgroup and filtering out noise covariates
- Performance of SIDEScreen substantially deteriorated when all noise covariates were correlated with true predictors
- Further improvement may be achieved by introducing randomness in generating candidate subgroups for the first step of SIDEScreen:

References

- Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. (2001). Subgroup analyses in randomized controlled trials: quantifying the risks of false-positives and false-negatives. Health Technology Assessment. 5:1-33
- Rothwell PM. (2005). Subgroup analysis in randomized controlled trials: importance, indications, and interpretation. *Lancet*. 365:176–86
- Sun X, Briel M, Walter SD, Guyatt GH. (2010) Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses, *British Medical Journal*. 340:850-854

References

- Kehl V and Ulm K. (2006). Responder identification in clinical trials with censored data. Computational Statistics & Data Analysis. 50:1338-1355
- Su X, Tsai CL, Wang H, Nickerson DM, Li B. (2009). Subgroup analysis via recursive partitioning. Journal of Machine Learning Research. 10:141-158.
- Lipkovich I, Dmitrienko A, Denne J, Enas G. (2011). Subgroup Identification based on Differential Effect Search (SIDES) – A recursive partitioning method for establishing response to treatment in patient subpopulations. Statistics in Medicine. 30:2601-2621
- Lipkovich, I., Dmitrienko, A. (2014). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. Journal of Biopharmaceutical Statistics. 24, 130-153.
- Foster JC, Taylor JMC, Ruberg SJ. (2011). Subgroup identification from randomized clinical trial data, Statistics in medicine. 30:2867–2880
- Berger, J., Wang, X., Shen, L. (2014). A Bayesian approach to subgroup identification. Journal of Biopharmaceutical Statistics. 24, 110-129.

References

- Jones HE, Ohlssen DI, Neuenschwander B, Racine A and Branson M. (2011). Bayesian models for subgroup analysis in clinical trials. Clinical Trials. 8:129–143
- Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. (2012). Estimating optimal treatment regimes from a classification perspective. Statistics. 1, 103-114.
- Zhao Y, Zheng D, Rush AJ, Kosorok MR. (2012). Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association. 107:1106-1118.
- Tian L, Alizaden AA, Gentles AJ, Tibshirani R. (2012). A Simple Method for Detecting Interactions between a Treatment and a Large Number of Covariates. http://arxiv.org/abs/1212.2995

Thank You!