Regression Trees for Subgroup Identification

Wei-Yin Loh
Department of Statistics
University of Wisconsin, Madison
www.stat.wisc.edu/∼loh/

Joint work with Xu He, Chinese Academy of Sciences, Beijing,

and Michael Man, Eli Lilly, Indianapolis

The high cost of drugs

- It used to take $200 million and 7 years to bring a new drug to market
- Now the cost is $1.5–2 billion and the time-line can be as long as 15 years

Prices of some cancer drugs

Perjeta (*Genentech*). Breast cancer. $188K for one course. Can delay cancer growth for 6 months.

Yervoy (*Bristol-Myers Squibb*). Skin cancer. $120K for 4 injections. Extends survival by 4 months.

Provenge (*Dendreon*). Prostate cancer. $93K. Extends survival by 4 months.

Tarceva (*Astellas and Genentech*). Pancreatic cancer. $15K when combined with gemcitabine. Extends survival by 2 weeks.

Avastin (*Genentech*). Colorectal cancer. $10K per month. Extends survival by 5 months.

Ipilimumab (*Bristol-Myers Squibb*). Lung cancer and melanoma. $120K per course of treatment. Increases survival by 4 months compared to a different treatment.

Tailored therapeutics/Personalized medicine

- **Tailored therapeutic** is a treatment that is **shown to be more effective on average** in one subgroup of patients than in its complementary subgroup.
- "shown" = based on adequate and well controlled trials.
- "more effective on average" = still comparing average responses, just on smaller subgroups.

Examples of tailored therapeutics

Herceptin (trastuzumab). For HER-2 positive breast cancer.

Gleevec (imatinib). For chronic myeloid leukemia (CML) and gastrointestinal stromal tumor (GIST) stomach cancer.

Erbitux (cetuximab). For colorectal cancer and head and neck cancer.

Regression trees are natural for subgroup identification

— subgroups are defined by terminal nodes of a tree

Two key steps in tree construction

1. How to split each node?
2. When to stop splitting?

Previous methods for censored responses

Let $Z = 0, 1$ be the treatment variable and let node t be split into t_L and t_R

RECPAM (Negassa et al., 2005) Choose split to maximize Cox partial likelihood ratio for testing H_0 vs. H_1:

$$H_0 : \quad \lambda(u, \mathbf{x}) = \lambda_{0,t}(u) \exp\{\beta_0 z I(\mathbf{x} \in t)\}$$
$$H_1 : \quad \lambda(u, \mathbf{x}) = \lambda_{0,t}(u) \exp\{\beta_1 z I(\mathbf{x} \in t_L) + \beta_2 z I(\mathbf{x} \in t_R)\}$$

IT: Interaction trees (Su et al., 2008, 2009) Choose split to minimize p-value for testing $H_0 : \beta_3 = 0$ in the model

$$\lambda(u, \mathbf{x}) = \lambda_{0,t}(u) \exp\{\beta_1 z + \beta_2 I(\mathbf{x} \in t_L) + \beta_3 z I(\mathbf{x} \in t_L)\}$$

Weaknesses:

1. Compute intensive: one or more Cox models fitted for *each candidate split*
2. Biased toward selecting variables that allow more splits
3. Baseline hazard function $\lambda_{0,t}(u)$ depends on t and hence on $\mathbf{x}$

Previous methods for binary responses

VT: Virtual twins (Foster et al., 2011) Assume $Y, Z = 0, 1$.

1. Random forest (RF) to estimate $\tau = P(Y = 1|Z = 1) - P(Y = 1|Z = 0)$ with $Z, X_1, \ldots, X_M, ZX_1, \ldots, ZX_M, (1 - Z)X_1, \ldots, (1 - Z)X_M$.
2. RPART to predict τ. Subgroups are terminal nodes with large τ.

Weaknesses:

1. Selection biases of CART and random forest.
2. No good way to deal with missing values (RF needs prior imputation).
3. Not extensible to three or more treatments and to censored responses.
4. Random: subgroups depend on choice of random seed in RF.

SIDES: (Lipkovich et al., 2011)

1. Find 5 splits to minimize p-value (e.g., differential treatment effects or difference in efficacy and safety between child nodes).
2. For each split, repeat on most promising child node.

Performance largely unknown; not extensible to three or more treatments.

QUINT: Qualitative interaction tree (Dusseldorp and Van Mechelen, 2013)

Split each node to optimize a weighted sum of measures of effect size and subgroup size.

Strength: Allows simultaneous control of effect size and subgroup size

Weaknesses:

1. Selection bias.
2. Needs one treatment to be better in one subgroup and worse in other.
3. Not easily extensible to three or more treatments.
4. Not easily extensible to censored responses.

Key idea #1:
use piecewise-*linear* models

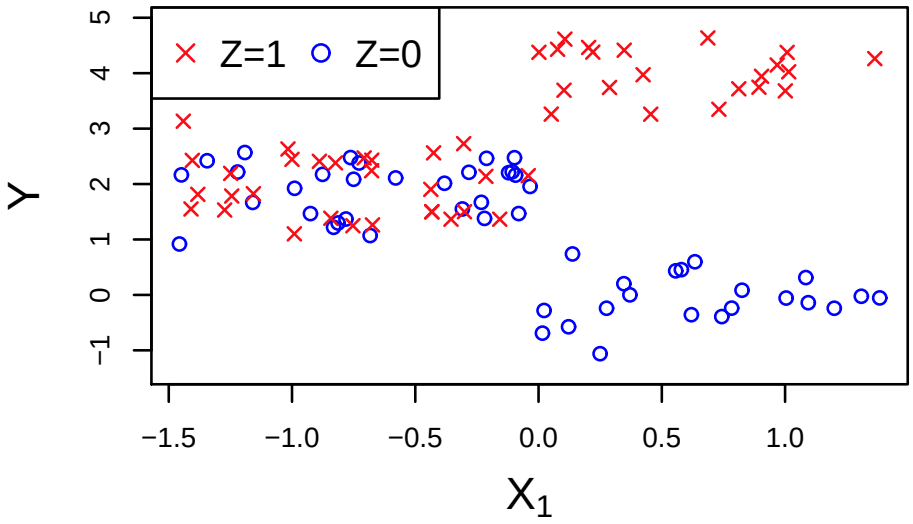
- Suppose Z takes values 0, 1, . . .
- Fit the model $EY = \eta + \sum_k \beta_k I(Z = k)$ in each node (so that treatment effects can be estimated)
- CART, RPART, and other piecewise-constant trees inapplicable

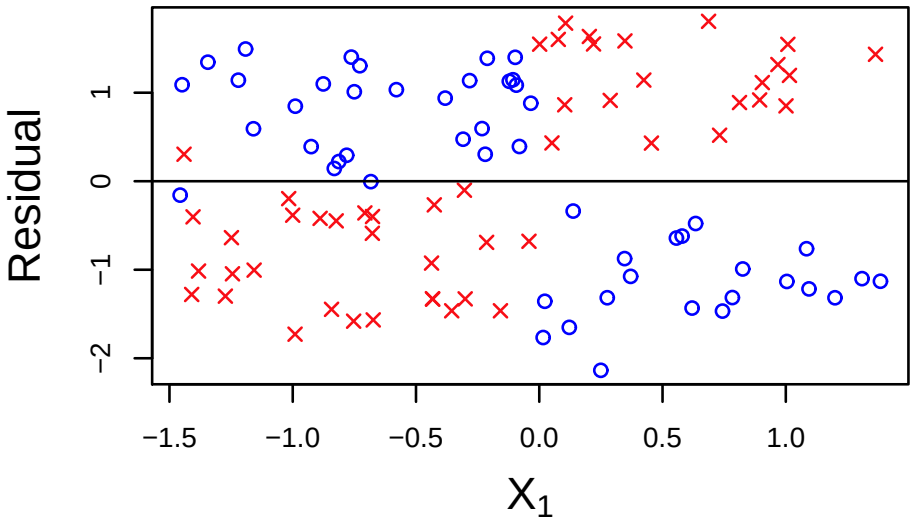
GUIDE (Loh, 2002, 2009) and MOB (Zeileis et al., 2008)

- These algorithms use significance tests to select variables for splitting data
- GUIDE uses chi-squared tests of residual signs vs. each predictor variable
  - missing values are *included*
- CTREE and MOB use permutation tests on score functions
  - missing values are *excluded* (implies missing completely at random)

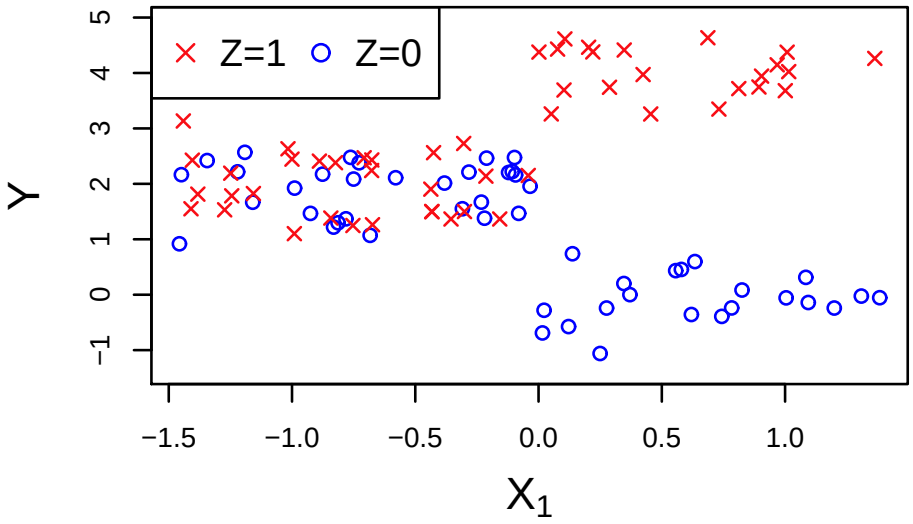
Example with treatment Z = 0, 1

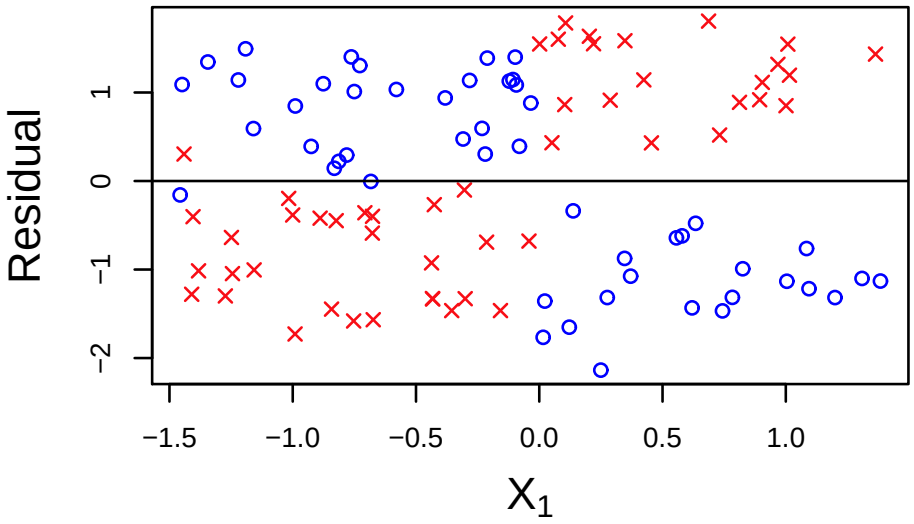
- True model:
  $Y = 1.9 + 0.2I(Z = 1) - 1.8I(X_1 > 0) + 3.6I(X_1 > 0, Z = 1) + \varepsilon$
- $X_2, X_3, \ldots$ are noise
- Fit $EY = \beta_0 + \beta_1 Z$ to data in each node

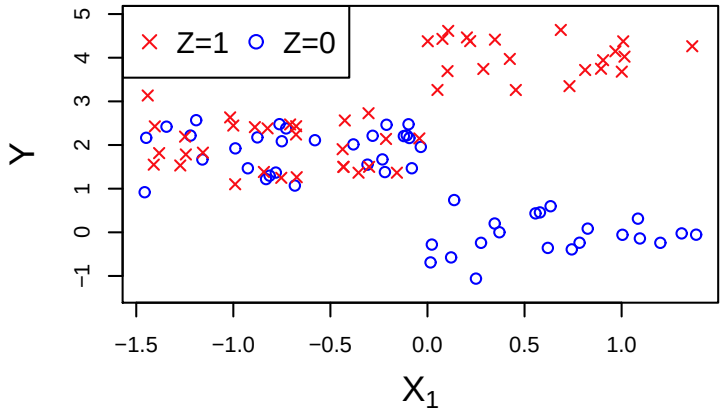


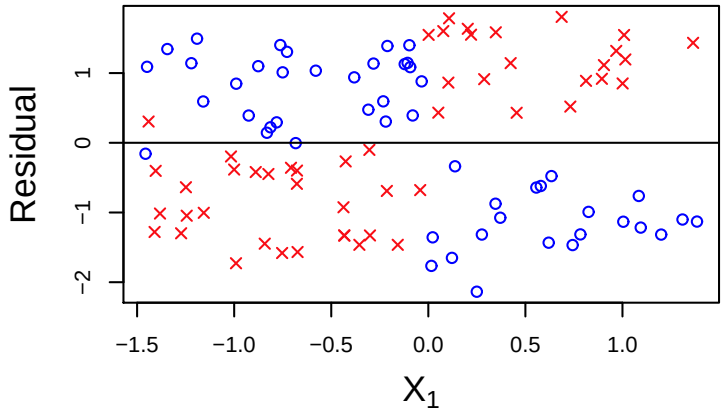


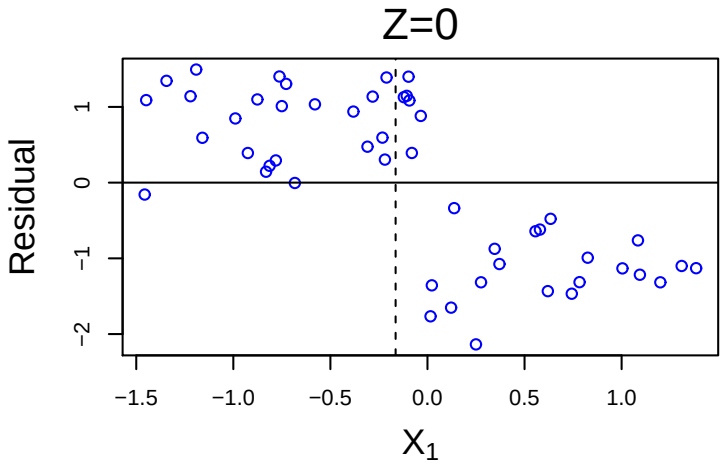
Key idea #2:
examine residual patterns
for each treatment level

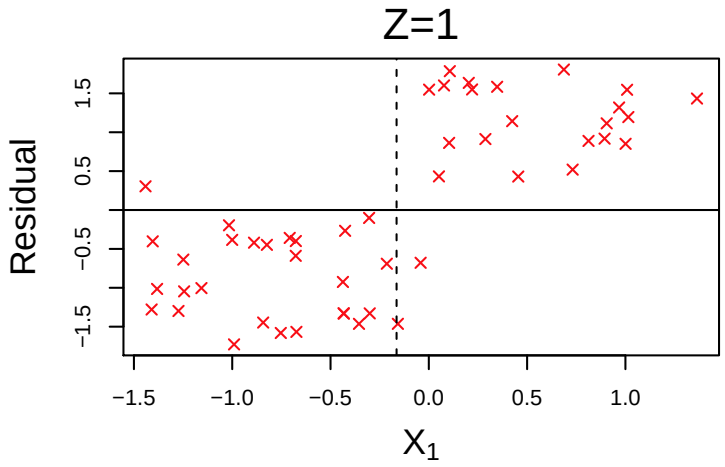












$Z=0$	$X_1 \le \bar{x}_1$	$X_1 > \bar{x}_1$
resid > 0	21	6
resid ≤ 0	2	21

$\chi^2 = 21.2,\ p = 4 \times 10^{-6}$

$Z=1$	$X_1 \le \bar{x}_1$	$X_1 > \bar{x}_1$
resid > 0	1	21
resid ≤ 0	26	2

$\chi^2 = 35.2,\ p = 3 \times 10^{-9}$

Key idea #3:
why group ordinal variables?

- Grouping values of ordinal X variables may result in power loss
- But grouping allows missing values to be used!

Gs method ("s" for "sum")

1. Obtain the residuals from the model $EY = \eta + \sum_k \beta_k I(Z = k)$
2. Do for each X variable:
   (a) Do for each value of Z:
      i. Crosstab residual signs vs. grouped values of X
      ii. Add one more group for missing values in X if there are any
      iii. Compute chi-squared statistic of the table
      iv. Convert chi-squared value to one with a single df
   (b) Sum converted chi-squareds over values of Z to get test statistic
3. Let X^* have largest test statistic
4. Find split $X^* \in S$ that minimizes sum of squared residuals in subnodes
5. Partition data and recursively apply procedure to each subnode

Predictive vs. prognostic variables

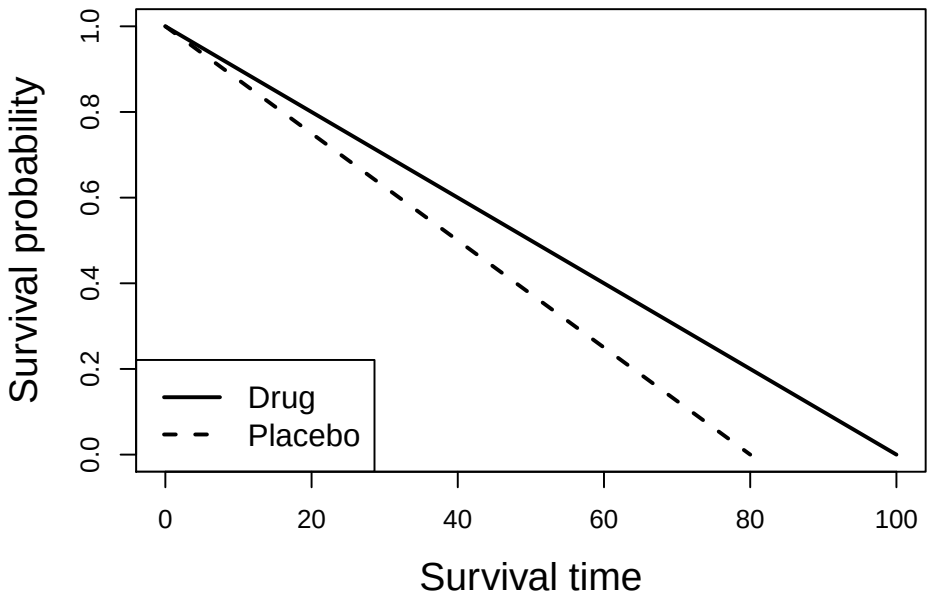
1. A **prognostic** variable is a clinical or biologic characteristic that is objectively measurable and that provides information on the likely outcome of the disease in an untreated individual.
   Examples are patient age, family history, disease stage, and prior therapy.

2. A **predictive** variable is a clinical or biologic characteristic that provides information on the likely benefit from treatment. Such variables can be used to identify subgroups of patients who are most likely to benefit from a given therapy.

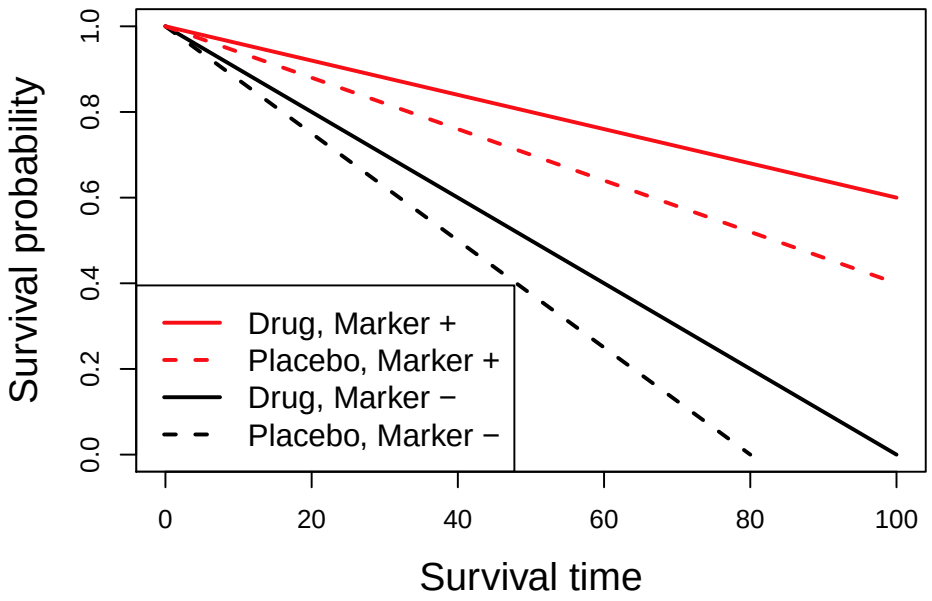
3. **Prognostic variables** define the effects of patient or tumor characteristics on the patient outcome, whereas **predictive variables** define the effect of treatment on the tumor.

— Italiano (2011)

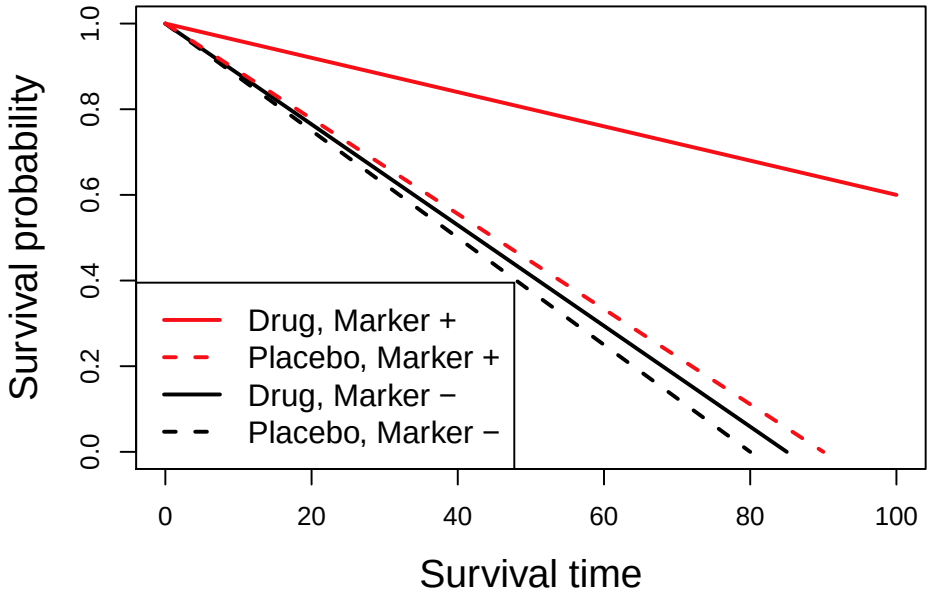
Whole population



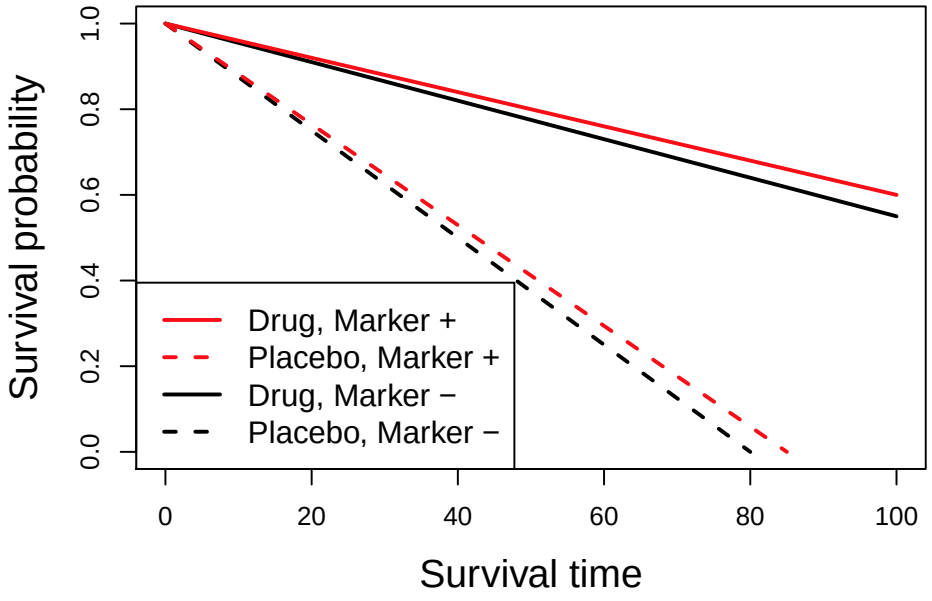
Prognostic biomarker



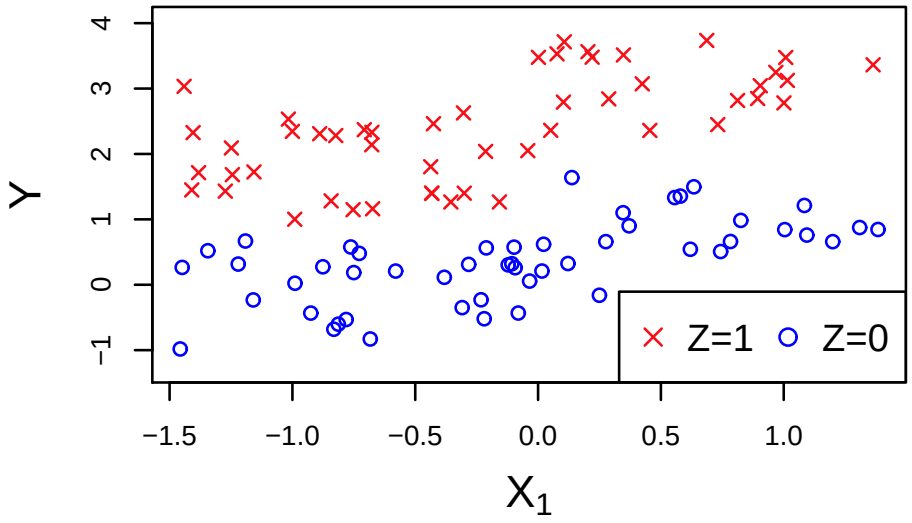
Predictive biomarker

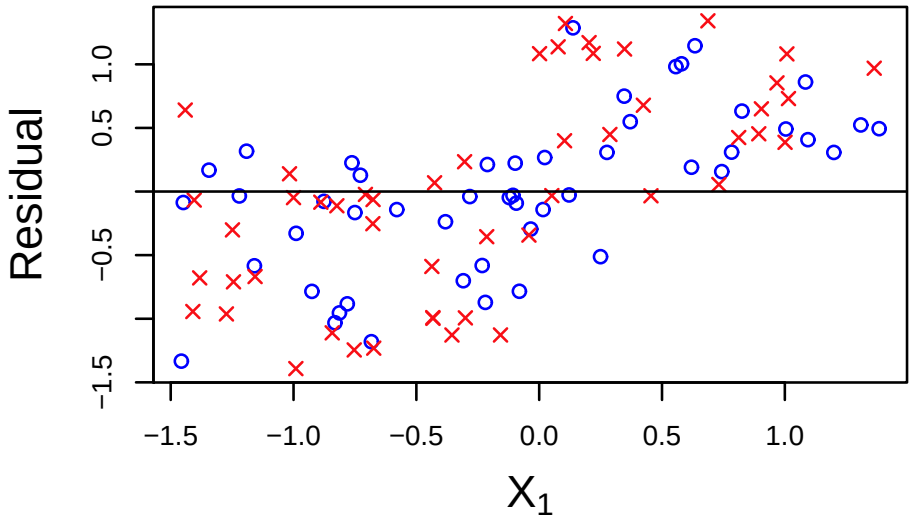


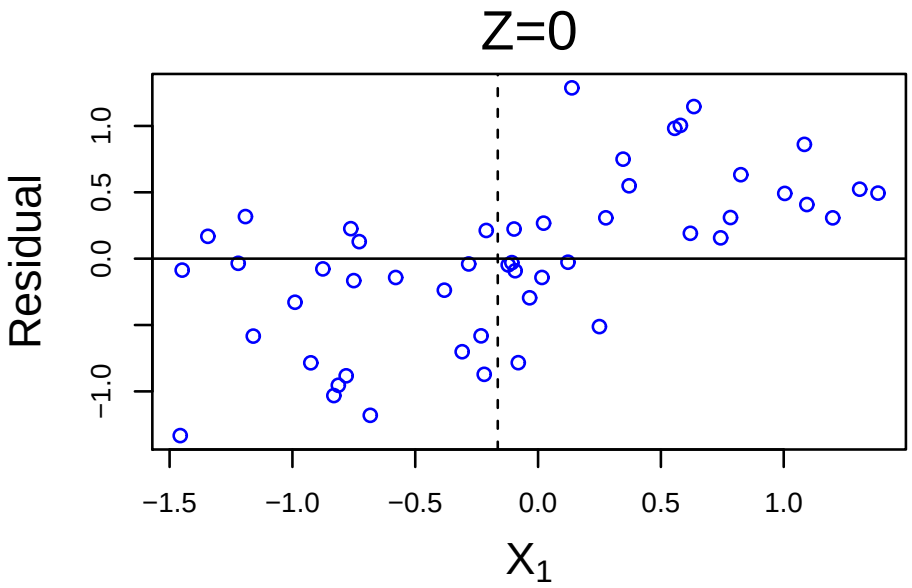
No biomarker effect

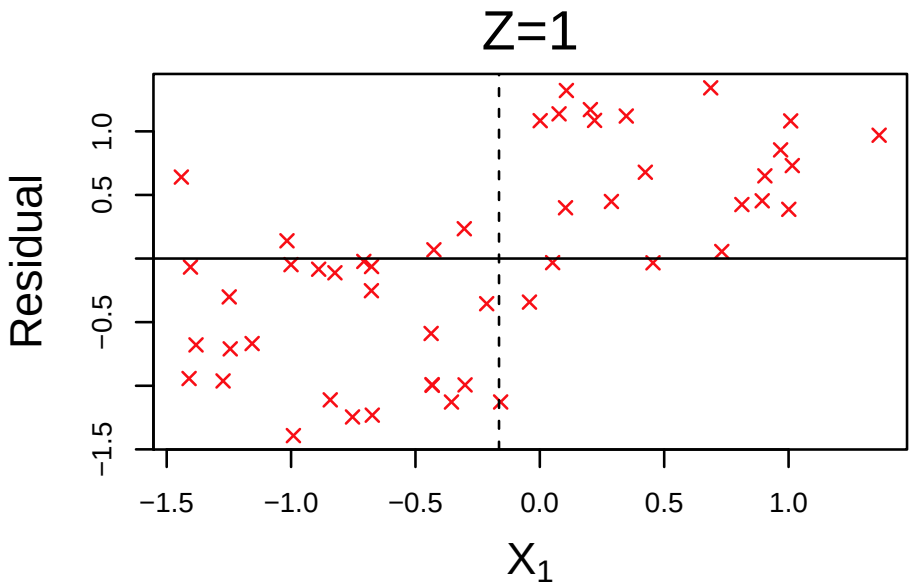


Problem: Gs is sensitive to prognostic variables









Key idea #4:
test for treatment interactions

1. Usual approach: add cross-product "interaction" terms if X is ordinal:

$$EY = \eta + \sum_k \beta_k I(Z = k) + \sum_k \gamma_k X I(Z = k)$$

2. Two problems with this:
   (a) Cross-products $XI(Z = k)$ do not represent every kind of interaction
   (b) Cross-products do not allow missing values in X

3. Solution: Use interaction model for categorical variables

$$EY = \eta + \sum_j \alpha_j I(X = j) + \sum_k \beta_k I(Z = k) + \sum_j \sum_k \gamma_{jk} I(X = j, Z = k)$$

with a category for missing values. If X is ordinal, group its values.

Solution: Gi method ("i" for "interaction")

Test for lack of fit of *model without interactions*:

1. Do for each X at each node:
   (a) If X is ordinal, categorize it into two groups at its mean
   (b) If X is categorical, let its values form the groups
   (c) Add a group for missing values
   (d) Let H be the factor variable created from the groups
   (e) Test lack of fit of the model $EY = \beta_0 + \sum_j \alpha_j I(H = j) + \sum_k \beta_k I(Z = k)$
2. Let X^* be the variable with the most significant chi-squared
3. Find the split on X^* that minimizes the sum of squared residuals of the model $EY = \eta + \sum_k \beta_k I(Z = k)$ fitted to each of the two subnodes

Comparison with other methods

IT. Interaction trees (Su et al., 2008)

QU. Qualitative interaction trees (Dusseldorp and Van Mechelen, 2013)

SI. SIDES (Lipkovich et al., 2011)

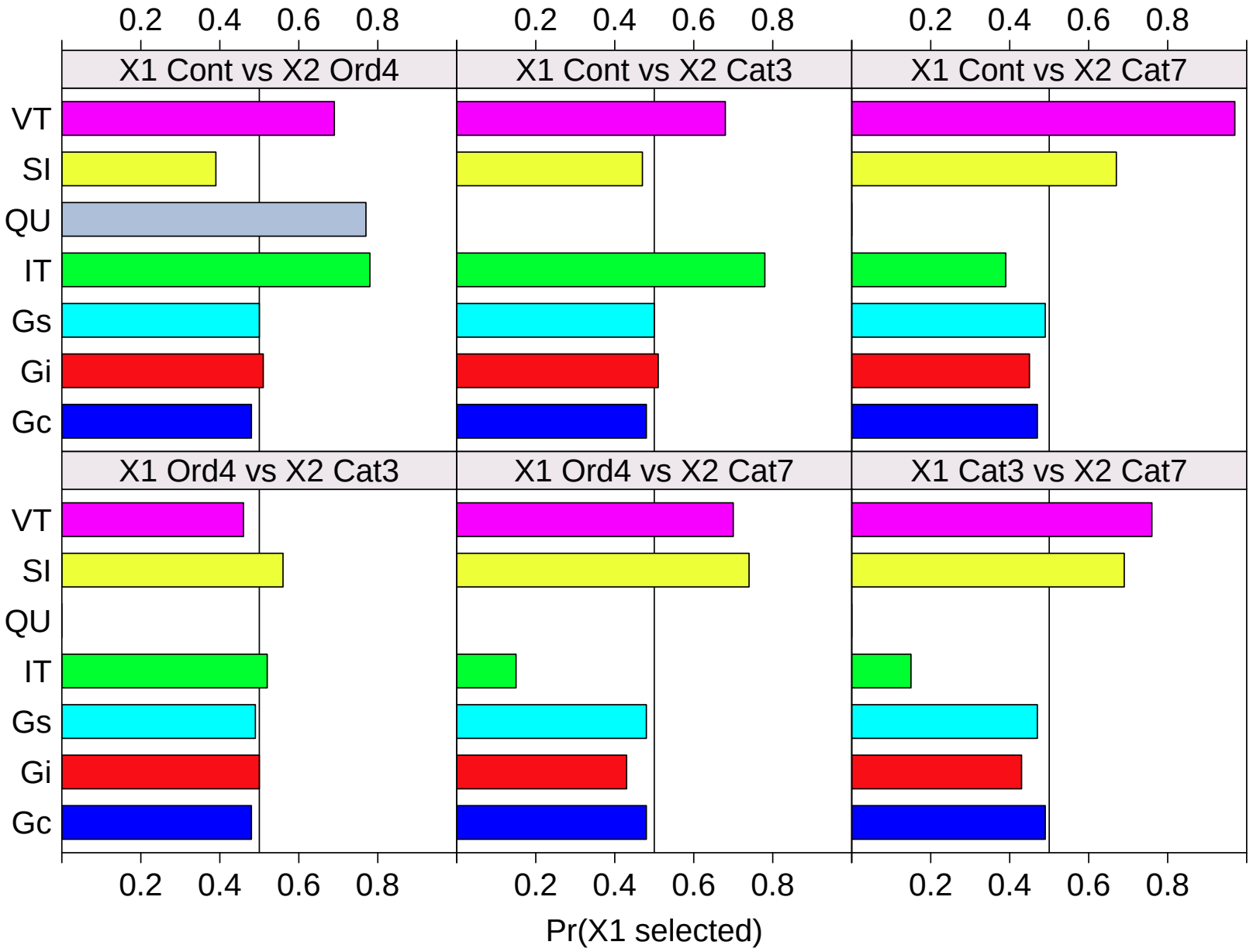
VT. Virtual twins (Foster et al., 2011)

Selection bias

- Bernoulli Y and Z with 0.50 success probabilities
- Two predictor variables X_1 and X_2 with distributions given below
- All variables are mutually independent
- Simulate 2500 data sets of 100 observations each
- Find frequency that X_1 is selected (0.50 if unbiased)

Notation	Type	Distributions of X_1 and X_2
Cont	Continuous	Standard normal
Ord4	Ordinal	Discrete uniform with 4 levels
Cat3	Categorical	Discrete uniform with 3 levels
Cat7	Categorical	Discrete uniform with 7 levels

Prob(X_1 is selected to split node)



Measuring accuracy of subgroups

- For any subgroup S, define its effect size as

$$R(S) = |P(Y = 1|Z = 1, S) - P(Y = 1|Z = 0, S)|$$

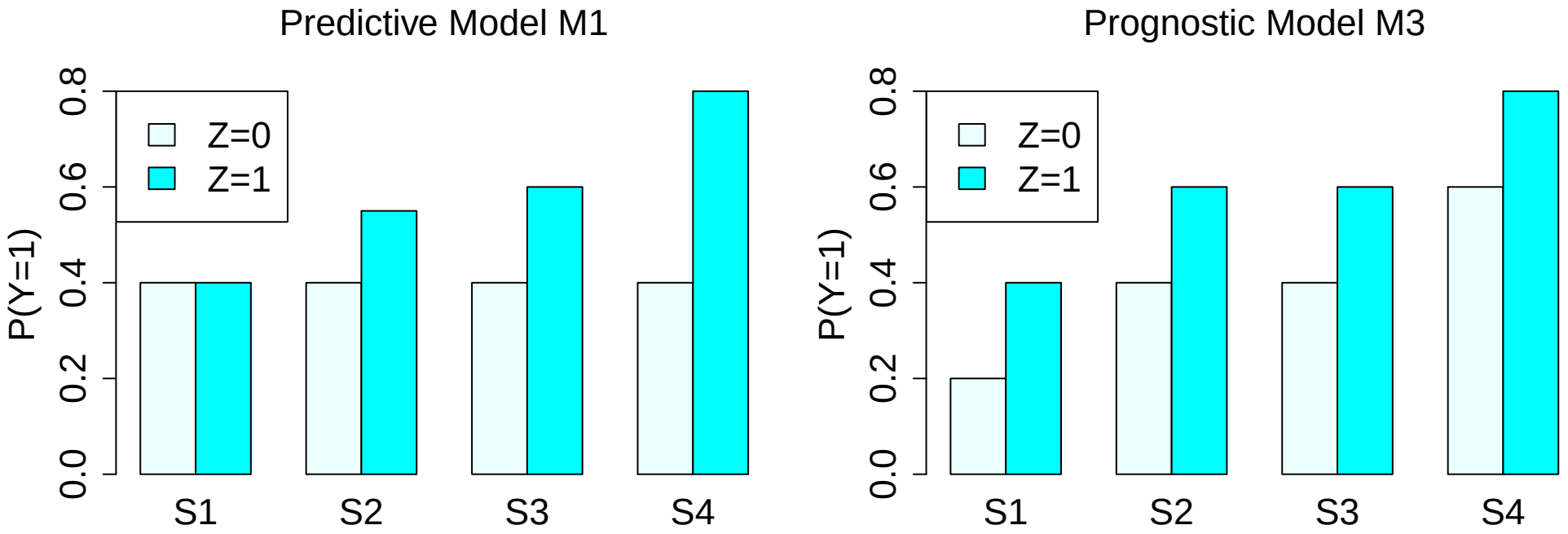
- "Correct" subgroup S^* is maximal (in probability) S with largest $R(S)$
- Let $n(t, y, z)$ be number of samples in node t with $Y = y$ and $Z = z$
- Define $n(t, +, z) = \sum_y n(t, y, z)$ and $n_t = \sum_y \sum_z n(t, y, z)$
- For any terminal node t, let S_t be the subgroup defined by t
- Estimate $R(S_t)$ with

$$\hat{R}(S_t) = \left| \frac{n(t, 1, 1)}{n(t, +, 1)} - \frac{n(t, 1, 0)}{n(t, +, 0)} \right|$$

- Selected subgroup is $\hat{S}$ that maximizes $\hat{R}(S_t)$; take union if not unique
- Accuracy of $\hat{S}$ is $P(\hat{S})/P(S^*)$ if $\hat{S} \subset S^*$, 0 otherwise.

Simulation Models M1 and M3

$X_i = 0, 1, 2\ (i = 1, 2, \ldots, 100)$ are genetic markers

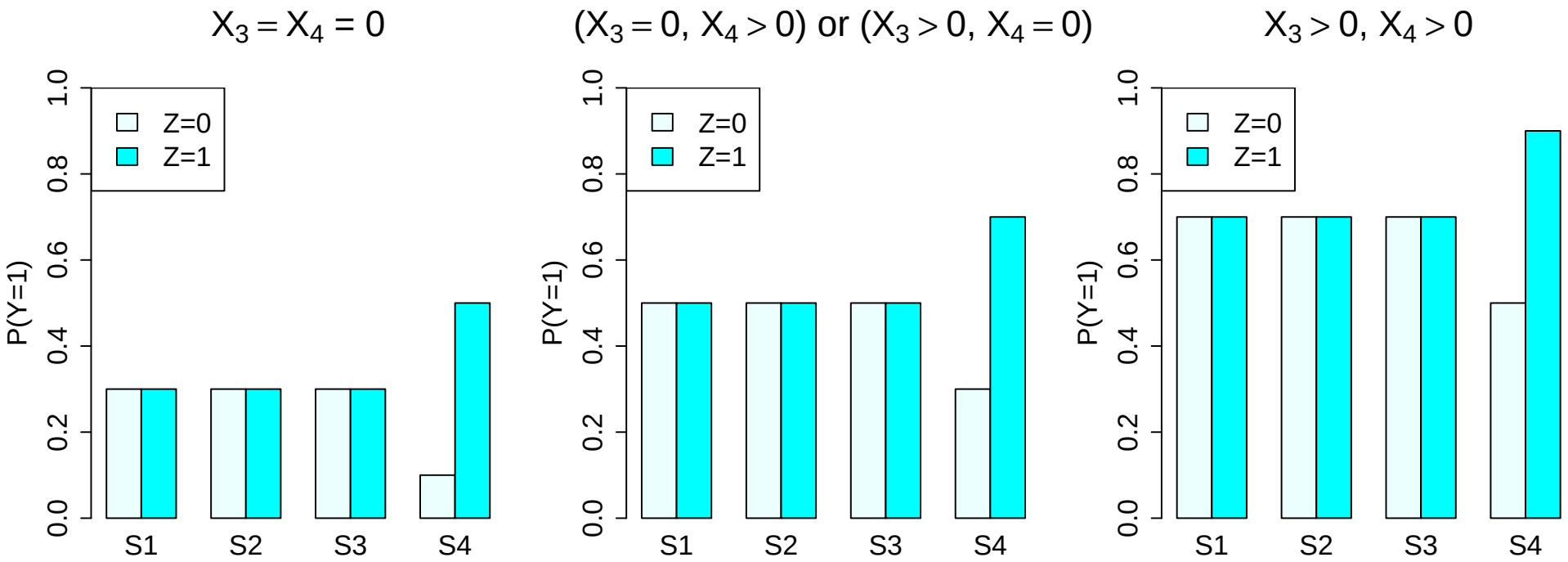


S1 = $\{X_1 = 0, X_2 = 0\}$, S2 = $\{X_1 = 0, X_2 > 0\}$,
S3 = $\{X_1 > 0, X_2 = 0\}$, S4 = $\{X_1 > 0, X_2 > 0\}$

Correct subgroup is S4 for Model M1 and entire space for M3

Model M2: $X_i = 0, 1, 2\ (i = 1, 2, \ldots, 100)$

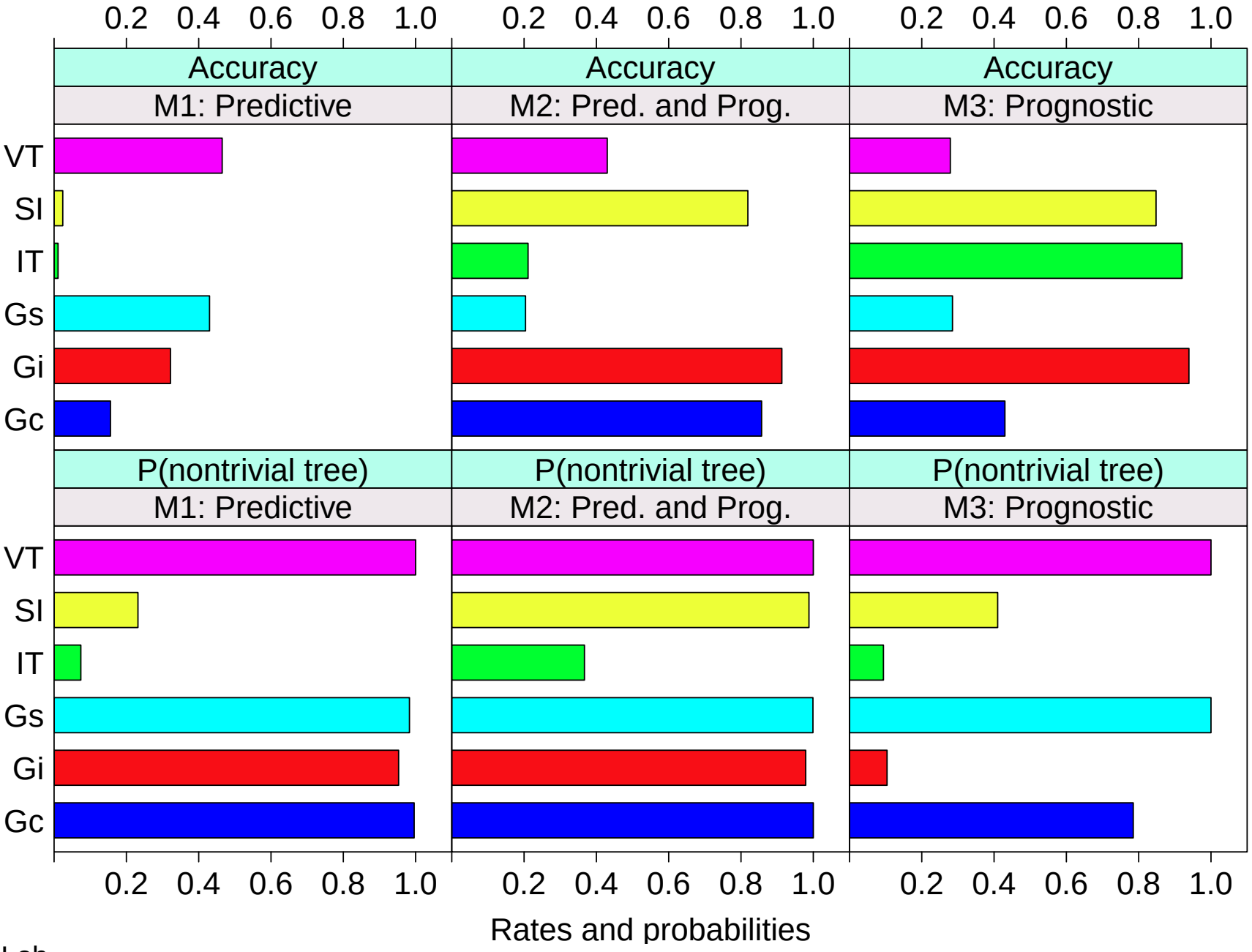
X_1 and X_2 predictive, X_3 and X_4 prognostic



S1 = $\{X_1 = 0, X_2 = 0\}$, S2 = $\{X_1 = 0, X_2 > 0\}$,
S3 = $\{X_1 > 0, X_2 = 0\}$, S4 = $\{X_1 > 0, X_2 > 0\}$

Correct subgroup is S4

Accuracy and probability of nontrivial trees



How to extend Gi and Gs to censored data?

1. Difficulties:
   (a) Gs employs residuals
   (b) Gi requires lack-of-fit test for additive model
2. Censored responses often fitted with proportional hazards model
3. What proportional hazards residuals to use?
4. How to test proportional hazards model for lack of fit?

Key idea #5

1. Use Poisson regression to fit a proportional hazards model to each node
2. Use Poisson residuals for Gs
3. Use chi-squared lack-of-fit test of additive Poisson model for Gi

Proportional hazards trees

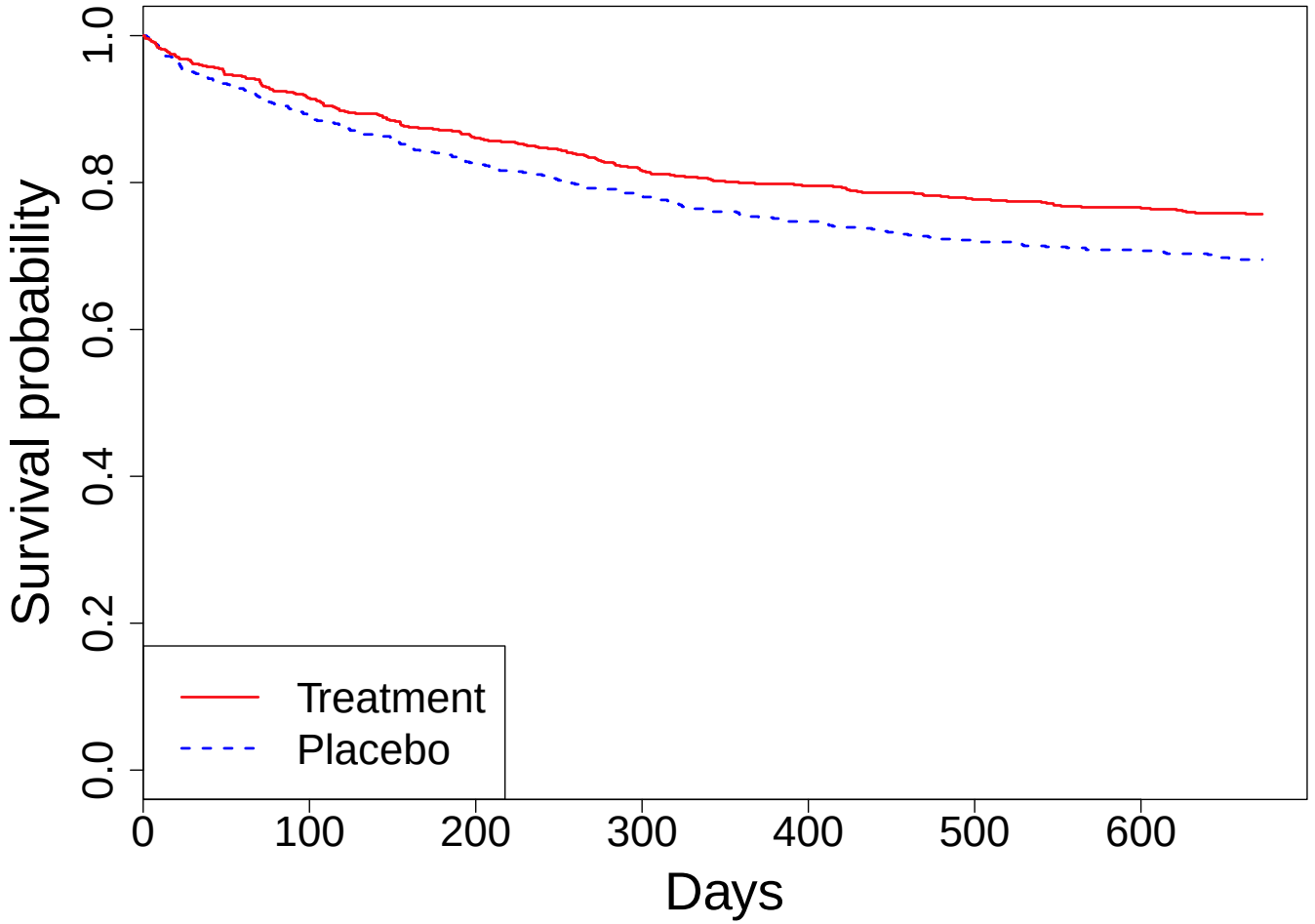
1. Let T and C be the true survival time and censoring time, respectively
2. Let $Y = \min(T, C)$ and $\delta = I(T < C)$ be the event indicator
3. Let $\Lambda_0(.)$ be the baseline cumulative hazard function
4. Estimate coefficients of proportional hazards model by iteratively fitting a Poisson regression model with δ_i as response and $\log \Lambda_0(y_i)$ as offset:
   (a) Use the Nelson-Aalen method to get an initial estimate of $\Lambda_0(y_i)$
   (b) Apply Gs or Gi method to construct a tree
   (c) Re-estimate $\Lambda_0(y_i)$ from the tree
   (d) Repeat steps (b) and (c) four more times

Proportional hazards and Poisson likelihoods

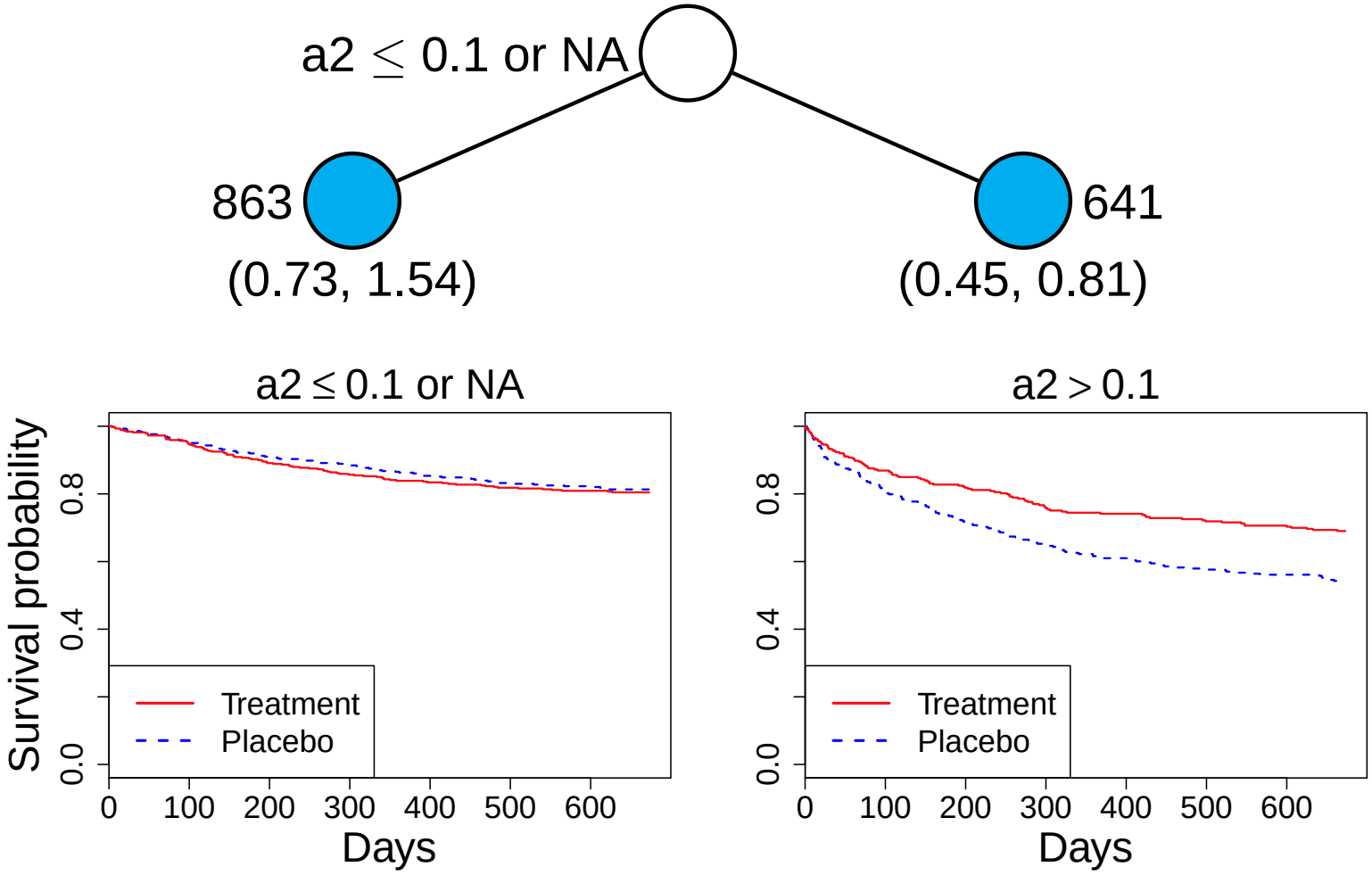
- Let u_i and $\mathbf{x}_i$ denote the survival time and covariate vector of subject i
- Let s_i be independent censoring time, $\delta_i = I(u_i < s_i)$ and $y_i = \min(u_i, s_i)$
- Let $F(u, \mathbf{x})$ and $\lambda(u, \mathbf{x})$ denote the distribution and hazard functions
- Suppose that $\lambda(u, \mathbf{x}) = \lambda_0(u) \exp(\beta' \mathbf{x})$
- Let $\Lambda(u, \mathbf{x}) = \int_{-\infty}^{u} \lambda(z, \mathbf{x})\, dz$, $\Lambda_0(u) = \Lambda(u, \mathbf{0})$, and $f(u, \mathbf{x}) = F'(u, \mathbf{x})$
- Then $f(u, \mathbf{x}) = \lambda_0(u) \exp\{\beta' \mathbf{x} - \Lambda_0(u) \exp(\beta' \mathbf{x})\}$. Let $\mu_i = \Lambda_0(y_i) \exp(\beta' \mathbf{x}_i)$
- Loglikelihood is $\sum_{i=1}^{n} (\delta_i \log \mu_i - \mu_i) + \sum_{i=1}^{n} \delta_i \log\{\lambda_0(y_i)/\Lambda_0(y_i)\}$
- 1st term is kernel of loglikelihood for Poisson variables δ_i with means μ_i; 2nd term is independent of $\mathbf{x}_i$ (Aitkin and Clayton, 1980)
- If $\Lambda_0(y_i)$ known, estimate β using δ_i as Poisson with means $\Lambda_0(y_i) \exp(\beta' \mathbf{x}_i)$

A retrospective gene study

- 1504 subjects randomized to treatment or placebo
- 23 baseline (17 ordered, 6 categorical) and 282 genetic (cat.) variables
- 95% of subjects have missing values; only 7 variables are complete



Gs model (Gi gives no tree)



At each node, a case goes to the left child node if stated condition is satisfied.

Sample sizes are beside terminal nodes.

95% bootstrap intervals for relative risk of treatment vs. placebo below nodes.

Confidence interval estimation

Fact: 95% naïve interval for node treatment mean μ

$$\bar{y} \pm 2\text{SE}(\bar{y}) = \bar{y} \pm 2\hat{\sigma}/\sqrt{n}$$

is often too short

Reason: $\hat{\sigma}$ ignores variance due to split selection

Solution: Find another answer for SE$(\bar{y})$ that includes this variance

Idea: Use bootstrap to estimate SE$(\bar{y})$

Difficulty: Every bootstrap sample gives a different tree

Question: How to relate terminal nodes (subgroups) from different trees?

Coverage of 95% CIs for treatment means and diff

		Naïve intervals			Bootstrap intervals		
n	Expt	$\mu(t,0)$	$\mu(t,1)$	$d(t)$	$\mu(t,0)$	$\mu(t,1)$	$d(t)$
162	M1-Gi	0.821	0.811	0.818	0.892	0.955	0.934
	M1-Gs	0.819	0.800	0.857	0.907	0.952	0.935
	M2-Gi	0.835	0.846	0.836	0.937	0.947	0.941
	M2-Gs	0.871	0.861	0.907	0.953	0.965	0.942
324	M1-Gi	0.880	0.874	0.889	0.903	0.972	0.957
	M1-Gs	0.869	0.862	0.888	0.916	0.967	0.955
	M2-Gi	0.896	0.915	0.911	0.966	0.967	0.963
	M2-Gs	0.888	0.913	0.916	0.968	0.973	0.950

Based on 1000 simulations with 100 bootstraps per trial; $d(t) = \mu(t,1) - \mu(t,0)$

Key idea #6:
perturb population instead of data

1. Let T be the constructed tree model and t_0 be a fixed subgroup (node)
2. If population is known, we can find $\mu(t_0, z) = E(Y|Z = z, t_0)$
3. Since population is unknown, use bootstrap to estimate it
4. For each bootstrap sample, construct a tree model T^* (that generates Y^*)
5. Find $\mu^*(t_0, z) = E(Y^*|Z = z, t_0)$
6. Repeat bootstrap many times to get SD of $\mu^*(t_0, z)$
7. Use SD instead of $\hat{\sigma}/\sqrt{n}$ in naïve interval for t_0

Computational times (sec.) for Model 1

Gs	Gi	Gc	IT	VT	SI	QU
4.3	7.0	17.5	130.1	341.1	1601.5	NA

1. Average times over 500 trials to construct 1 tree on 2.66GHz Intel *i*3
2. X variables take values 0, 1, 2
3. Y and Z are binary
4. QU does not allow categorical X variables
5. Relative speeds of Gs, Gi and Gc faster if X's have more distinct values
6. Only Gi, Gs and IT are applicable to censored Y, but IT software for it is not available
7. Only Gi and Gs allow Z variables with 3 or more levels

Acknowledgments

- Elise Dusseldorp, Netherlands Organisation for Applied Scientific Research (for QUINT R code)
- Jared Foster, University of Michigan, Ann Arbor (for VT R code)
- Jue Hou, University of Illinois, Urbana (for SIDES R code)
- Xiaogang Su, University of Texas, El Paso (for IT R code)
- Research partially supported by U.S. Army Research Office, National Science Foundation, National Cancer Institute, and Eli Lilly & Co.

References

Aitkin, M. and Clayton, D. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics*, 29:156–163.

Dusseldorp, E. and Van Mechelen, I. (2013). Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. *Statistics in Medicine*. In press.

Foster, J. C., Taylor, J. M. G., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30:2867–2880.

Italiano, A. (2011). Prognostic or predictive? It's time to get back to definitions! *Journal of Clinical Oncology*, 29:4718.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search — a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30:2601–2621.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12:361–386.

Loh, W.-Y. (2009). Improving the precision of classification trees. *Annals of Applied Statistics*, 3:1710–1737.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., and Boivin, J. R. (2005). Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. *Statistics and Computing*, 15:231–239.

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., and Bogong, L. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10:141–158.

Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S. (2008). Interaction trees with censored survival data. *International Journal of Biostatistics*, 4. Article 2.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17:492–514.