Identification of Promising Subgroups in the Retrospective Analysis of Clinical Trials

Ilya Lipkovich, Alex Dmitrienko, Eric Su, Jonathan Denne, Gregory Enas

Eli Lilly and Company

Answers That Matter.

Basic Idea

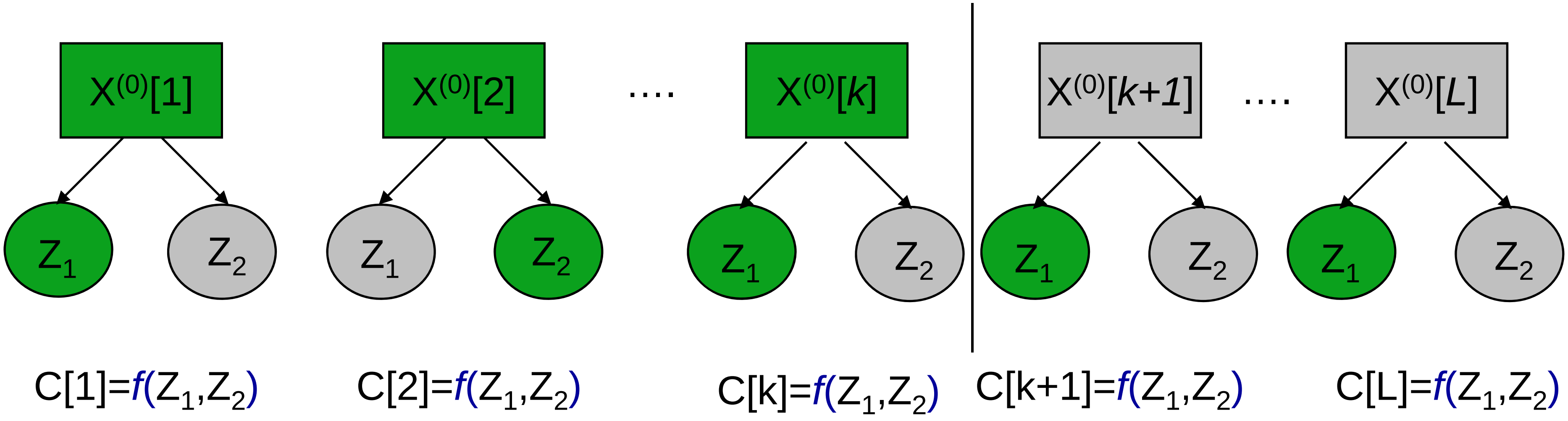
- Retrospective Data are formed:
  - Outcome (Y), treatment (T) (Drug vs. Placebo) and various subject characteristics
  - Potentially, multiple studies can be pooled
  - We assume overall treatment effect is not significant or very small ("failed studies")
- Goal: Find subgroup (s) where treatment effect is large
- Divide full data into 3 subsets of equal size, balanced with respect to treatment groups and patient characteristics
- Apply search algorithm to the exploratory data set and find best subgroup defined by subject characteristics
- Validate findings using 2 confirmatory datasets, ensuring that the overall type I error rate of the entire procedure is $<(0.05)^2=0.0025$

Pocock & Simon Allocation Procedure

- Allocate a proportion of subjects (f%) randomly into 3 subgroups
- Add subjects one by one and for each new subject:
  - Consider covariate X (with level X* for that subject)
    - compute the imbalance scores $IS_1(X^*)$, $IS_2(X^*)$, $IS_3(X^*)$, if that subject is allocated to sets 1, 2, or 3, respectively.
  - Compute total scores over all covariates: $IS_1=\Sigma_x IS_1(X^*)$, $IS_2=\Sigma_x IS_2(X^*)$, $IS_3= \Sigma_x IS_3(X^*)$
  - Allocate subject to the subgroup with smallest total imbalance score among $\{IS_1,IS_2,IS_3\}$
- The procedure guarantees with high probability that imbalance of the resulting sets with respect to the covariates will be minimal

Selecting Promising Covariates. A Tree Based Approach

- Assume there are L covariates x_i with m_i levels (i=1,..,L)
- for each candidate covariate identify the best binary split in terms of criterion C and identify *k* best covariates with promising splits such that C[*i*] > c_{cutoff}

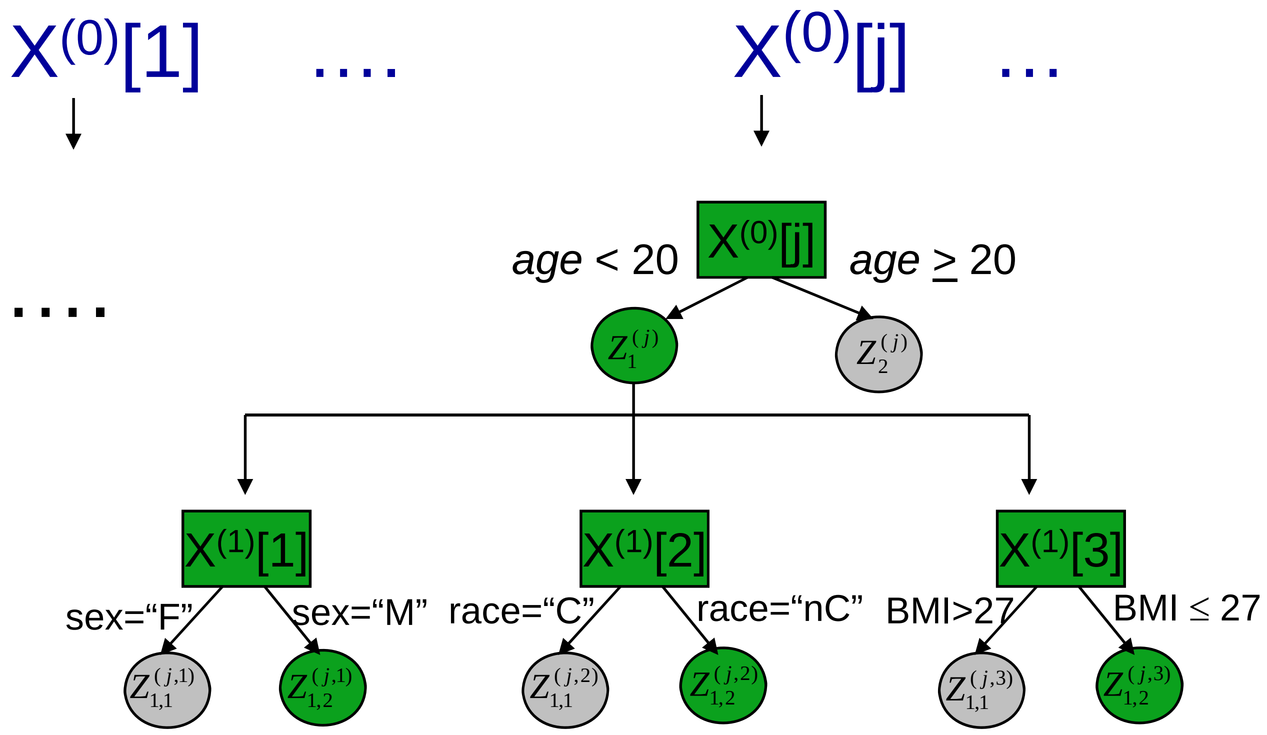


Z_1 and Z_2 are standardized treatment effects in subgroups, *f*() is discussed later

$$z_j = (\bar{y}_{j,T} - \bar{y}_{j,C})/(\bar{\sigma}_j \sqrt{1/n_{j,T} + 1/n_{j,C}}), \quad j = 1,2$$

Growing Multiple Trees

- Each of *k* selected covariates serves as a root of a tree constructed by recursively splitting the data using remaining covariates from the original set (excluding covariates already used in the current tree)



Comparing With Classical Regression Tree Methodology

- C&RT approaches look for subgroups with high level of outcome (Y)

- We are looking for subgroups with large TE

- C&RT can miss a subgroup with TE when trivial predictors that are common for treated and untreated subjects dominate the outcome

$$Y_i = f_1(X_{1i}) + f_2(X_{2i}) + TE(\boldsymbol{X}_i) + \varepsilon_i \ , \varepsilon_i \sim N(0, \sigma^2)$$

$$TE(\boldsymbol{X}_{\boldsymbol{i}}) = \{b_1 I(X_{1i} = X_1^*) + b_2(I(X_{2i} = X_2^*) - b\} I(T_i = 1)$$

Trees: How Many and How Large?

- Specifying cut-offs of splitting criterion at each level
  - Splitting criterion statistic at each level = adjusted p-value for *treatment-by-split interaction*:
    - $C=f(Z_1,Z_2) = 2(1- \Phi\{|Z_1-Z_2|/\sqrt{2}\})*\{\#\text{of possible splits}\}$,
  - Nominal alphas at levels 1,2,3 (say $\alpha_0 = .1$, $\alpha_0 =.05$, $\alpha_0 =.01$)
  - Then level-specific cut-offs for criterion *C* are based on null distribution of criterion statistic

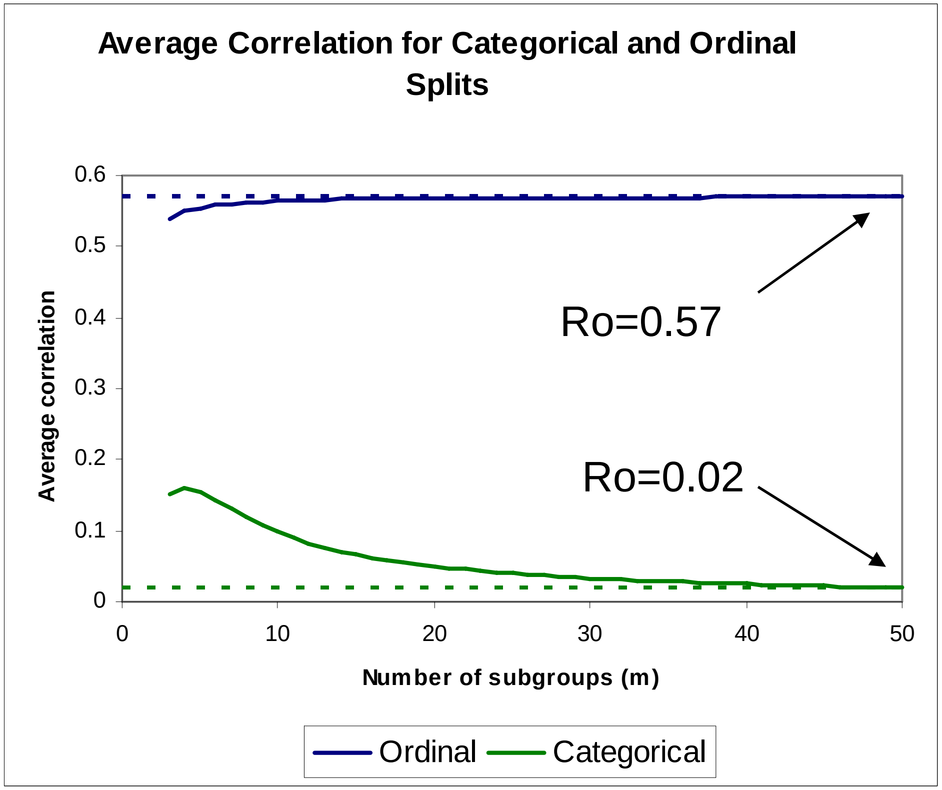
- Imposing constraints on:
  - upper limit on number of variables that serve as new roots (e.g. =5)
  - upper limit of nesting (e.g. =3)
  - lower limit on size of a subgroup (e.g. N=30)
  - upper limit on total number of comparisons

Adjusting for Multiple Comparisons per Covariate

$$Bonf\,(m) = M_{eff} = M^{1-\bar{\rho}}$$

$$M_{categorcial} = 2^{m-1} - 1,\ \text{e.g.}\,(1,0,1,0,1,..,0)$$

$$M_{ordinal} = m - 1,\qquad \text{e.g.}\,(1,1,1,0,0,..,0)$$



Number of levels (m)	Categorical splits		Ordinal Splits	
	# of splits, M	effective number of splits, $M^{(1-\rho)}$	# of splits, M	effective number of splits, $M^{(1-\rho)}$
3	3	2.5	2	1.4
4	7	5.1	3	1.6
5	15	9.9	4	1.9
6	31	19.0	5	2.0
7	63	36.8	6	2.2
8	127	71.6	7	2.3
9	255	140.6	8	2.5
10	511	277.3	9	2.6

Obtaining the Null Distribution for Splitting Criteria

- Data sets under H_0 are constructed by standardizing within treatment groups and permuting treatment labels
  - This is consistent with randomization: it only breaks relationship between y and treatment while preserving relationship between y and covariates
    - Note that any relationship between treatment and covariates should be irrelevant due to randomization

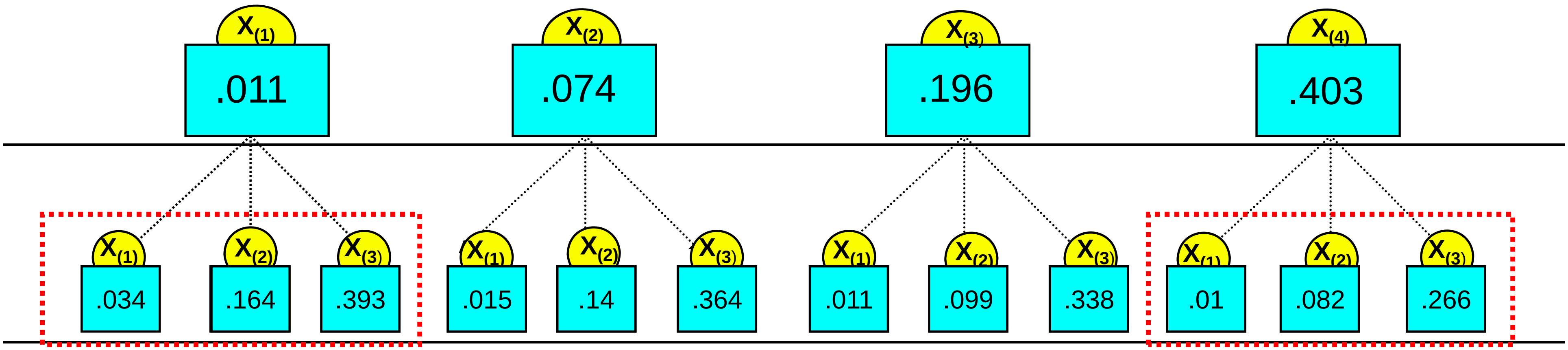
- Compute adjusted criterion *C** for every possible configuration (defined by order $j_0, j_1, ..j_{lev}$ of covariates selected at current and previous levels)

- Repeat many (1,000) times and compute cut-offs at each level for any desired nominal alpha

$$\frac{\#(C^{*}(j_0,.., j_{curlevel}) < cutoff)}{N_{perm}} = \alpha(culevel)$$

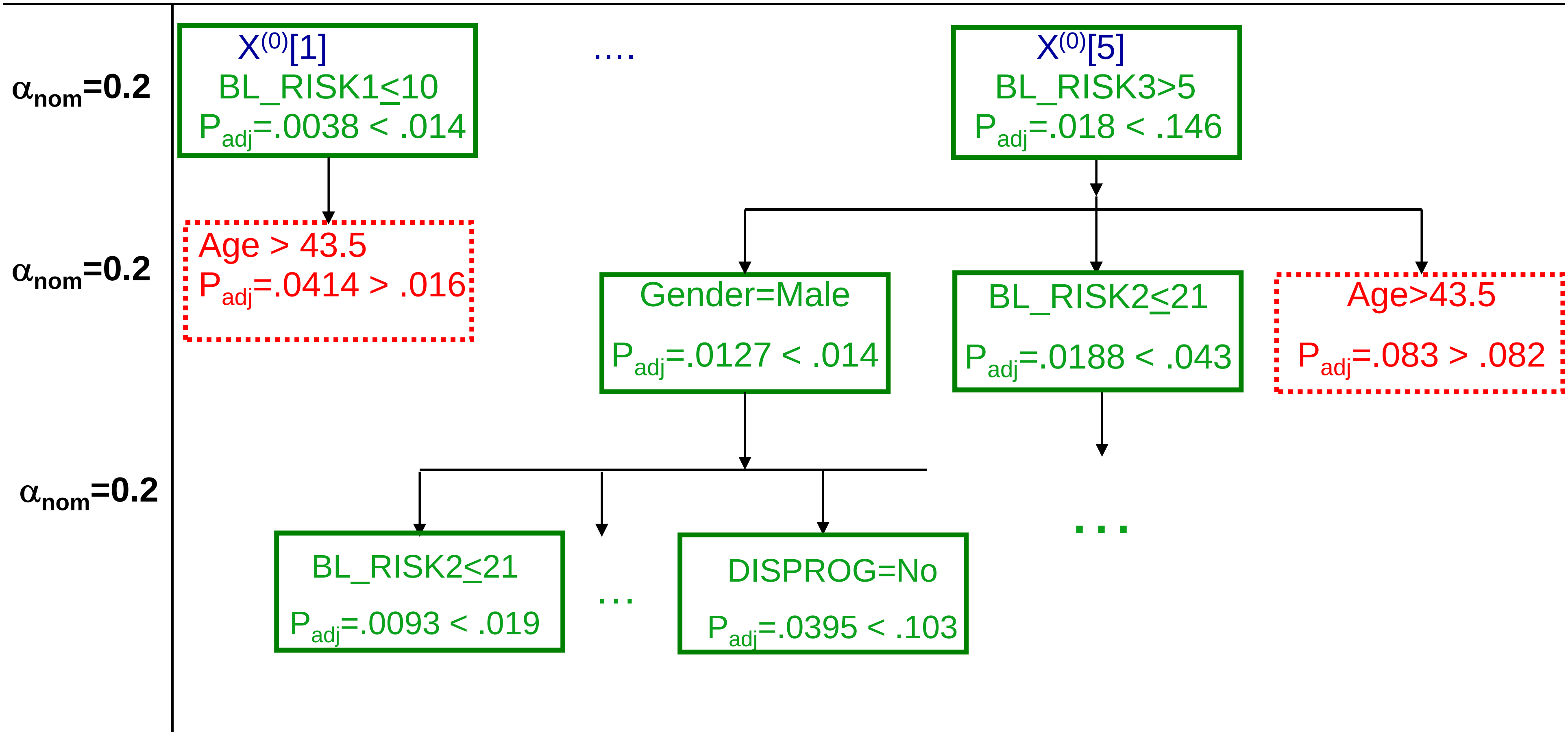
Splitting criterion Cut-offs from Permutation Null Distribution

Test statistic (criterion) is $p=2(1-\Phi\{|Z_1-Z_2|/\sqrt{2}\})$



- Nominal α = 0.05 was used at all 3 levels, all variables have 2 categories
- $X_{(1)}$, $X_{(2)}$, etc refer to variables ordered by the criterion, from best to worst;
- the same variable cannot appear more than once along the same path
- The cut-offs are conditional on the current level and order of covariate selected at higher level(s)

Illustrating Recursive Partitioning Procedure With Clinical Trial Data



α_{nom}**=0.2**

$X^{(0)}[1]$
BL_RISK1≤10
$P_{adj}=.0038 < .014$

….

$X^{(0)}[5]$
BL_RISK3>5
$P_{adj}=.018 < .146$

α_{nom}**=0.2**

Age > 43.5
$P_{adj}=.0414 > .016$

Gender=Male
$P_{adj}=.0127 < .014$

BL_RISK2≤21
$P_{adj}=.0188 < .043$

Age>43.5
$P_{adj}=.083 > .082$

α_{nom}**=0.2**

BL_RISK2≤21
$P_{adj}=.0093 < .019$

…

DISPROG=No
$P_{adj}=.0395 < .103$

…

Drug A vs. Placebo. P values based on a Chi-square test for categorical outcome
Number of max sub trees at every level is limited to 5

Top Subgroups Identified

Subgroups found in exploratory set	Exploratory set			Test set
	N (sub-group)	asymptotic Z score	P value	P value
BL_RISK2 ≤ **21 and BL_RISK3** > **5 and GENDER**=(**Male**)	43	3.30	.00049	.03898
AGE > **43.51 and BL_RISK2** ≤ **21 and BL_RISK4** ≤ **3**	183	3.29	.00049	.26106
BL_RISK5 ≤ **25 and BL_RISK3** > **5 and GENDER** =(**Male**)	45	3.28	.00052	.01204
BL_RISK2 ≤ **21 and BL_RISK4** ≤ **3 and ORIGIN**=(**Caucasian**)	169	3.16	.00080	.34738
….				

Drug A vs. Placebo. P values based on a Chi-square test for categorical outcome
Data divided into exploratory and a single test set

Simulating Data With Treatment Effect Within Subgroups

True subgroup: X1={1}, X2={2}

	X1=1	X1=2
X2=1	0.25	-0.75
X2=2	1.25	0.25
X2=3	0.25	-0.75
X2=4	0.25	-0.75

- TE is the sum of effects from each “contributing” variable:
- Overall TE is zero

$$Y_i = TE(X_i) + \varepsilon_i \ , \varepsilon_i \sim N(0, \sigma^2)$$

$$TE(X_i) = a \sum_{j=1}^{m_e} \left\{ I(X_{ij} = X_j^*)(1 - \frac{n_j}{N}) - I(X_{ij} \neq X_j^*) \frac{n_j}{N} \right\} I(T_i = 1)$$

Quantifying “Success” for Simulation Study. Proportion of TE Recovered

$$\%TE(captured\,/\,true) = 100\% \frac{|S_{found}|^{-1} \sum_{i \in S_{found}} TE_i}{|S_{true}|^{-1} \sum_{i \in S_{true}} TE_i}$$

S_{found} the set of all *treated* subjects identified as the best subgroup by the algorithm and confirmed by 2 validation sets

S_{true} the set of *treated* subjects in the “true best subgroup”

Simulation Results

Correct group: “X1=0”, n=150, corr(X)=0

Total # of covariates	Assumed treatment effect in correct subgroup	Multiple δ for TE with 80% power on full data	Power for TE in the correct subgroup,% **(1-β)**	Power for confirmed TE,% $(\mathbf{1-\beta})^3$	Proportion of effective runs,%	Proportion of confirmed runs,%	TE Recovered/ TE in correct subgroup, %	Size of confirmed subgroup
5	0	0	2.5	0.002	3.5	0.00		
	0.364	1.95	60	21.6	70.94	21.8	100	145.39
	0.460	2.46	80	51.2	91.44	51.58	100	143.84
10	0	0	2.5	0.002	5.0	0.02		
	0.364	1.95	60	21.6	65.40	20.14	100	144.20
	0.460	2.46	80	51.2	87.02	46.44	100	144.02
20	0	0	2.5	0.002	4.68	0.00		
	0.364	1.95	60	21.6	54.64	15.18	100	145.59
	0.460	2.46	80	51.2	82.34	42.94	100	144.52

Assumed TE in full data = 0
Assumed TE for correct subgroup = δ **x** (TE that would give 80% power in full data)
N (full data) =900, number of simulated data sets =5,000

Simulation Results

Correct group: “X1=0,X2=0,X3=0”,n=150, corr(X)=0

Total # of covariates	Assumed treatment effect in correct subgroup	Multiple δ for TE with 80% power on full data	Power for TE in the correct subgroup,% **(1-β)**	Power for confirmed TE,% **(1-β)^3**	Proportion of effective runs,%	Proportion of confirmed runs,%	TE Recovered/ TE in correct subgroup, %	Size of confirmed subgroup
5	0	0	2.5	0.002	3.5	0.00		
	0.364	1.95	60	21.6	50.66	9.94	92.39	156.38
	0.460	2.46	80	51.2	76.44	33.76	94.83	153.64
10	0	0	2.5	0.002	5.0	0.02		
	0.364	1.95	60	21.6	43.12	5.76	86.94	157.58
	0.460	2.46	80	51.2	61.24	20.94	90.88	156.35
20	0	0	2.5	0.002	4.7	0.00		
	0.364	1.95	60	21.6	25.90	2.08	79.46	164.56
	0.460	2.46	80	51.2	45.34	11.60	87.38	157.09

Assumed TE in full data = 0

Assumed TE for correct subgroup = **δ x** (TE that would give 80% power in full data)

Full data =900, number of simulated data sets =5,000

Simulation Results

Correct group: “X1=0”, n=150, corr(X)=0.3

Total # of covariates	Assumed treatment effect in correct subgroup	Multiple δ for TE with 80% power on full data	Power for TE in the correct subgroup,% **(1-β)**	Power for confirmed TE,% **$(1-\beta)^3$**	Proportion of effective runs,%	Proportion of confirmed runs,%	TE Recovered/ TE in correct subgroup, %	Size of confirmed subgroup
5	0	0	2.5	0.002	4.96	0.00		
	0.364	1.95	62	23.8	75.96	24.82	100	142.70
	0.460	2.46	82	55.1	92.38	52.28	100	139.00
10	0	0	2.5	0.002	4.62	0.00		
	0.364	1.95	62	21.6	69.96	21.30	99.93	140.23
	0.460	2.46	82	51.2	88.98	46.18	99.99	138.33
20	0	0	2.5	0.002	4.80	0.00		
	0.364	1.95	62	23.8	59.12	17.12	99.88	140.7
	0.460	2.46	82	55.1	85.68	42.78	99.95	136.5

Assumed TE in full data = 0
Assumed TE for correct subgroup = δ **x** (TE that would give 80% power in full data)
Full data =900, number of simulated data sets =5,000

Simulation Results

Correct group: "X1=0,X2=0,X3=0", n ≈ 168,corr(X)=0.3

Total # of covariates	Assumed treatment effect in correct subgroup	Multiple δ for TE with 80% power on full data	Power for TE in the correct subgroup,% **(1-β)**	Power for confirmed TE,% **$(1-\beta)^3$**	Proportion of effective runs,%	Proportion of confirmed runs,%	TE Recovered/ TE in correct subgroup, %	Size of confirmed subgroup
5	0	0	2.5	0.002	4.96	0.00		
	0.364	1.95	62	23.8	87.86	26.64	90.00	173.53
	0.460	2.46	82	55.1	97.30	58.70	92.29	171.29
10	0	0	2.5	0.002	4.62	0.00		
	0.364	1.95	62	21.6	78.34	16.84	84.13	173.59
	0.460	2.46	82	51.2	93.42	46.34	87.52	169.90
20	0	0	2.5	0.002	4.80	0.00		
	0.364	1.95	62	23.8	70.44	11.80	79.32	173.01
	0.460	2.46	82	55.1	92.32	36.58	82.60	170.08

Assumed TE in full data = 0
Assumed TE for correct subgroup = **δ x** (TE that would give 80% power in full data)
Full data =900, number of simulated data sets =5,000

Simulation Results. Distribution of Confirmed Runs

Correct group: “X1=0, X2=0”, n ≈ 161,corr(X)=0.3

Total # of covariates	Assumed treatment effect in correct subgroup	Proportion of confirmed runs,%	Size of confirmed subgroup	% complete match	% undershoot **choosing $x_1=0$ or $x_2=0$**	% overshoot **choosing $x_1=0\&x_2=0\&x_3=\{0 \text{ or } 1\}$**	% overlap	% complete miss
5	0.364	21.7	155.6	48.21	12.05	34.77	4.97	0.00
	0.460	55.5	149.1	49.42	4.66	41.79	4.13	0.00
10	0.364	14.8	154.1	28.34	17.68	40.35	13.50	0.93
	0.460	45.0	144.2	29.35	6.26	53.42	10.79	0.22
20	0.364	10.8	151.8	20.37	18.15	41.85	18.70	0.93
	0.460	36.1	147.0	23.88	11.19	47.81	16.90	0.22

Next Steps

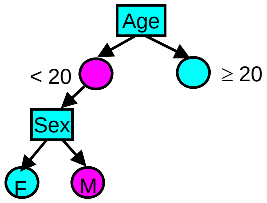
- The performance of the algorithm can be improved by calibrating various tuning parameters
  - Nominal alphas at each level, α_1, α_2, α_3
  - Number of covariates (levels) γ in defining the best subgroup
- Tuning parameter can be calibrated via bootstrap or cross-validation
- The solution (optimal subgroup) given change in tuning parameters should be obtained fast (without re-computing permutation distribution for the criterion)

Illustration of k-fold Cross-validation For Choosing γ=Number of levels

Best subgroup
on training folds

Exploratory data

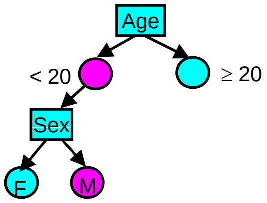
$\gamma=2$



train | train | train | train | valid

P-value for TE

$\gamma=2$

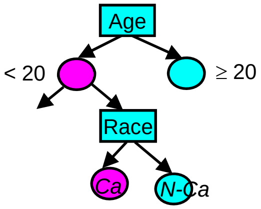


train | train | train | valid | train

P-value for TE

. . . . .

$\gamma=2$



valid | train | train | train | train

P-value for TE

Compute average P-value(γ) from all validation folds
Choose γ*=***argmin*****{**av_Pvalue(γ)**}**

Discussion

- A novel tree-based procedure is proposed as a “salvaging strategy” for failed studies. This approach can also can be used as an exploratory tool for hypothesis generation
- The rate of treatment effect recovered in confirmed subgroup is $\approx$90% of the maximal TE
- When the number of potential covariates is small ($\leq$ 5) the rates of confirmed sub-groups are comparable with the rates of success using 2 confirmation data sets, if the true subgroup were known (an ideal benchmark)
- With larger number of candidate covariates ($\geq$ 10) the rates of confirmed runs may drop substantially compared with the “ideal benchmark”
- The effect of correlation in covariates appears to
  - improve the rate of “confirmed subsets”, however
  - at the expense of poorer match with the true subsets (confirmed subsets may partially overlap the true subset)