Recursive Partitioning Method on Survival Outcomes for Personalized Medicine

Wei Xu, Ph.D

Dalla Lana School of Public Health, University of Toronto
Princess Margaret Cancer Centre

2nd International Conference on Predictive, Preventive and Personalized Medicine & Molecular Diagnostics

Outline

- Statistical consideration on personalized medicine
- Recursive partitioning method: prognostic tree and predictive tree on cancer research
- Model construction and simulations
- Application to a GWAS study on randomized clinical trial data
- Conclusions and further directions

Personalized Medicine

- Personalized Medicine is the idea of getting the right treatment on the right people based on their demographic, clinical, genetic, and genomic characteristics

- This has seen by some medical researchers and pharmaceutical industries as the future of healthcare

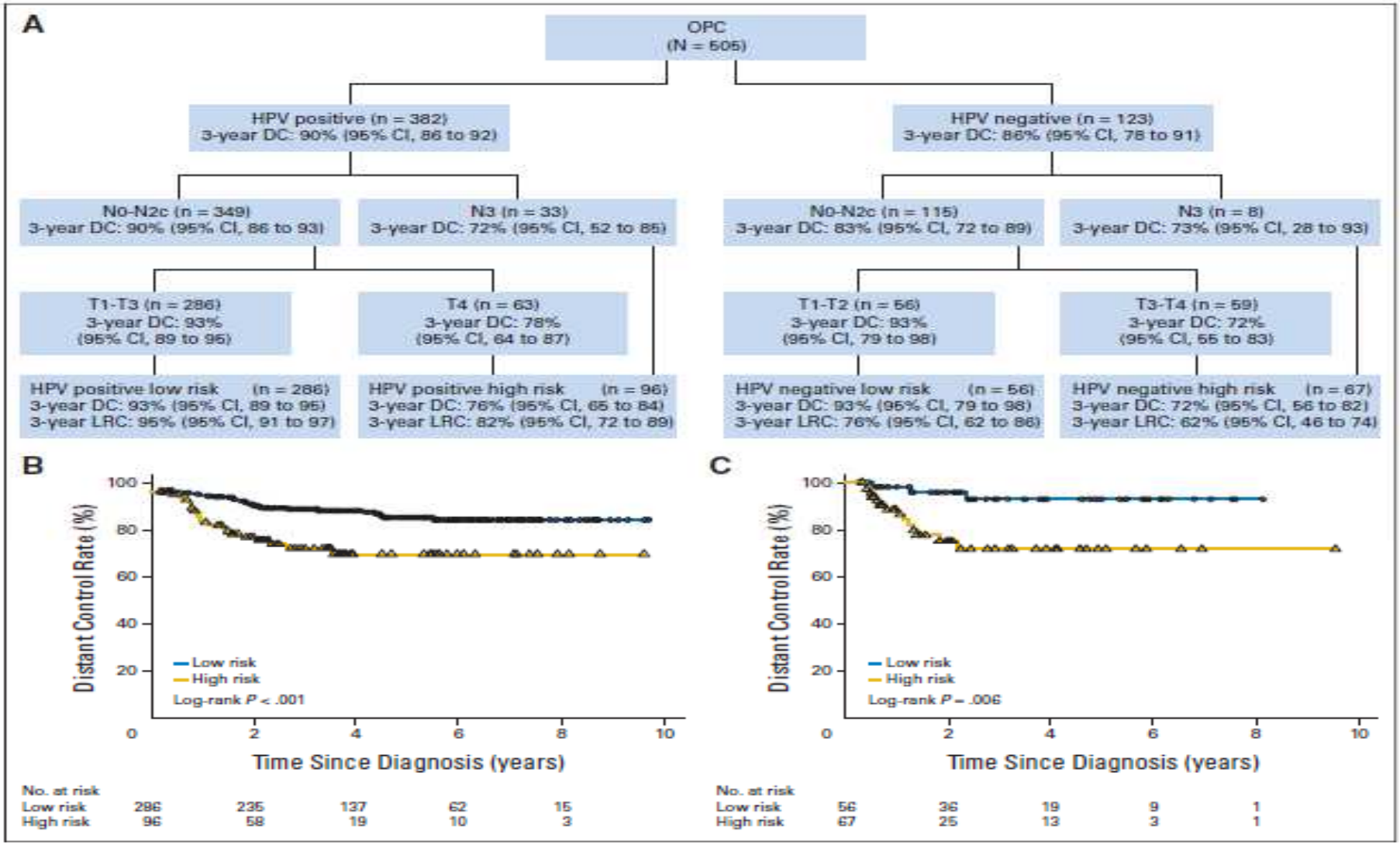
Statistical Consideration

- Statistical methods to help enable personalized medicine in cancer research face many challenges.
  - High dimensional genetic and clinical data
  - Work on censoring outcomes such as OS and PFS
  - Treatment interactions with multiple risk factor on cancer outcomes
  - Classify the patient population into homogeneous subgroups based on the covariate space
  - The analytic results should have clear clinical interpretation

Recursive Partitioning Methods

- This model is natural for personalized medicine as they partition the covariate space in a way that mimics the clinicians' natural decision making process.

- They have been used to create interpretable tools to classify prognosis (i.e. prognostic survival tree).



Deintensification Candidate Subgroups in Human Papillomavirus–Related Oropharyngeal Cancer According to Minimal Risk of Distant Metastasis

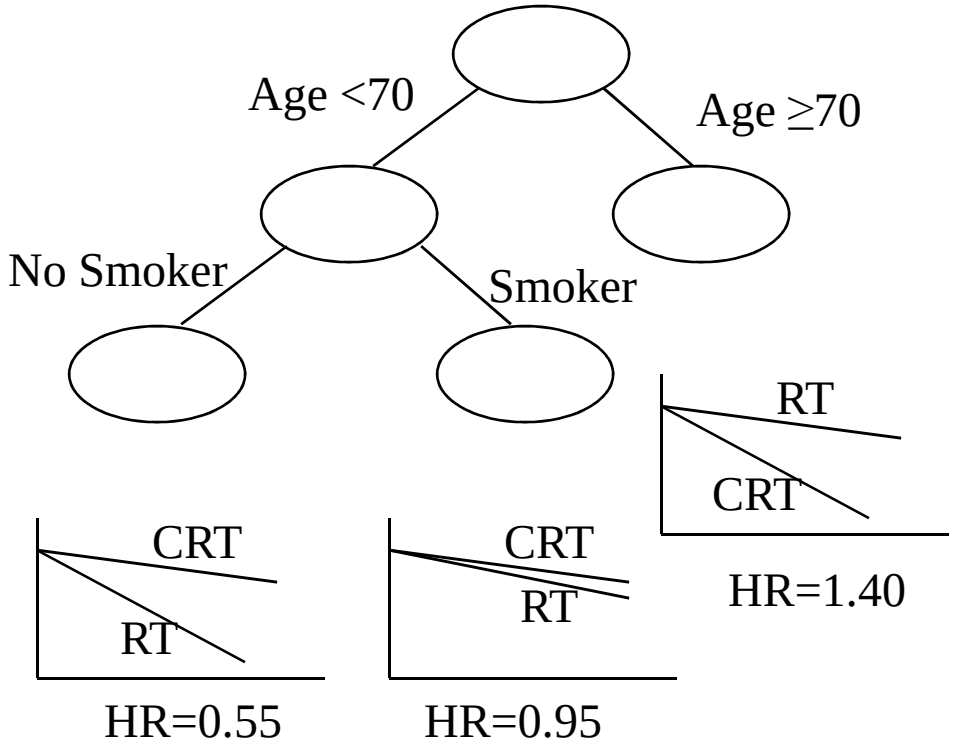
Brian O'Sullivan, Shao Hui Huang, Lillian L. Siu, John Waldron, Helen Zhao, Bayardo Perez-Ordonez, Ilan Weinreb, John Kim, Jolie Ringash, Andrew Bayley, Laura A. Dawson, Andrew Hope, John Cho, Jonathan Irish, Ralph Gilbert, Patrick Gullane, Angela Hui, Fei-Fei Liu, Eric Chen, and Wei Xu

VOLUME 31 · NUMBER 5 · FEBRUARY 10 2013

JOURNAL OF CLINICAL ONCOLOGY

Predictive Survival Tree

- We extend survival trees to partition the covariate space based on having large differences in response to ***treatment***. These methods could be used to create interpretable tools that help on the best treatment decision of patients (treatment interaction).

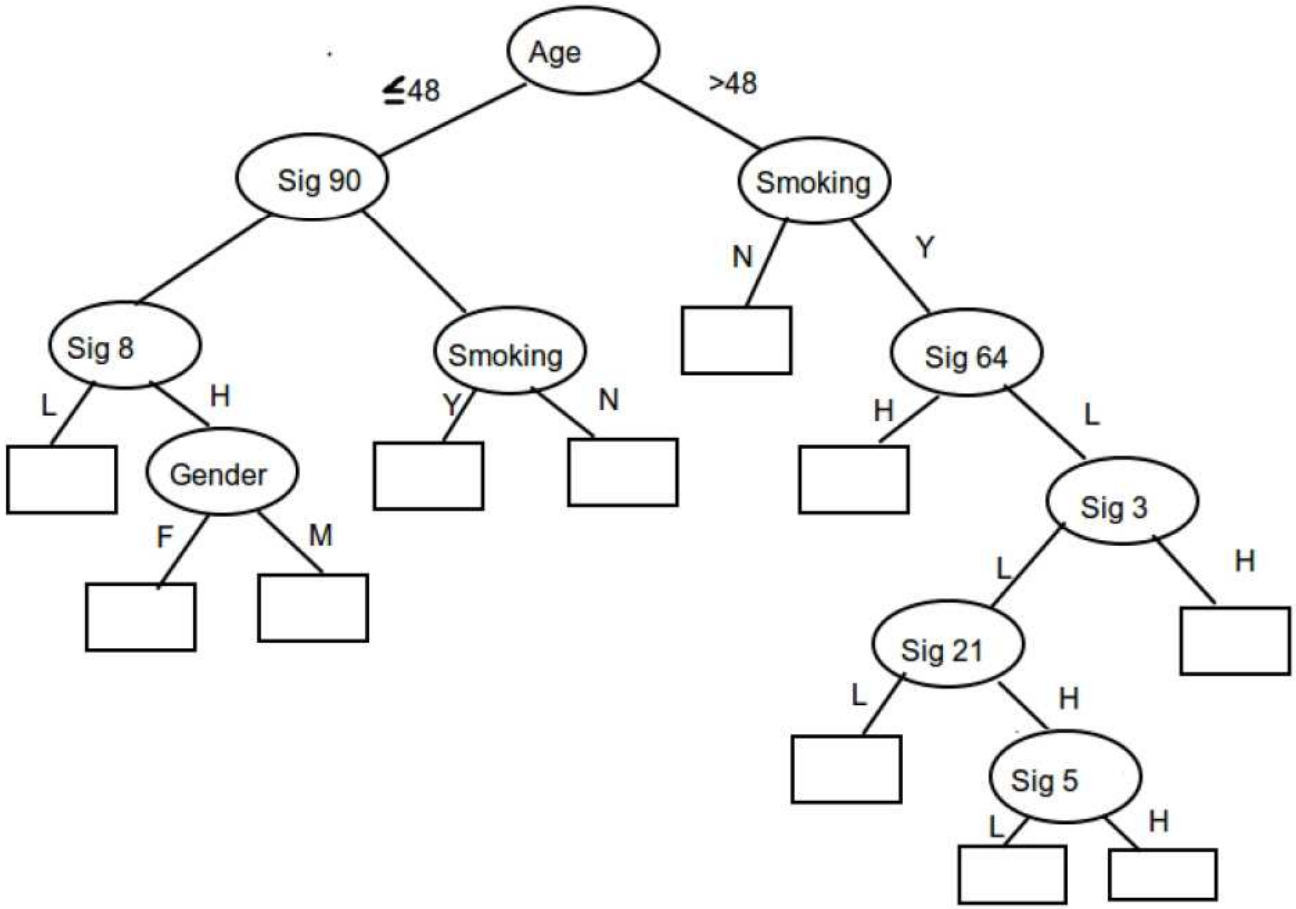


Recursive Partitioning Algorithm

- Tree based models usually consist of three parts:
  - Splitting rule
  - Pruning algorithm
  - Final tree selection
- For the predictive survival tree, all three parts need to be re-developed to reflect the difference of treatment effect for the subgroups.

Splitting Rule

- The splitting rule partitions the samples into many groups. It is applied recursively until there are very few samples in each group, or large number of groups are created.



Splitting Rule

To partition a node h, find the split s such that some measure of improvement $G(s, h)$ is maximized.

$$G(s^*, h) = \max_{s \in S_h} G(s, h)$$

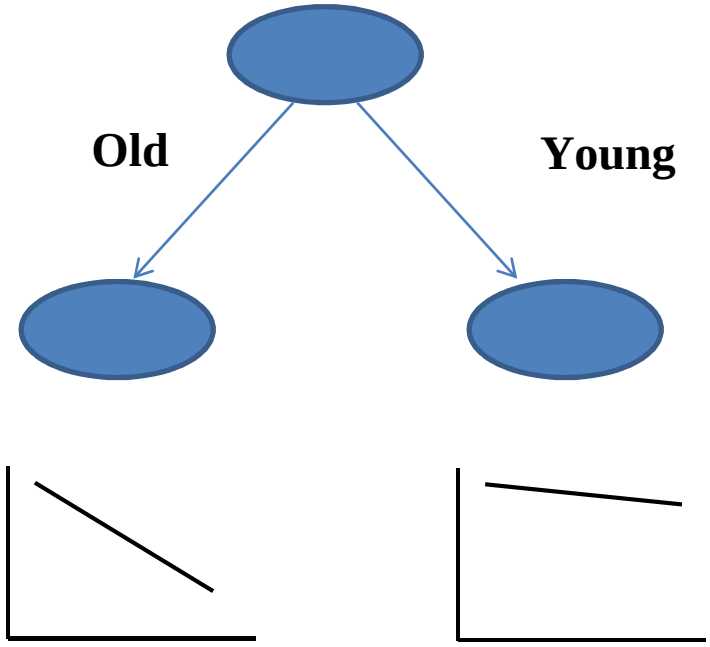
If there is more than one terminal node to partition then find the best split s^* for each $h \in H$ and split the node with the maximal improvement.

$$G(s^{**}, h^*) = \max_{(s^*, h) \in S_H^*} G(s^*, h)$$

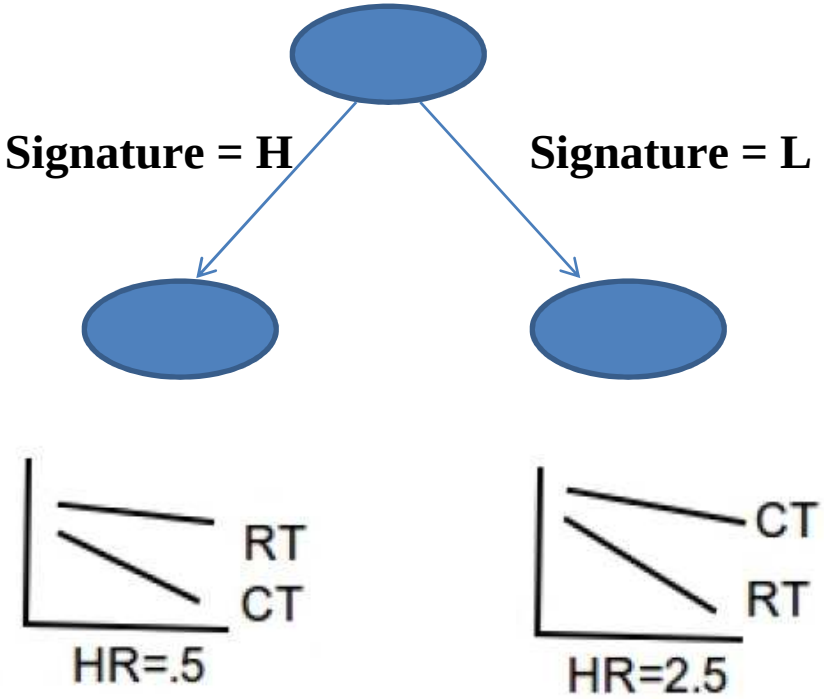
If either of these maximum are not unique, randomly select one of the tied splits.

Different Splitting Rules

Prognostic Tree Split



Predictive Tree Split



Splitting Rule for Predictive Survival Tree

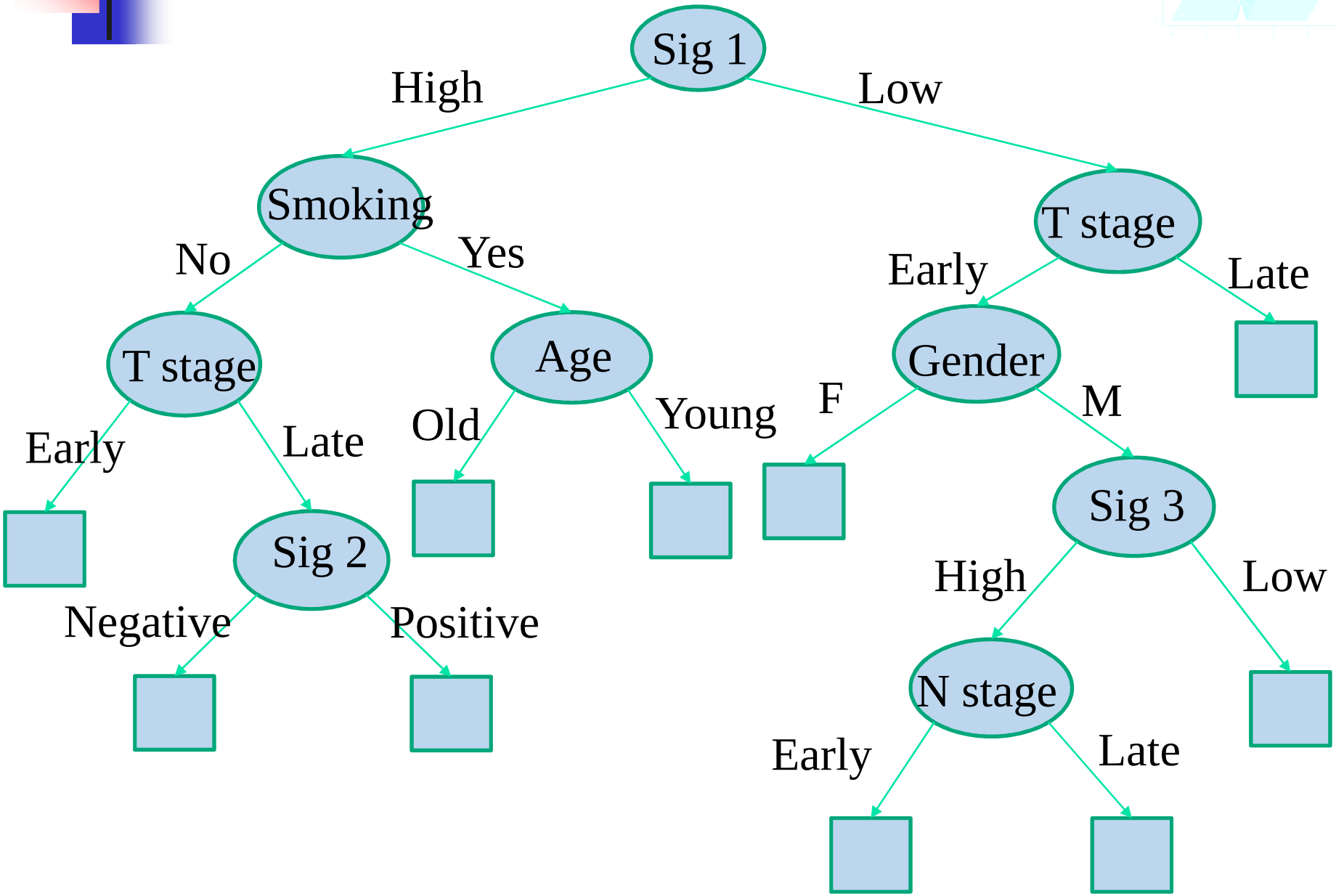
The best split can be interpreted as the one that creates the two child nodes with the most statistically significant difference in response (i.e. OS) to treatments.

Definition

$G_{ci}(s,h)$ is the LRT statistic corresponding to $H_0 : \beta_{ts} = 0$ in the Cox model $\log(\lambda) = \vec{\beta_c}\vec{X}_c + \beta_t X_t + \beta_s X_s + \beta_{ts}(X_t \times X_s)$

- $\vec{X}_c$ is a vector of confounding variables
- X_s is an indicator of the potential split s, which is a binary partition of some covariate c
- X_t is a treatment with ≥ 2 levels
- $X_t \times X_s$ is the interactive term of the treatment and the split s.

Full Tree


Sig 1
High
Low
Smoking
T stage
No
Yes
Early
Late
T stage
Age
Gender
Early
Late
Old
Young
F
M
Sig 3
Sig 2
High
Low
Negative
Positive
N stage
Early
Late

Pruning

This large tree overfits the data and will perform poorly out-of-sample, thus a subtree must be chosen as the final tree. The space of all subtrees is large, and an efficient pruning algorithm is used to find all optimal subtrees.

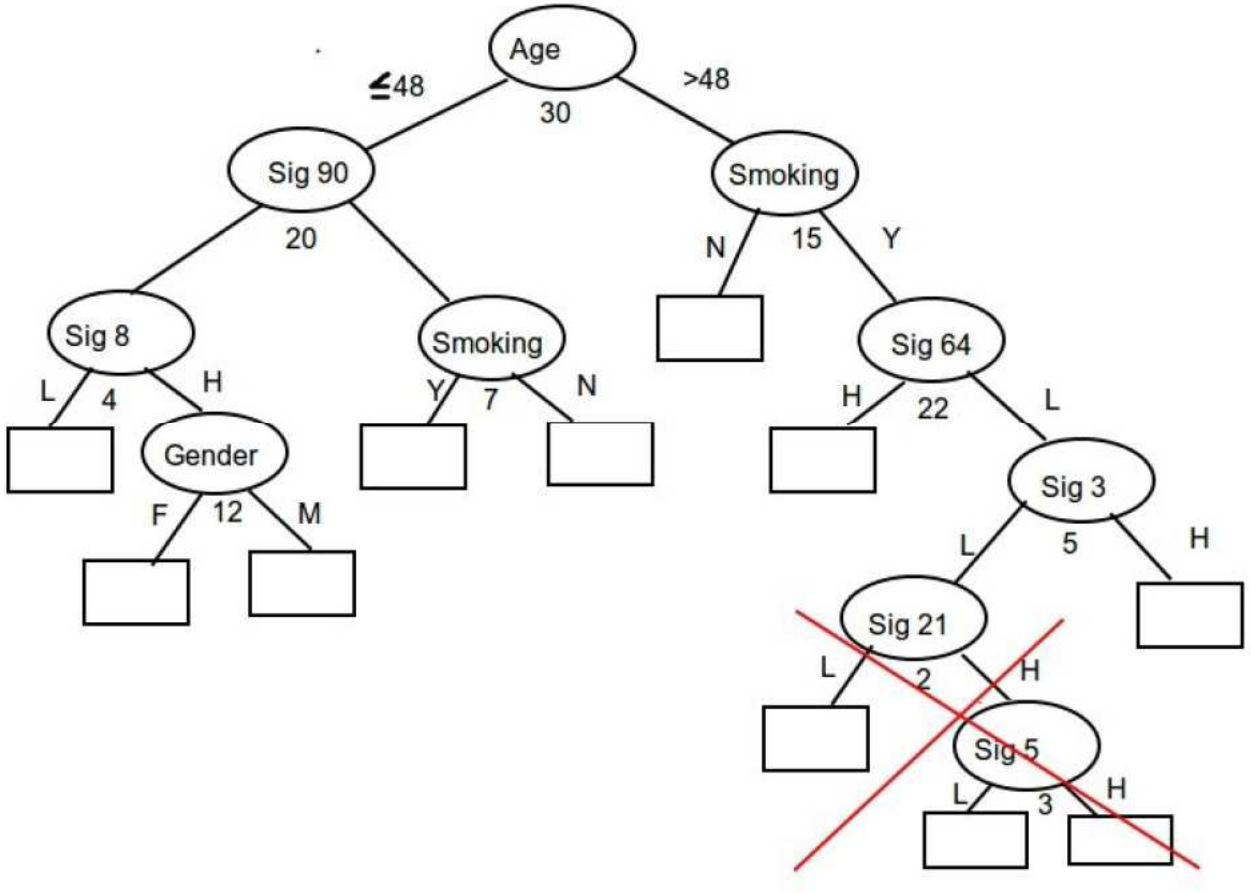
The the split-complexity $G_\alpha(T)$ can be defined as

$$G_\alpha(T) = G(T) - \alpha|S|$$

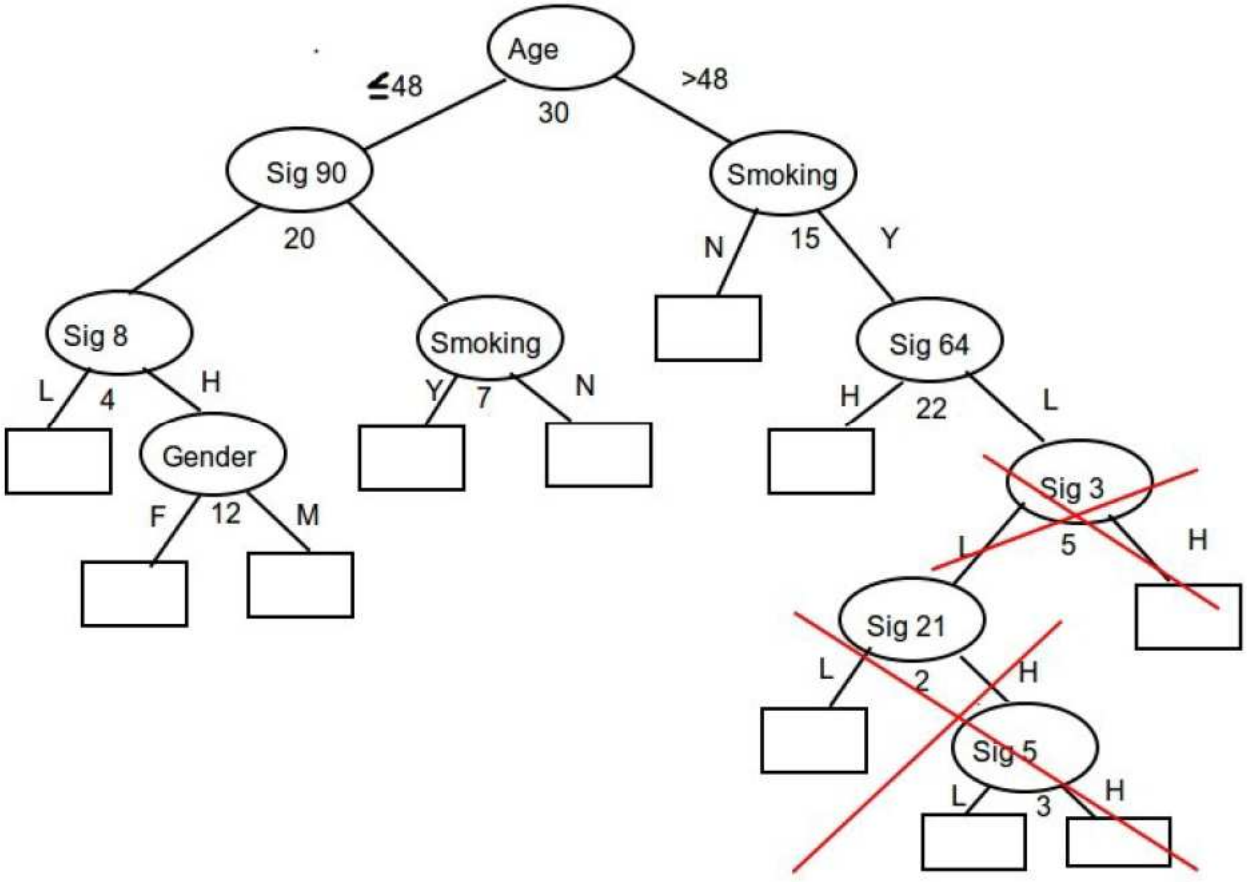
Where S is the set of internal nodes of tree T, $|S|$ is the cardinality of S, $\alpha \geq 0$ is the complexity parameter, and $G(T)$, the goodness of split of tree T, is the sum of the split improvement statistics over the tree.

$$G(T) = \sum_{h \in S} G(h)$$

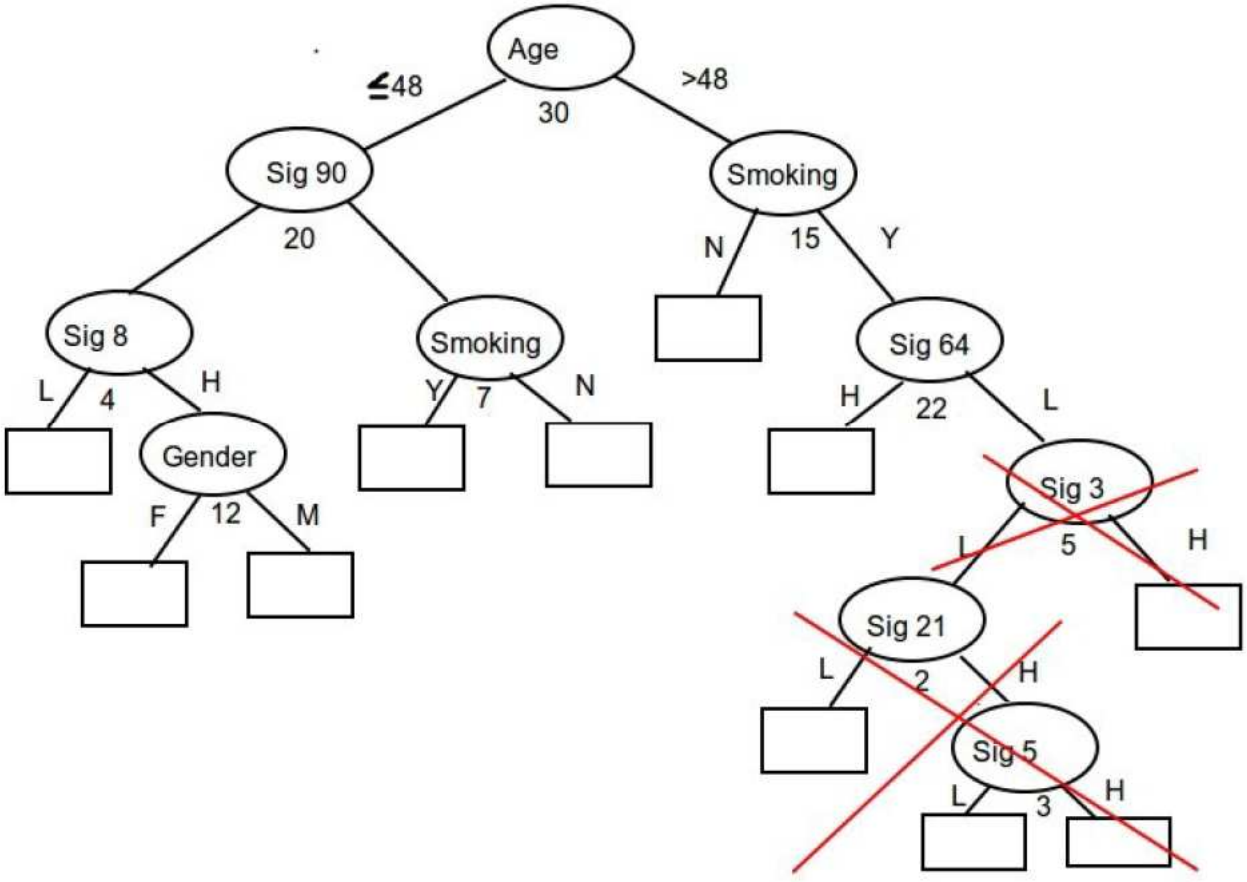
Pruning


Age
≤48
30
>48
Sig 90
20
Smoking
N
15
Y
Sig 8
L
4
H
Smoking
Y
7
N
Sig 64
H
22
L
Gender
F
12
M
Sig 3
L
5
H
Sig 21
L
2
H
Sig 5
L
3
H

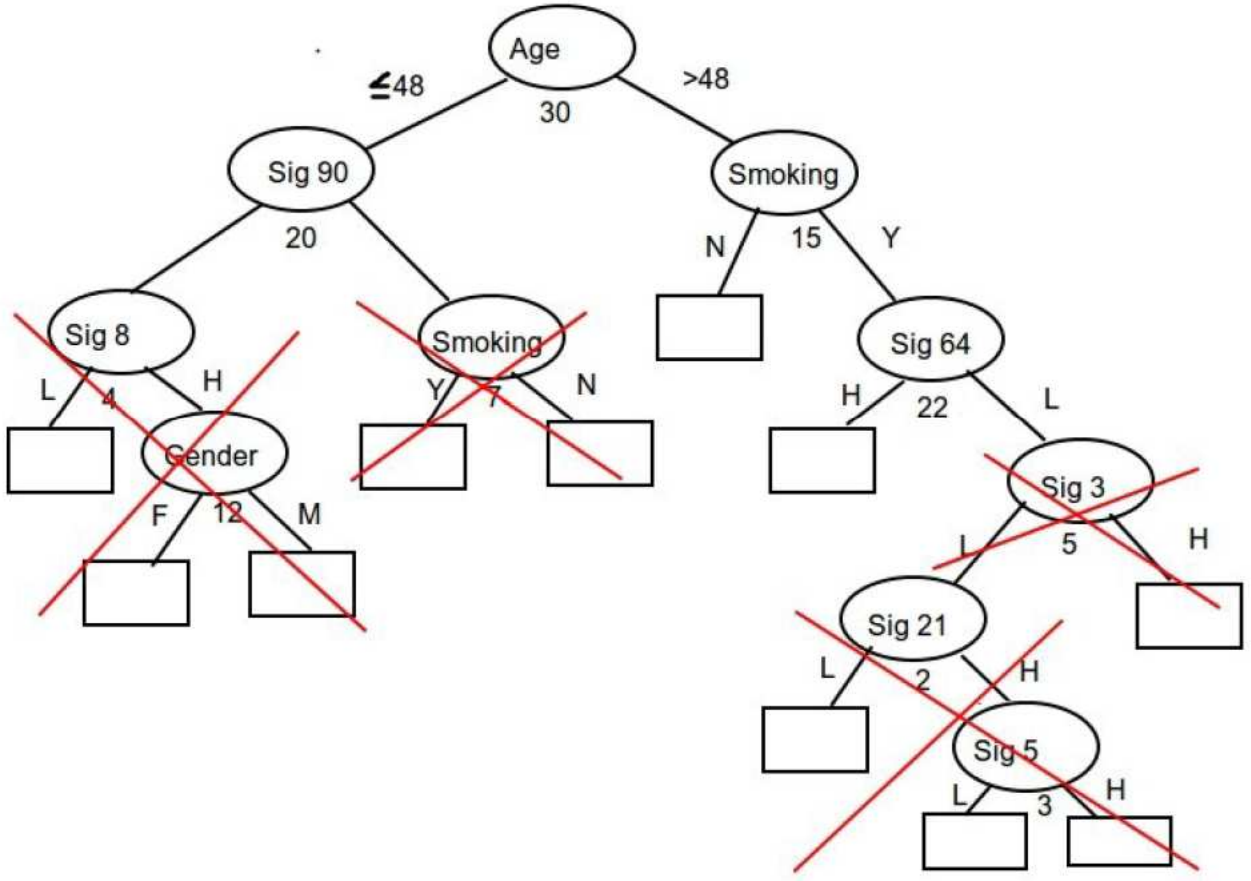
Pruning


Age
≤48
30
>48
Sig 90
20
Smoking
N
15
Y
Sig 8
L
4
H
Smoking
Y
7
N
Sig 64
H
22
L
Gender
F
12
M
Sig 3
L
5
H
Sig 21
L
2
H
Sig 5
L
3
H

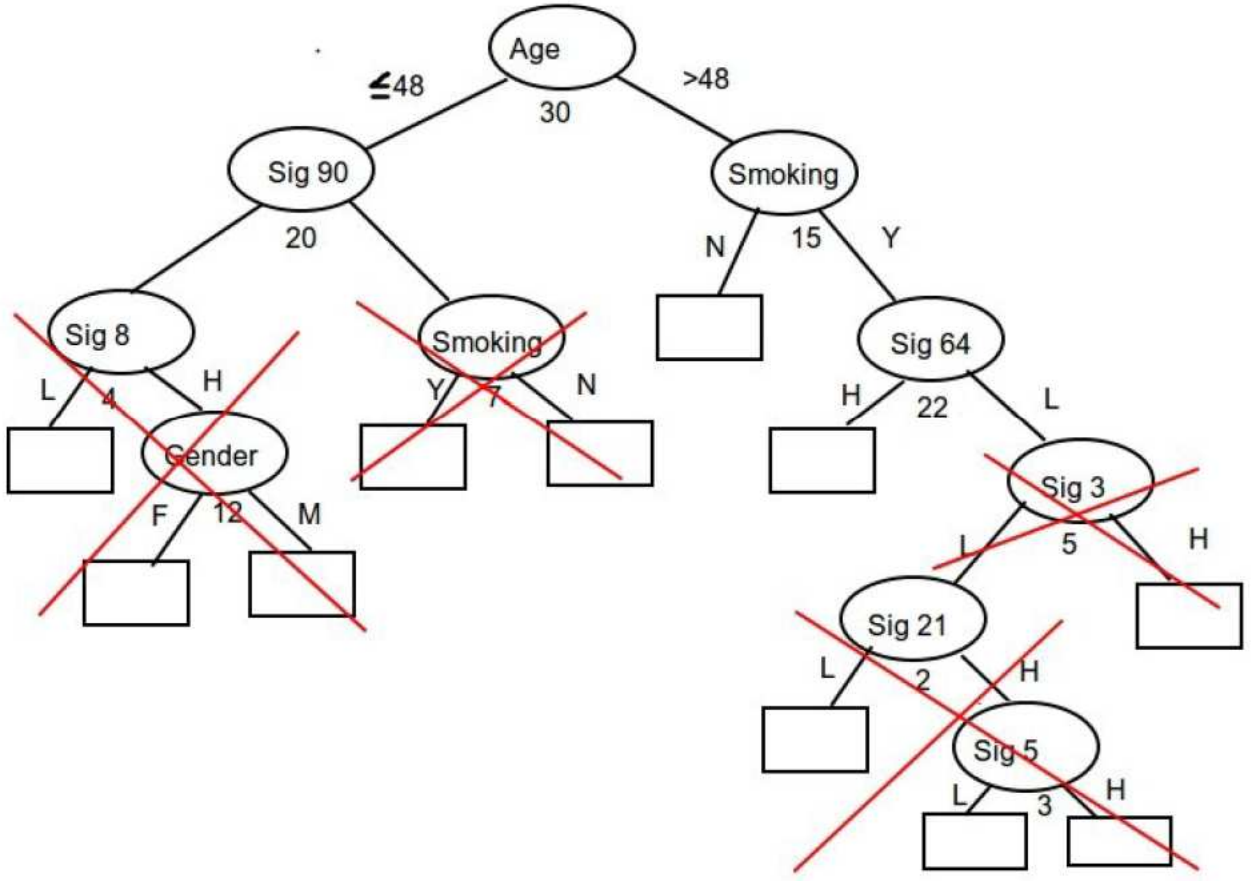
Pruning


Age
≤48
30
>48
Sig 90
20
Smoking
N
15
Y
Sig 8
L
4
H
Smoking
Y
7
N
Sig 64
H
22
L
Gender
F
12
M
Sig 3
L
5
H
Sig 21
L
2
H
Sig 5
L
3
H

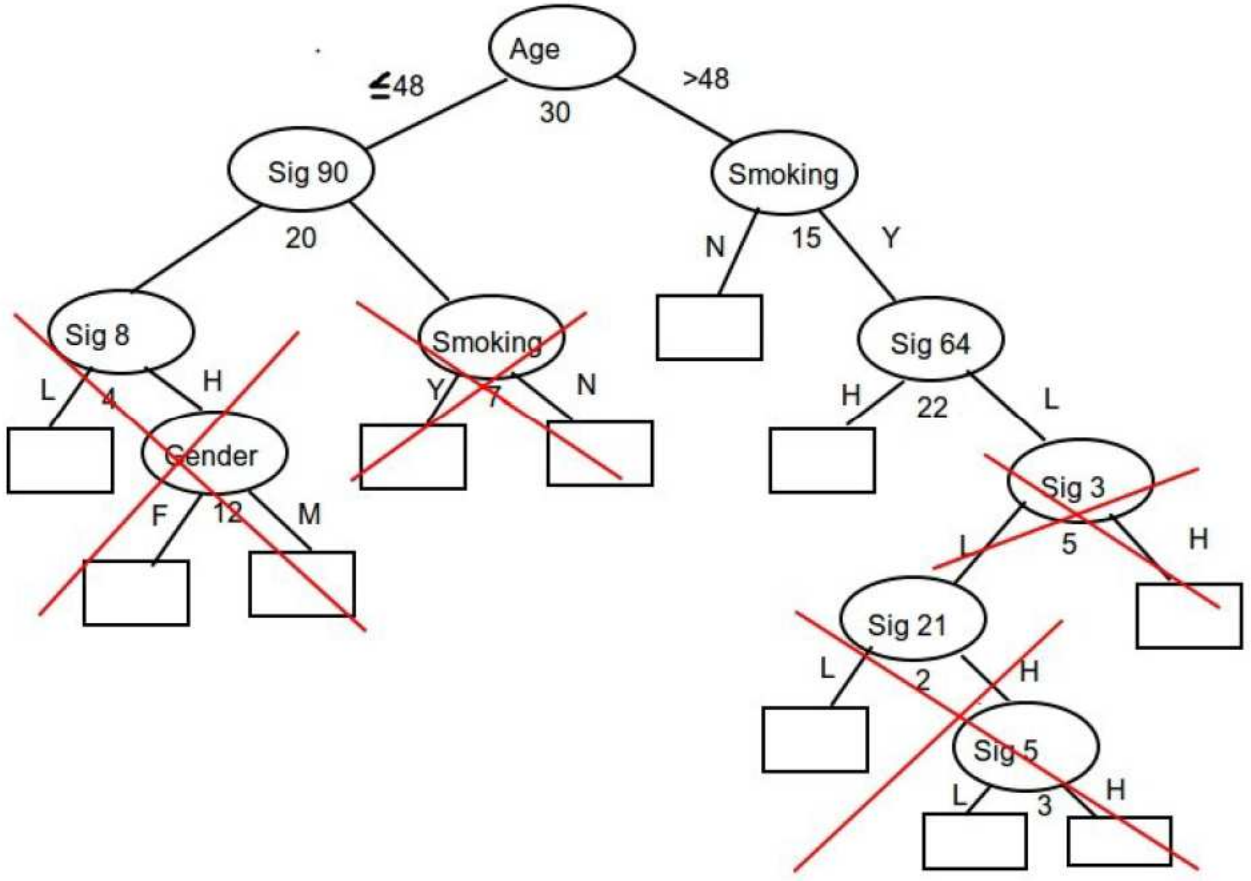
Pruning


Age
≤48
>48
30
Sig 90
Smoking
20
N
15
Y
Sig 8
Smoking
Sig 64
L
4
H
Y
7
N
H
22
L
Gender
Sig 3
F
12
M
L
5
H
Sig 21
L
2
H
Sig 5
L
3
H

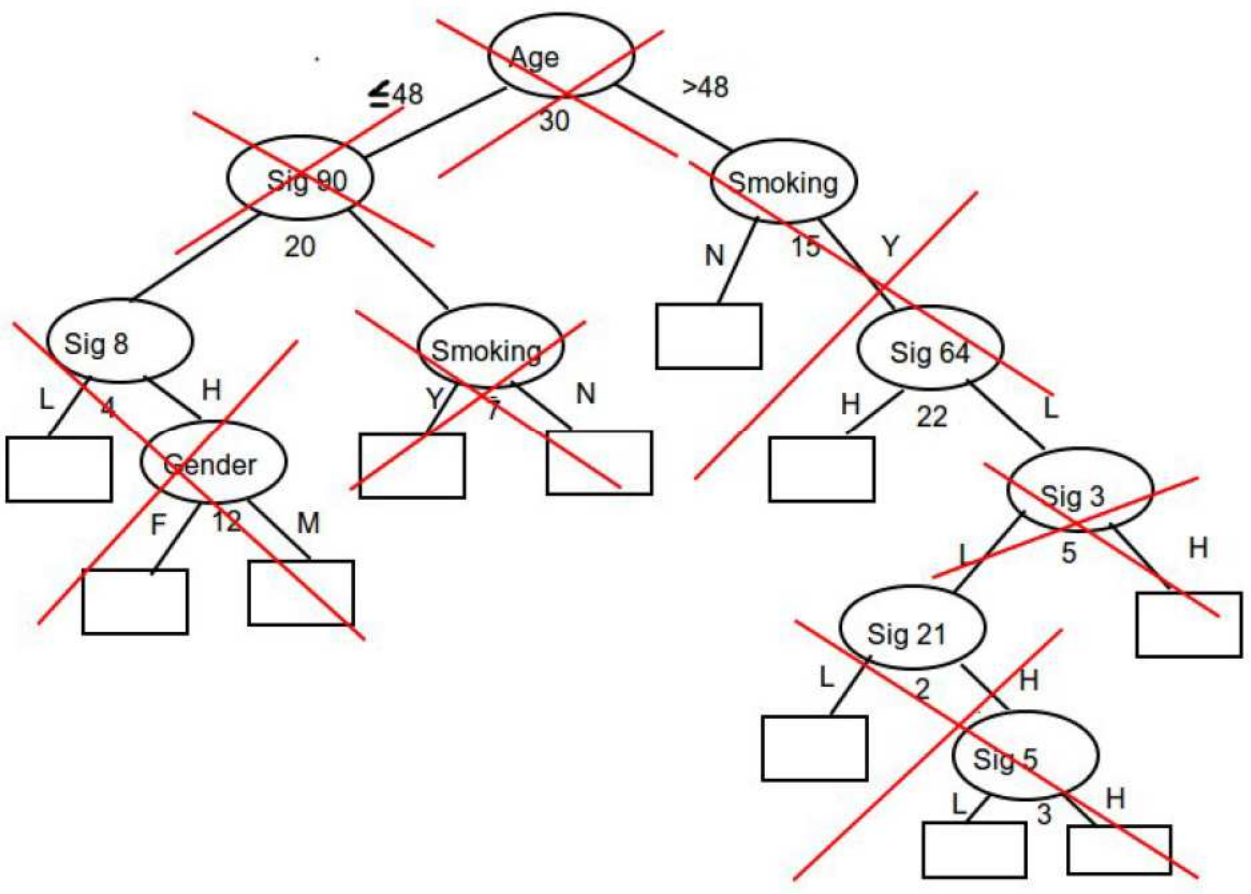
Pruning


Age
≤48
>48
30
Sig 90
Smoking
20
N
15
Y
Sig 8
Smoking
Sig 64
L
4
H
Y
7
N
H
22
L
Gender
Sig 3
F
12
M
L
5
H
Sig 21
L
2
H
Sig 5
L
3
H

Pruning


Age
≤48
>48
30
Sig 90
Smoking
20
N
15
Y
Sig 8
Smoking
Sig 64
L
4
H
Y
7
N
H
22
L
Gender
Sig 3
F
12
M
L
5
H
Sig 21
L
2
H
Sig 5
L
3
H

Pruning


Age
≤48
>48
30
Sig 90
Smoking
20
N
15
Y
Sig 8
Smoking
Sig 64
L
4
H
Y
7
N
H
22
L
Gender
Sig 3
F
12
M
L
5
H
Sig 21
L
2
H
Sig 5
L
3
H

Final Tree Selection with Validation Set

Since the splits used to make a tree are adaptively chosen as the maximum of several potentially correlated LRT statistics, the split complexity $G_\alpha(T) = G(T) - \alpha|S|$ is larger than would be expected if the splits were chosen a-priori.
We need to get an 'honest' estimate of $G_{\alpha_c}(T)$. We can do this using:

- A training and test (validation) set
- A resampling method

If we have a large sample we can get an 'honest' estimate of $G_{\alpha_c}(T)$ by using the following method.

- Split the data into a training set and test set
- Build a large tree with the training set
- Find the optimal subtrees and corresponding complexity parameters with the algorithm in the above section
- Finally force the test set down each of the subtrees

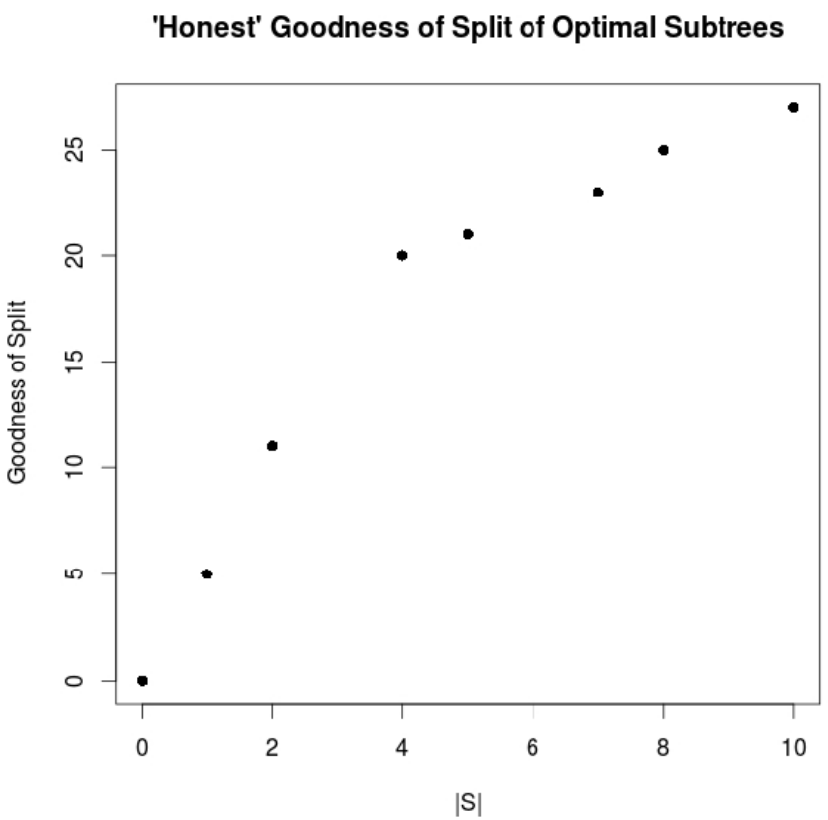
The final tree is the one that maximizes $G_{\alpha_c}(T)$ where $G_{\alpha_c}(T)$ is calculated using the test set. We recommend using $\alpha_c = 4$. This roughly corresponds to the 0.05 significance level of the split

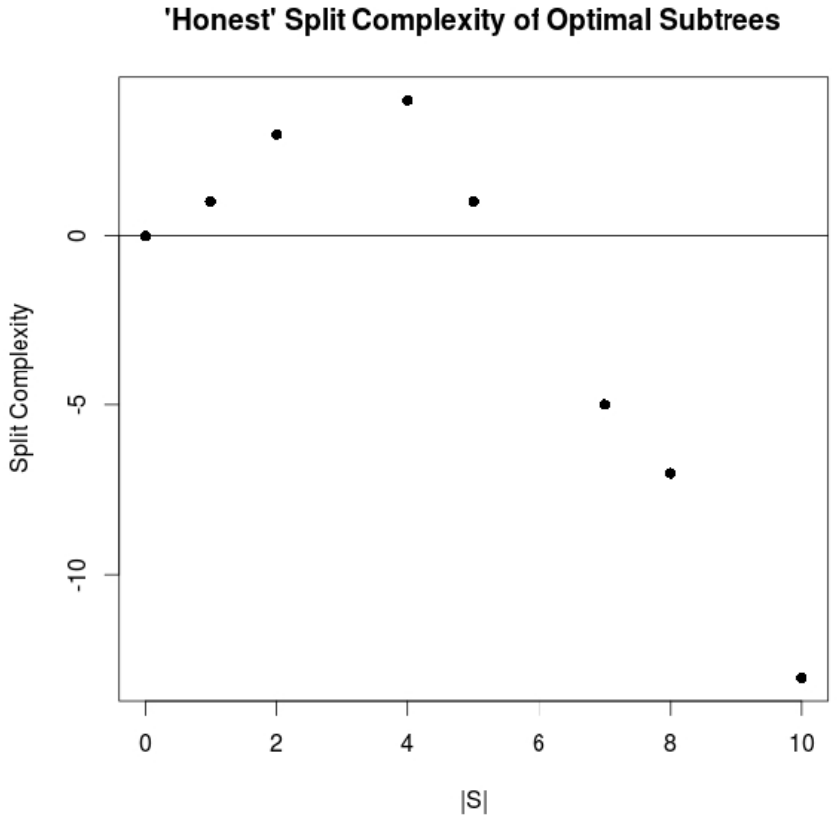
Final Tree Selection with Cross Validations

We propose choosing the final tree with a a 10×5-fold cross validation based method.

- Build a large tree and find the optimal subtrees and corresponding complexity parameters
- Partition the observations into 5 folds $\mathcal{L}_j, j = (1, ..., 5)$
- Build 5 trees $\mathcal{T}_{(-j)}$ on samples $\mathcal{L}_{(-j)}$
- For each α_k and $\mathcal{T}_{(-j)}$ find the optimal subtree $\mathcal{T}_{(-j),k}$
- Force $\mathcal{L}_j$ on $\mathcal{T}_{(-j),k}$ obtaining trees $\mathcal{T}_{j,k}$
- For each $\mathcal{T}_{j,k}$ calculate the goodness of split $G(\mathcal{T}_{j,k})$
- Take the mean over the folds to get $G(\mathcal{T}_{.,k})$
- Repeat this process 10 times with a new set of folds generated each time and take the average of the $G(\mathcal{T}_{.,k})$ to get $\bar{G}(\mathcal{T}_{.,k})$.
- Find $k^* = \max_k \bar{G}_{\alpha_c}(T_{.,k})$ and if $\bar{G}_{\alpha_c}(T_{.,k^*}) > 0$ the final tree is T_{k^*}, otherwise it is the null tree with no nodes.

Final Tree Selection Example





Evaluation of the Decision Tree: Simulations

- We simulate the following tree structures:
  - The 'null' tree with no predictive factor (no split)
  - A tree with a single true predictive factor (one split)
  - A tree with two true predictive factor (two splits)
- The number of potential genetic or clinical factors from 20 to 1000, and the risk factors are binary variables.
- Different effect sizes and sample sizes were simulated with four confounders created that are associated with survival outcome.

Evaluation of the Model: Type I Error

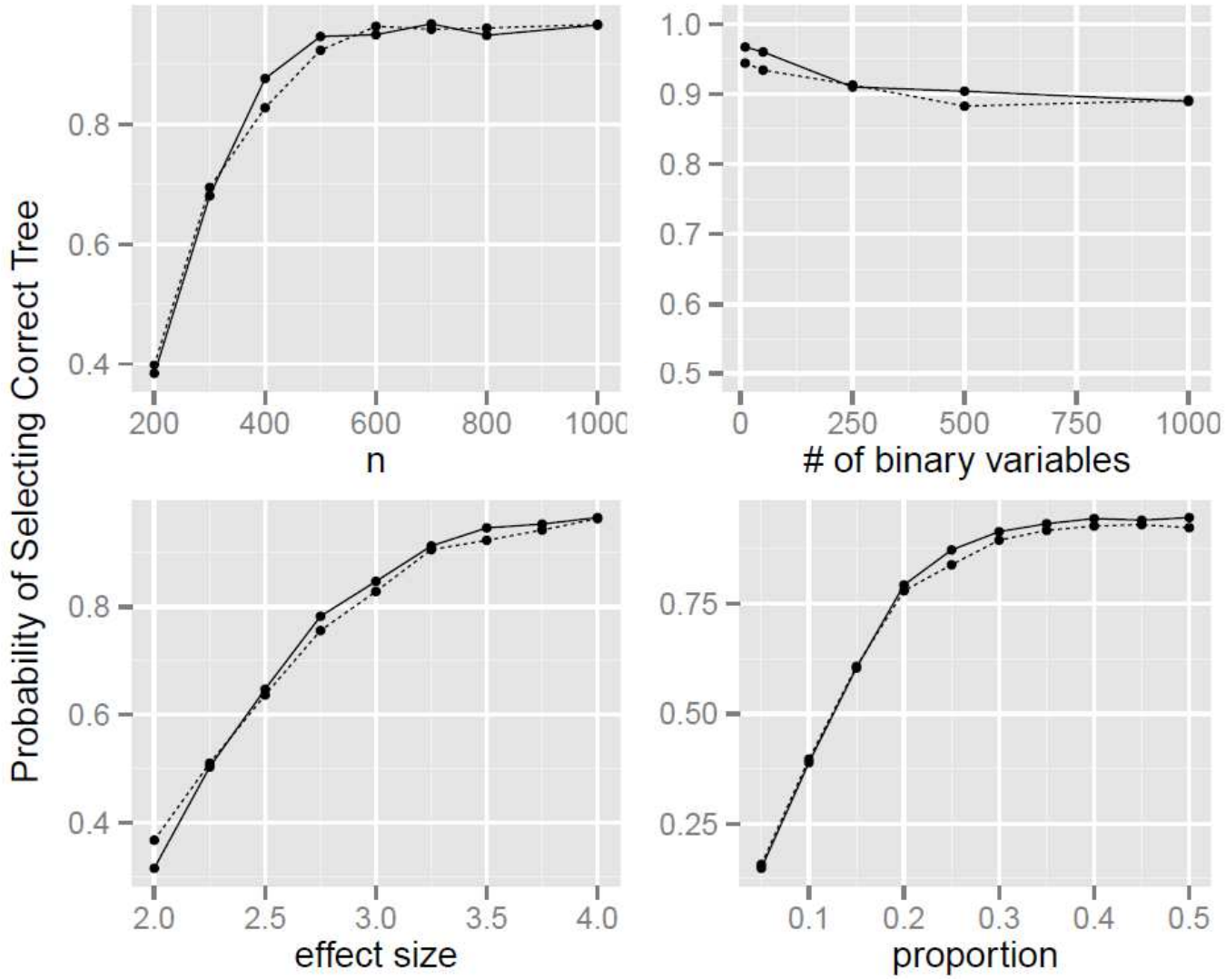
Table: Probability of selecting the wrong tree under the null hypothesis of no associated splits

$G(s,h)$	Sample Size (n)							
	1000	900	800	700	600	400	300	200
G_c	.049	.068	.054	.042	.047	.038	.032	.014
G_{ci}	.071	.070	.051	.069	.059	.047	.045	.022

$G(s,h)$	Number of Potential Splits					
	1000	500	250	100	50	10
G_c	.044	.054	.051	.048	.047	.035
G_{ci}	.039	.053	.058	.040	.058	.058

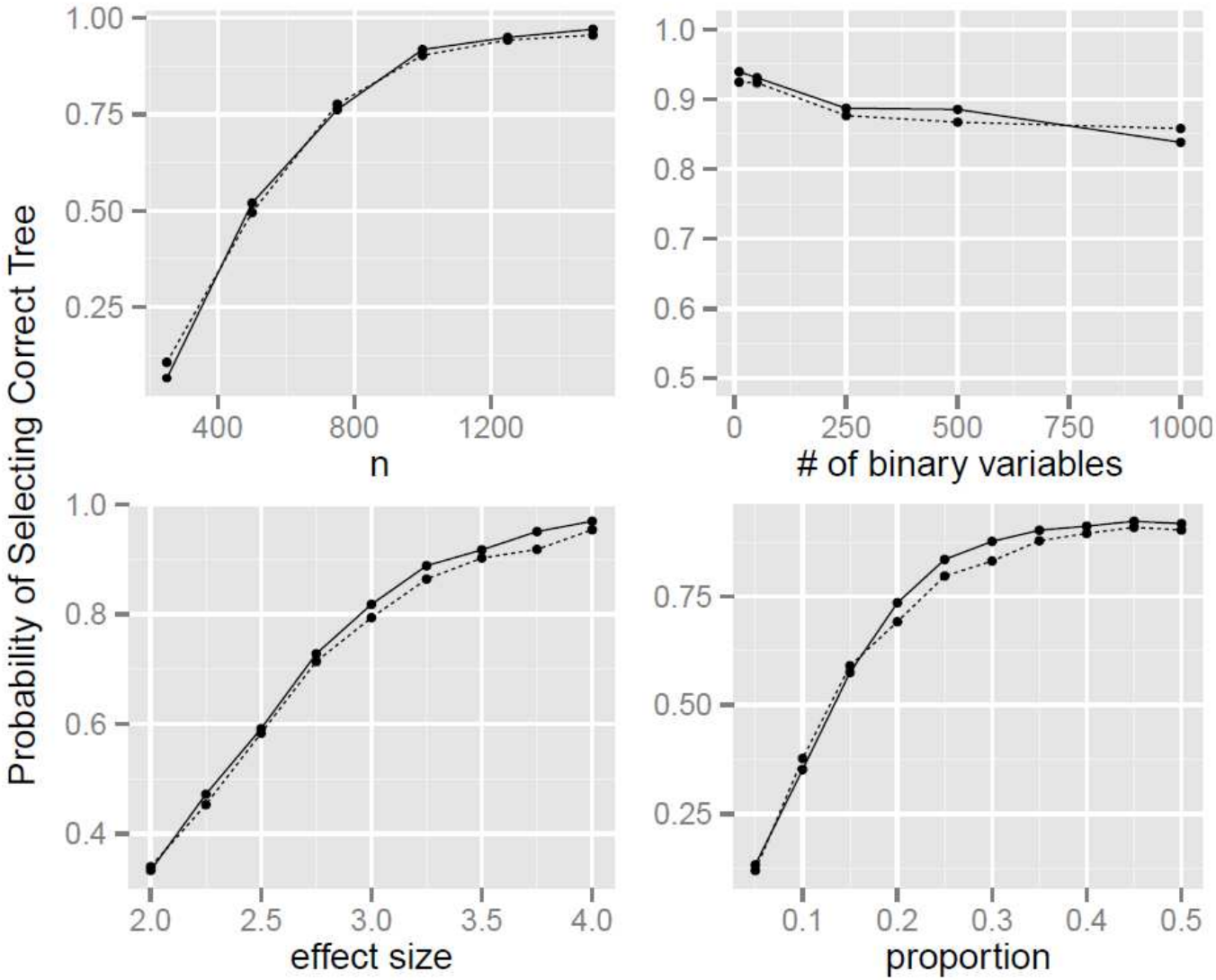
Evaluation of the Model: Statistical Power

Figure: Probability of selecting the correct tree with one split



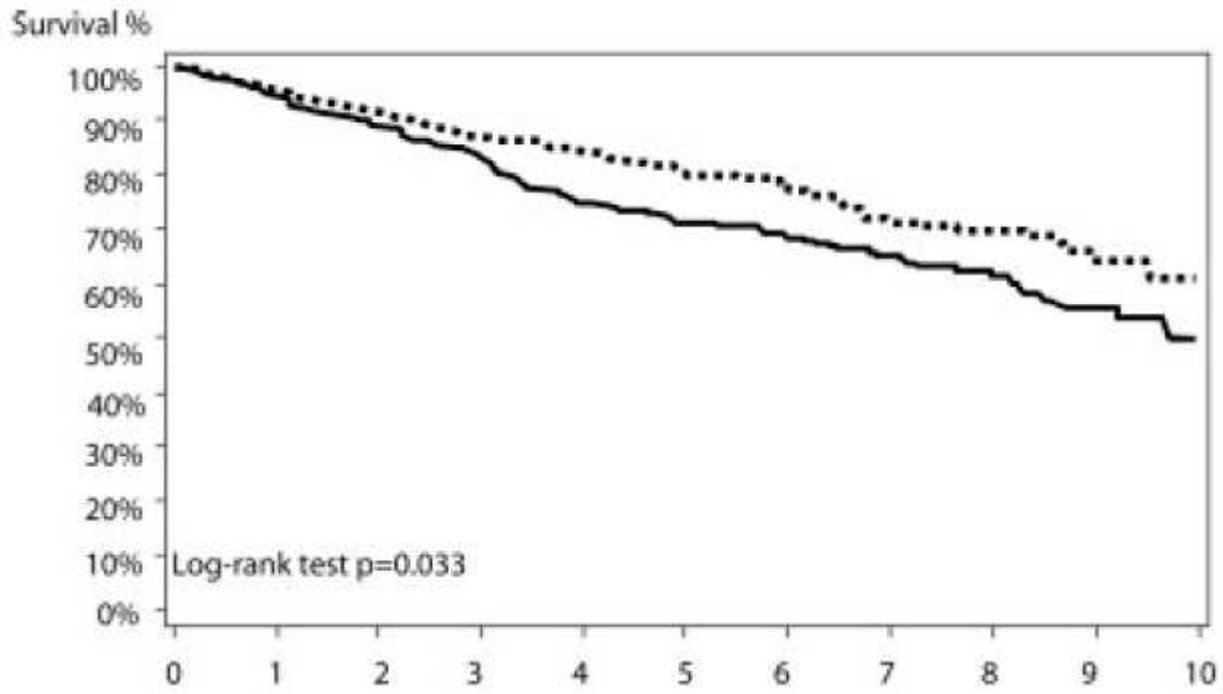
Evaluation of the Model: Statistical Power

Figure: Probability of selecting the correct tree with two splits



Application to a Clinical Trial Study

- A randomized Phase III α-tocopherol/β-carotene placebo-controlled trial with 540 early stage HNC patients (Quebec).



Int. J. Cancer: **119,** 2221–2224 (2006)

SHORT REPORT

Antioxidant vitamins supplementation and mortality: A randomized trial in head and neck cancer patients

Isabelle Bairati[1], François Meyer[1*], Edith Jobin[1], Michel Gélinas[2], André Fortin[3], Abdenour Nabid[4], François Brochet[5] and Bernard Têtu[1]

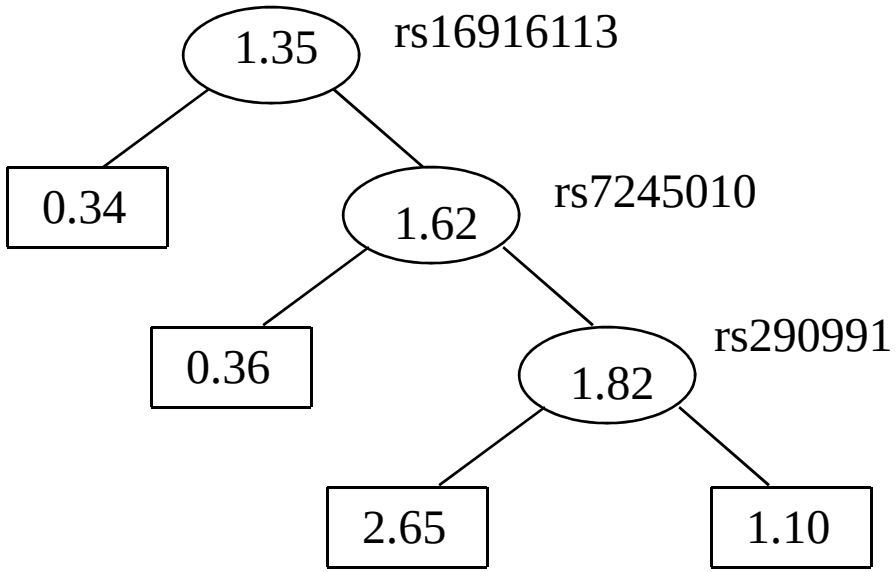
Application to a Clinical Trial Study

- GWAS data with 620,901 SNPs genotype information (Illumina 610K platform).
- After genetic quality control:

  515 patients (261 in the treatment arm and 254 in the placebo arm) with 543,873 SNPs.
- PFS is the primary outcome with top three genetic principal component as the confounders.
- Top 100 most prognostic significant SNPs were selected for predictive survival tree. Genetic dominant model was used for each SNP.
- 1000 validation data sets were used for pruning.

Application to a Clinical Trial Study

Final Decision Tree



- Subgroups are defined by SNP genotypes, wild type on the right, others on the left.
- Hazard ratios (HRs) are presented for each subgroup.

Conclusions

- The predictive tree model can be used to assess treatment interactive effect of multiple risk factors such as multiple genetic markers or signatures
- The method has well controlled type I error and is robust to the number of potential risk factors to be explored.
- The method can adjust for potential confounders
- The identified subgroups can help treatment decision on patients with specific characteristics.

Further Directions

- The methods can be extended to the studies where there are more than two treatment arms.
- The methods are based on parametric survival models. It can be extended to other clinical outcomes such as binary, ordinal, count.
- Further extension on computing risk outcomes such as cause specific survival.
- Further development on non-randomized retrospective clinical data for personalized medicine decision.

Acknowledgement

Princess Margaret Cancer Centre
University of Toronto

Ryan Del Bel
Colleen Kong
Osvaldo Espin-Garcia

Dr. Brian O'Sullivan
Dr. Sophie Huang
Dr. Geoffrey Liu
Dr. Fei-Fei Liu

Laval University

Dr. Isabelle Bairati
Dr. Francois Meyer

CIHR IRSC
Canadian Institutes of Health Research | Instituts de recherche en santé du Canada

CIHR STAGE
Strategic Training for Advanced Genetic Epidemiology