IPS / Translational Bioinformatics

Choosing a Threshold for Patient Selection

18 November 2021

Abraham Apfel, Jessica Landis, Sarah Hu, Scott Chasalow

Bristol Myers Squibb™

Summary of Conclusions from Dichotomania Talk

- Dichotomizing can obscure important patterns in data
- Dichotomous patterns usually are unrealistic
- Models with categorized continuous predictors are less efficient (waste information)
- Loss of information by categorizing can cause failure to adjust for imbalance in prognostic factors
- Categorizing continuous confounders can lead to inflation of Type I error rates
- Using an “optimal” or arbitrary cut point
  - Usually, will not generalize well to external data
  - Can inflate both Type I and Type II error rates
  - An “optimal” cut point may have large uncertainty (e.g., wide confidence interval)
- Dichotomizing a continuous endpoint reduces effective sample size (i.e., wastes samples)
  - This can increase risk of over-fitting and decrease power
- Common excuses for dichotomizing are based on fallacies

Nice! We have a robust, well-performing model. Now what?

- We do not recommend dichotomizing continuous markers during the model building process
- **However,** we acknowledge there may come a time when dichotomizing becomes necessary, (e.g., identifying a threshold for patient selection for a future trial)
- Therefore, we recommend categorizing a continuous marker, or model score, as late as possible: after a model already has been built and the only remaining task is threshold selection

Motivation

- Given a robust model that yields a reliable predictive score, f(**X**), for each subject, how can we use this model to aid in:

1. Patient stratification – Use model to divide patients into subgroups within which treatment will be randomly allocated
   - Stratify subjects according to predicted scores, f(**X**), from model
   - Group by quantiles of f(**X**)
   - We recommend allowing for multiple intervals

2. Patient selection – Use model to specify subject-inclusion criteria
   - **This is the focus of our discussion**

Approaches to Patient Subgroup Selection while Simultaneously Building Model

1. SIDES (Subgroup Identification based on Differential Effect Search)[1]
   - Uses recursive partitioning algorithm with numerous splits of data to find subgroup of patients likely to experience significant treatment benefit
2. GUIDE[2]
   - Identifies subgroup via regression trees
3. Virtual Twins[3]
   - Random Forests followed by CART to identify potential subgroups
4. Virtual Twins + GUIDE (VG)[4]
5. TSDT[5] (Treatment-Specific Subgroup Detection Tool)
   - Subsampling-based (without replacement) recursive partitioning
6. BATTing[6] (Bootstrapping and Aggregating of Thresholds from Trees)
   - Build single cutoff of biomarker on each of B bootstrap datasets and choose median from distribution of cutoffs

Issues with Above-Mentioned Approaches

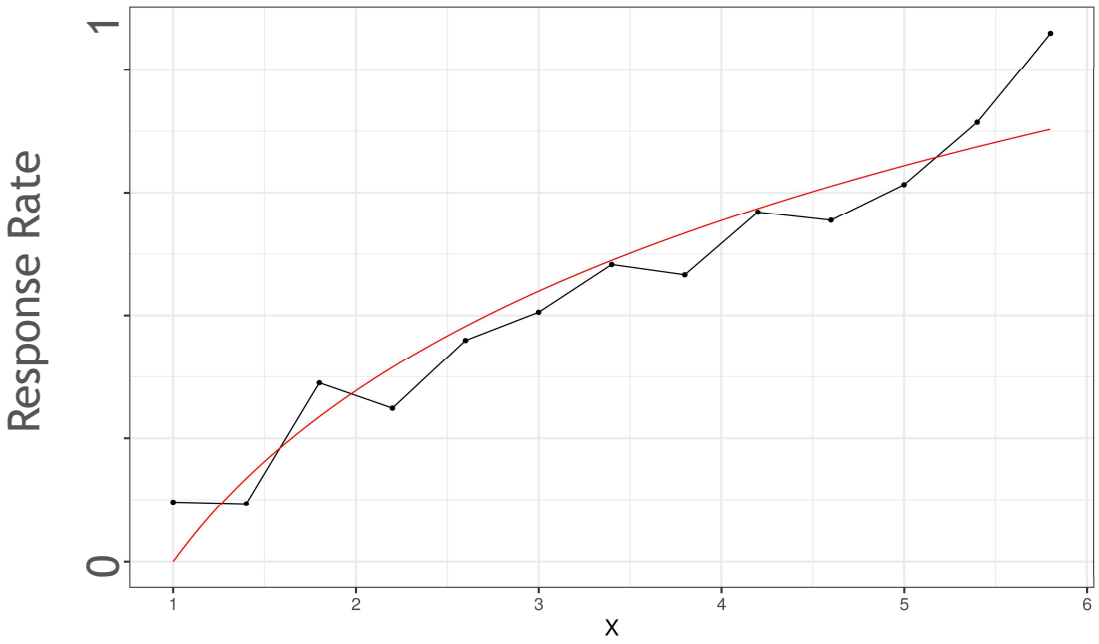
- All of the previously mentioned methods attempt to train model while simultaneously searching for subgroups via regression trees
  - Results in model built from dichotomized continuous predictor at “optimal” cutpoint
  - A model built to simultaneously determine a cutoff and obtain precise estimates will tend not to generalize to external data as well as properly planned regression modeling followed by cutoff determination (see Dichotomania talk regarding generalizability of cutoff “identification”)
- Better to use a method which makes use of model score and selects appropriate cutoff in a post-hoc manner (after model is already built)
  - This allows model to be built using continuous biomarkers in their original form to make the most use of information available.
  - The alternative approach of simultaneously searching for subgroups while building the model via Regression Trees implies dichotomizing the biomarkers which would entail throwing away information

Use of Model Score, Not Individual Biomarkers

- We recommend using the score from a multivariable model as a cutoff rather than using the value of a specific biomarker.
  - When other covariates are included in the model, it is misleading (biased) to apply cutoff to an individual risk factor, because there are other contributors to outcome predictions (i.e., risk is determined by all terms in model, not individual risk factor)
  - Simple analogy: If decision was being made based on BMI, we wouldn't use separate cutoffs for height and weight; rather would use a single cutoff for BMI
  - This is true whether or not there are interactions amongst covariates

Use of Model Score, Not Individual Biomarker (cont.)

- Even if model consists of only single biomarker interacting with treatment, still preferable to use model score
  - Smooth estimate from model will yield more realistic, less noisy estimates of performance at given thresholds
  - Under certain, usually reasonable, assumptions, can combine information from both treatment arms to yield more precise predictions



2 Potential Approaches

- P-value ranking approach
- Current Approach (recommended)

P-value Ranking Approach[7]

- Provides ranking of candidate cutoffs from continuous model score via hypothesis testing
- **Advantage** over earlier described methods since selects cutoff in a post-hoc manner from model with continuous predictors
  - Also makes use of full model score rather than cutoff of individual biomarkers
- **Disadvantage**:
  - Prevalence is factored into p-value calculation (low prevalence will naturally yield higher p-value), yielding an automatic selection of cutoff without using discretion of analyst
    - Acceptable levels of prevalence may vary per situation

Current Approach - Recommended

- Method:
  - Identify which performance metrics are most important to you (PPV, NPV, Sensitivity, Specificity, etc.)
    - If there is interaction between treatment and predictor(s), plot difference of performance metric between the 2 arms
  - Plot relevant performance metrics across range of continuous score from model (f(**X**))
    - In presence of interaction, plot difference in score between the 2 treatment arms on x-axis
  - Plot prevalence curve (% subjects with f(**X**) > than given cutoff) across range of f(**X**)
  - Decide which cutoff makes the most sense, based on optimizing most relevant metrics and prevalence
- Advantages:
  - Makes use of score from continuous, robust model and therefore not reliant on less robust regression tree approaches
  - Allows the user to decide how much weight to place on prevalence and relevant metrics

More detailed methods for a two-arm case

1. Use the model and predictor values to calculate predicted probabilities of response for each subject, if they receive placebo or drug:

Subject	Covariate 1	Covariate 2	Covariate 3	Covariate 4	Treatment
SUBJ-1	-0.5754	0.6384	-0.1506	0.0441	Placebo
SUBJ-1	-0.5754	0.6384	-0.1506	0.0441	Drug
SUBJ-2	2.0825	-1.4322	-0.2090	1.6164	Placebo
SUBJ-2	2.0825	-1.4322	-0.2090	1.6164	Drug

model → prediction

Subject	Treatment	P(Response)
SUBJ-1	Placebo	0.3584
SUBJ-1	Drug	0.6404
SUBJ-2	Placebo	0.0220
SUBJ-2	Drug	0.4642

difference of predicted probs →

Subject	Score
SUBJ-1	0.2820
SUBJ-2	0.4422

2. Calculate the actual/observed response rate in the placebo arm and drug arm, and the difference in response rates, among subjects having score ≥ cutoff across a range of cutoff values

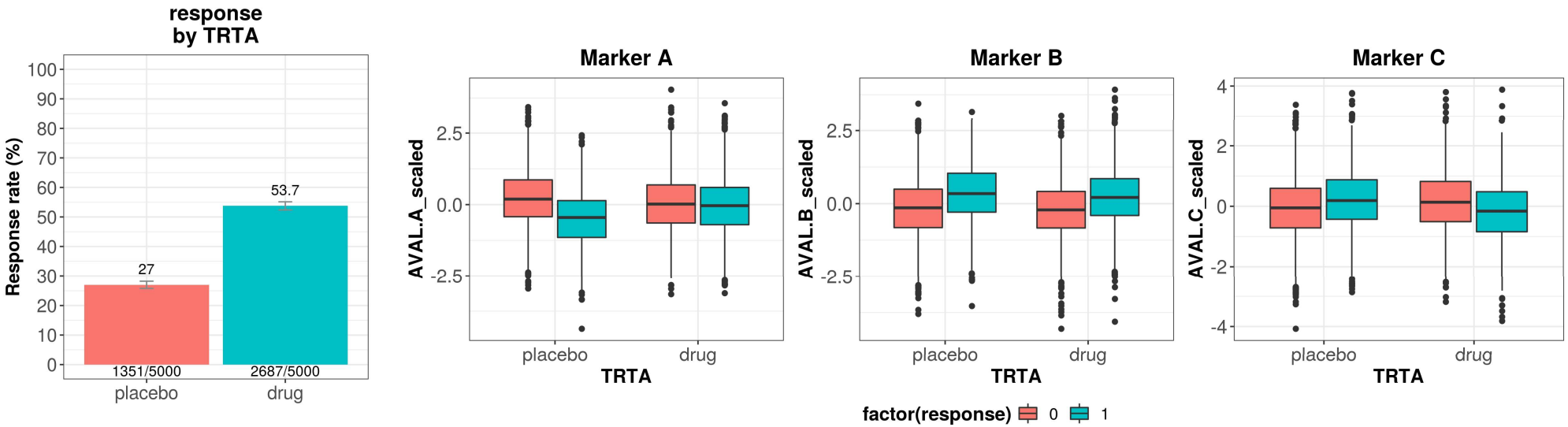
Subject	Actual Trt	Observed Response	Score
SUBJ-15	Drug	0	-0.466
SUBJ-41	Drug	0	-0.4246
SUBJ-19	Placebo	1	-0.3321
SUBJ-104	Drug	1	-0.3313
SUBJ-77	Placebo	1	-0.3161
...	...	...	...

Observed response rates →

Cutoff	N Placebo	N Pbo Responders	Pbo Response Rate	N Drug	N Drug Responders	Drug Response Rate
-0.5	250	68	0.272	250	135	0.54
-0.4	250	68	0.272	248	135	0.544
-0.3	246	64	0.26	247	135	0.547
...	...	...	...	...	...	...

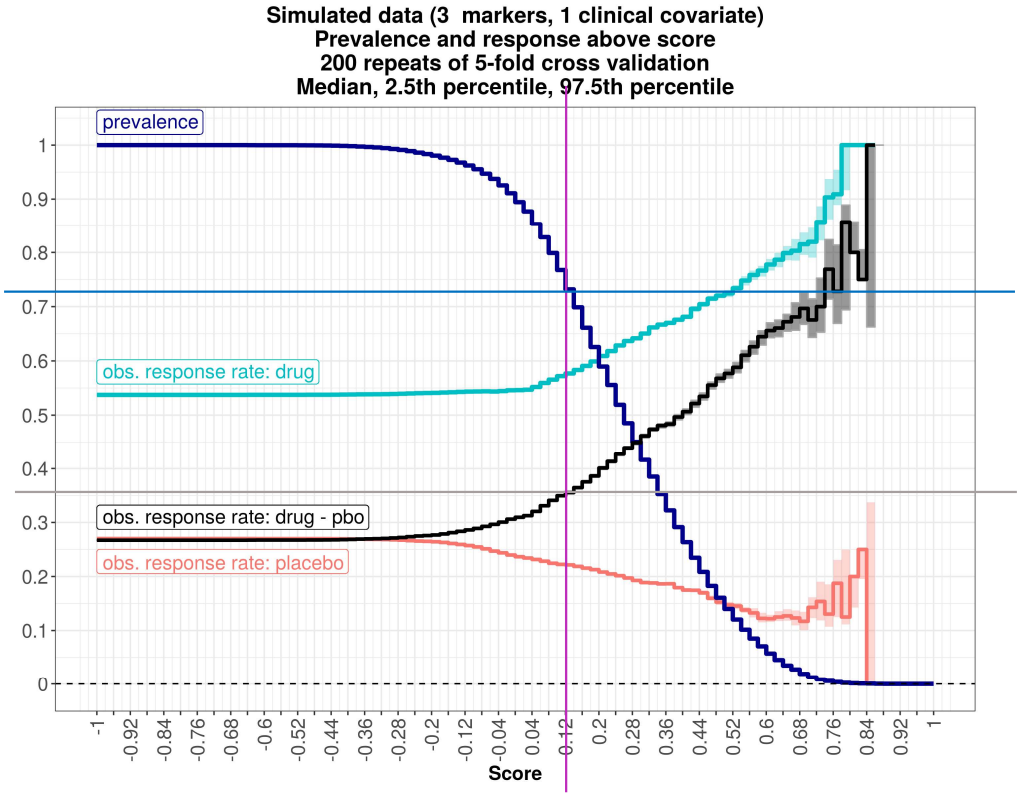
3. In consultation with collaborators, identify a target treatment benefit (depending on drug risk profile, competitive landscape, cost, etc.) and evaluate results:
   - Assess whether you can achieve the targeted drug-placebo response difference with an acceptable % of included patients
   - Assess whether a more stringent score cutoff would be justified by gains in treatment benefit

Simulated data example



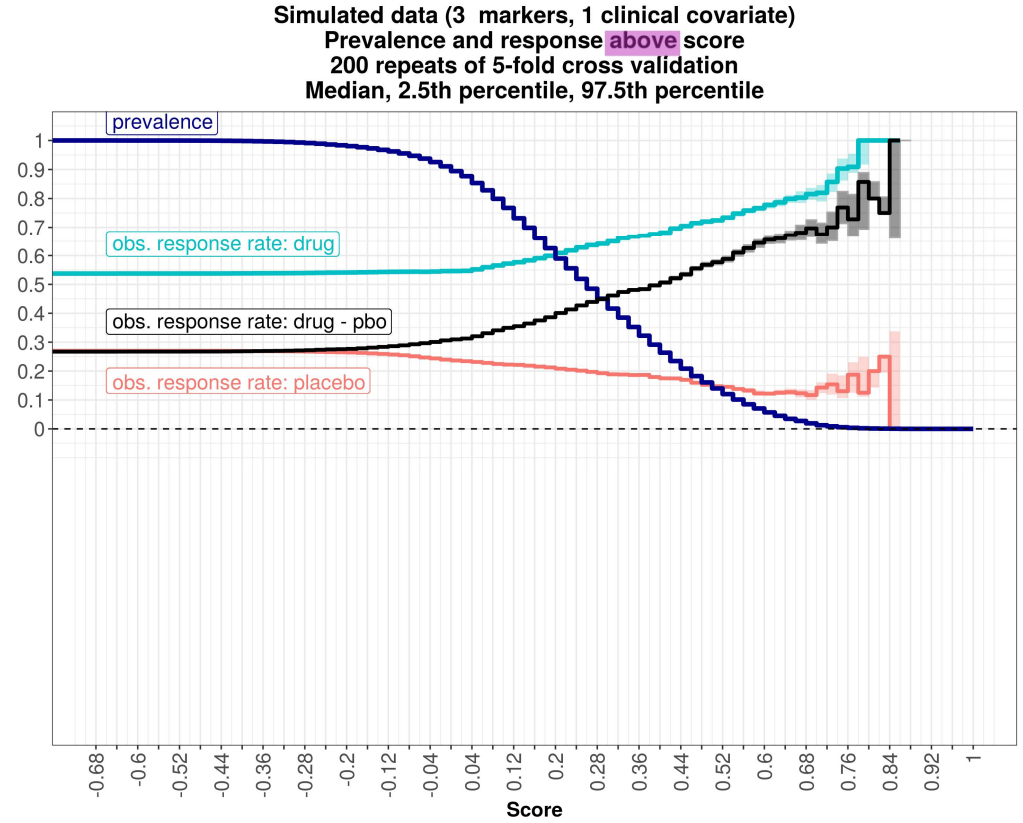
- 2 arms, 3 markers, 1 numeric covariate
- Model:
  log(response_odds) = -1.386294 + 1.586965*(arm == drug) + -0.9*markerA_scaled + 0.9*(arm == drug)*markerA_scaled + 0.5*markerB_scaled + 0.4*markerC_scaled + -0.8*(arm == drug)*markerC_scaled + -0.2*baseline_covariate_centered
- N = 5000/arm

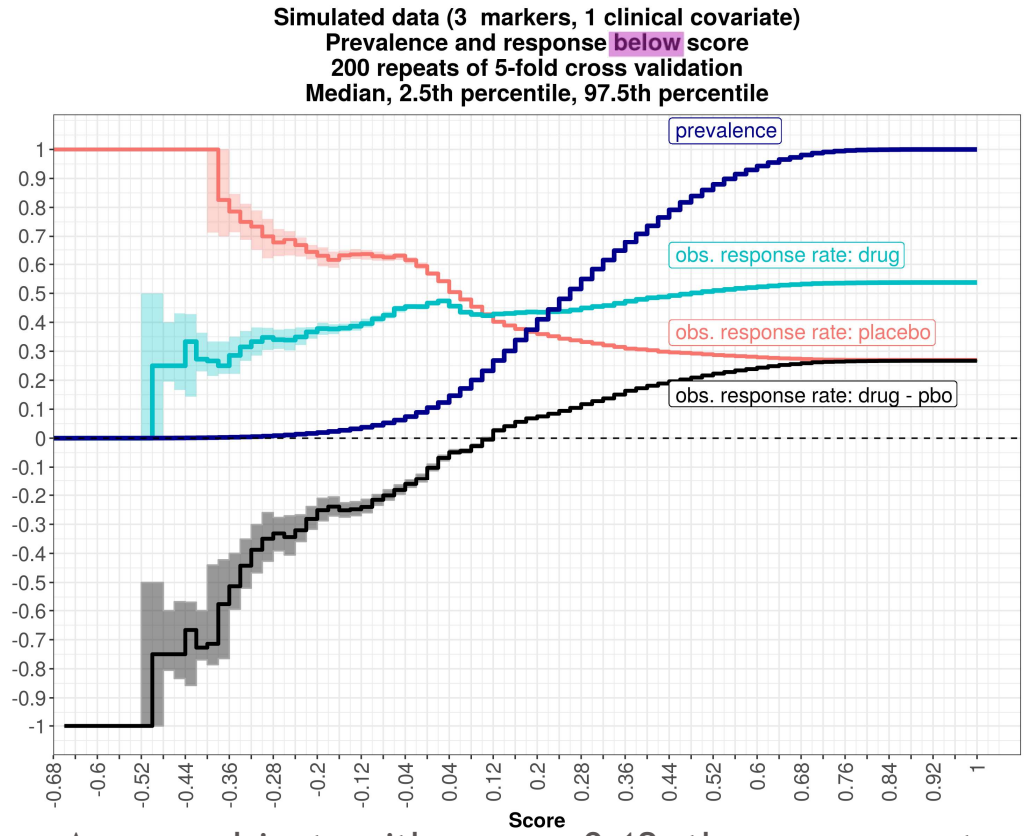
Simulated data – recommended approach



- Overall, the response rates including all subjects (i.e. score ≥ -1) are 53.7% and 27.0% for drug and placebo, respectively; the difference is 26.7%
- Among subjects with score ≥ 0.12, the response rate difference is 35.5% (35.5,35.7)
  - 73.2% (73.0,73.7) of subjects have score ≥ 0.12
- At prevalence 51.9% (51.8,52.1%) -- score ≥ 0.24 -- the observed response rate difference is 42.7% (42.4,43.0)

Another view of scores vs. metrics





- Among subjects with score ≥ 0.12, the response rate difference is 35.5% (35.5,35.7)
  - 73.2% (73.0,73.7) of subjects have score ≥ 0.12
- Among subjects with score < 0.12, the response rate difference is 2.7% (2.1,3.3)
  - 26.8% (26.6,27.0) of subjects have score < 0.12

Limitations of previous analyses

- Dichotomizing a continuous predictor wastes information
- Choice to dichotomize at the median value is somewhat arbitrary
  - Maybe we could achieve an acceptable treatment benefit while retaining >50% of the subjects
- Analyses have been performed on each biomarker separately; not clear how to combine multiple predictors

Summary

- Dichotomizing continuous markers is wasteful and usually unrealistic
- If you must dichotomize, e.g. for patient selection for a future trial, it is better to dichotomize the model predictor score rather than individual predictors
- We have demonstrated how the recommended approach can be applied to the two-treatment scenario

References

[1] Lipkovich, Ilya, et al. "Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations." *Statistics in medicine* 30.21 (2011): 2601-2621.

[2] Loh, Wei-Yin, Xu He, and Michael Man. "A regression tree approach to identifying subgroups with differential treatment effects." *Statistics in medicine* 34.11 (2015): 1818-1833.

[3] Foster, Jared C., Jeremy MG Taylor, and Stephen J. Ruberg. "Subgroup identification from randomized clinical trial data." *Statistics in medicine* 30.24 (2011): 2867-2880.

[4] Jia, Jia, Qi Tang, and Wangang Xie. "A Novel Method of Subgroup Identification by Combining Virtual Twins with GUIDE (VG) for Development of Precision Medicines." *Design and Analysis of Subgroups with Biopharmaceutical Applications* (2020): 167.

[5] Shen, Lei, et al. "Subgroup Identification for Tailored Therapies: Methods and Consistent Evaluation." *Design and Analysis of Subgroups with Biopharmaceutical Applications* (2020): 181.

[6] Huang, Xin, et al. "Patient subgroup identification for clinical drug development." *Statistics in medicine* 36.9 (2017): 1414-1428.

[7] Huang, Xin, et al. "Exploratory Subgroup Identification for Biopharmaceutical Development." *Design and Analysis of Subgroups with Biopharmaceutical Applications* (2020): 245.

Backup

P-value Ranking Approach[7]

- Prognostic case method:
  1. Optimize $Y \sim \alpha + \beta\omega(\mathbf{X})$
  2. Rank by p-value of β
  3. If $f(\mathbf{X})$ was not yet validated (e.g. via optimism-adjusted bootstrap), optimize $f(\mathbf{X})$ on training set to determine $\omega(\mathbf{X})$ in test set. Then run $\alpha + \beta\omega(\mathbf{X})$ on test set to select best cutoff
     - Repeated CV
     - We recommend using outer loop of CV to measure performance of method

Let:
Y := endpoint (e.g. OS, BOR, continuous biomarker)
$f(\mathbf{X})$:= score from model

$$\omega(\mathbf{X}) := \begin{cases} 1 \text{ if } f(\mathbf{X}) > \text{cutoff} \\ 0 \text{ if } f(\mathbf{X}) <= \text{cutoff} \end{cases}$$

t := treatment indicator

- Predictive case:
  - Same except use $Y \sim \alpha + \beta(\omega(\mathbf{X})*t) + \gamma t$ where t = treatment indicator

- Disadvantage:
  - Prevalence is factored into p-value calculation (low prevalence will naturally yield higher p-value), yielding an automatic selection of cutoff without using discretion of analyst
    - Acceptable levels of prevalence may vary per situation

Simulations – Continuous Outcome