Optimizing personalized treatment in dose ranging study

Xiwen Ma[1], Wei Zheng[1], Yang Liu[2] and Yuefeng Lu[1]

[1]Sanofi
[2]University of Connecticut

July 5, 2017

Table of contents

1. Introduction
    - Personalized medicine
    - Biomarker
2. Challenges
    - Missing data
    - Reproducibility
    - Population diversity
3. Existing Methods
    - Methods
    - Comparison of tree based methods
4. Subgroup analysis in dose ranging study
    - Model
    - The INIA method
    - Examples
5. Discussion

Personalized medicine

- For pharmaceutical interventions, it's well known that the strategy of "one-size fits all" is hardly applicable to most common diseases.
- It's been reported that the percentage of patients for whom drugs are ineffective ranges from 38% to 75% for several major diseases, due to the heterogeneity of patient population, complex underlying pathophysiology, and inadequate or inappropriate dosing regimens among other factors [9].
- Personalized medicine involves developing and validating evidence-based treatment algorithms to match a **right patient** with the **right treatment**, at the **right dose** and at the **right time**.

Biomarker

- **Bio**marker: Any measurable substance, structure, or process **in the body** that can influence or predict the incidence of treatment outcome or disease.

Biomarker can be anything

- Potential biomarkers: whole genome sequencing, RNA seq, microRNA, proteomics, metabolomics, and many others

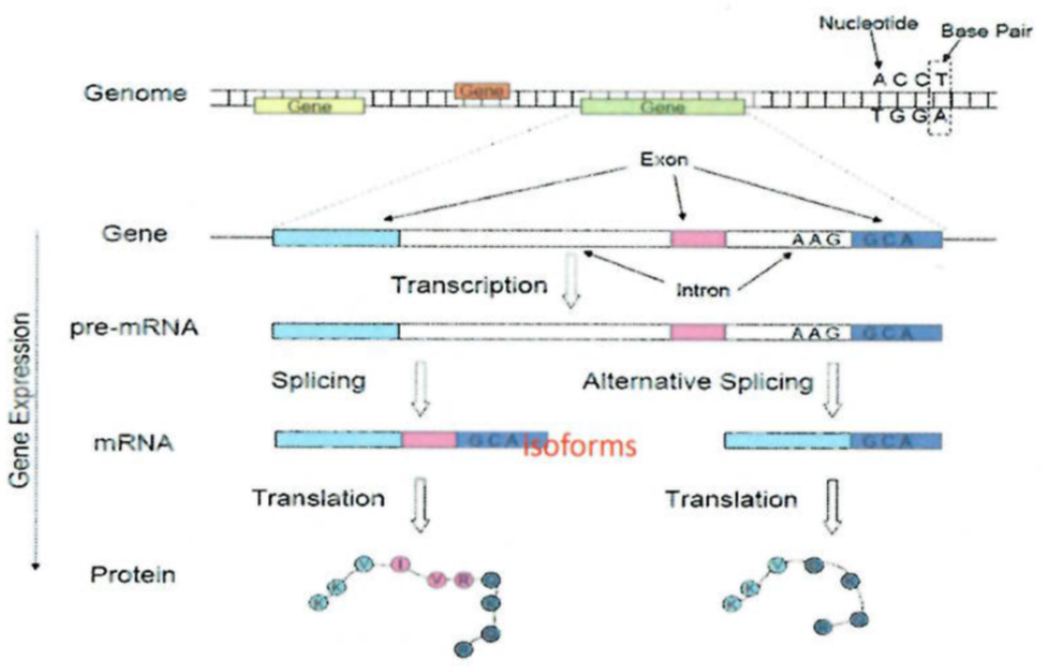
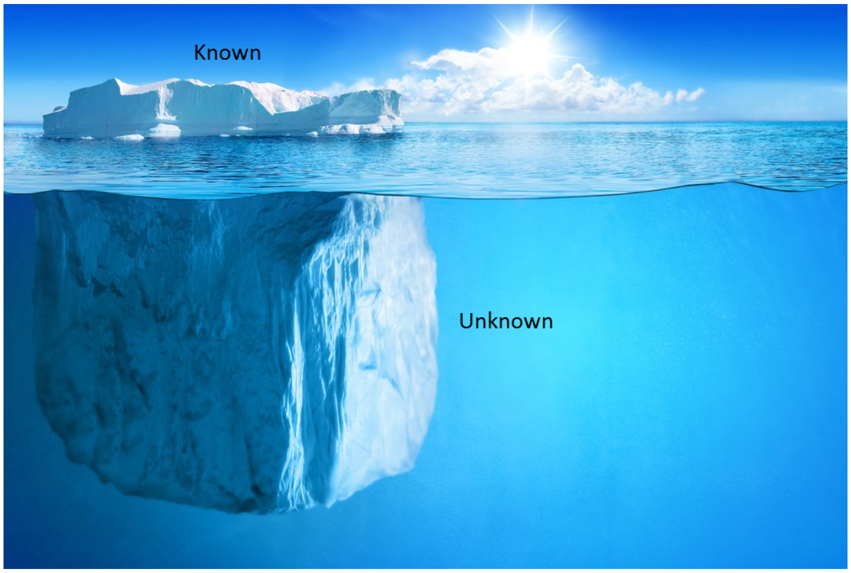


Figure: Dogma of molecular biology

The iceberg of biomarker

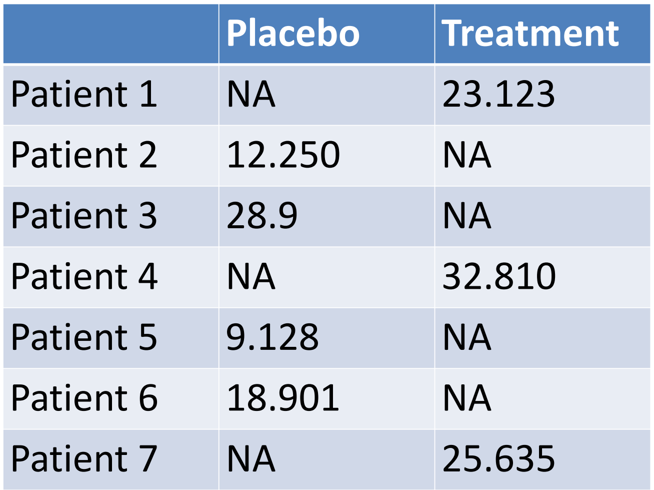
- Biomarker identification is so difficult.
- Very few biomarkers (of high prevalence) have been discovered.



Two arm randomized clinical trial

- In a clinical trial, patients are randomized into 2 treatments groups $\mathcal{A} = \{d_1, d_2\}$, where d_1 denotes placebo group.
- Let $X = (X_1, X_2, ..., X_p)' \in \mathcal{X}$ denote the vector of baseline candidate biomarkers.
- Let Y denote the clinical outcome (either continuous or binary). We assume that a larger value of Y indicates a more favorable clinical outcome.

A simple example



	Placebo	Treatment
Patient 1	NA	23.123
Patient 2	12.250	NA
Patient 3	28.9	NA
Patient 4	NA	32.810
Patient 5	9.128	NA
Patient 6	18.901	NA
Patient 7	NA	25.635

- Each patient was randomly assigned either placebo or treatment.
- The goal is to identify the real biomarkers in X and find the optimal decision rule based on biomarkers.

Personalized treatment

- A personalized treatment can be viewed as a decision rule from the biomarker space $\mathcal{X}$ to the treatment space $\mathcal{A}$:

$$D \ : \ \mathcal{X} \rightarrow \mathcal{A}. \tag{1}$$

- The optimal decision D^* of personalized treatment is to maximize the expected clinical outcome:

$$D^* = \text{argmax}_D E\left[E\left[Y|X, D(X)\right]\right], \tag{2}$$

  Or equivalently minimize the "error":

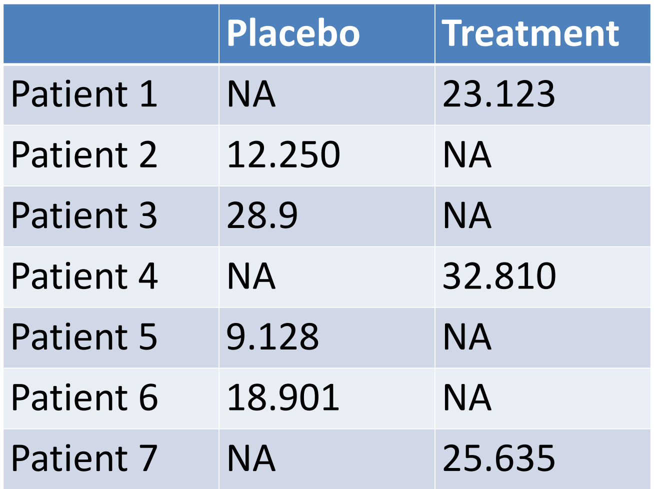
$$D^* = \text{argmin}_D E\left[E\left[Y|X, D^c(X)\right]\right], \tag{3}$$

  where $D^c(X)$ is the alternative assignment to $D(X)$.

- **Fact:** the solution D^* dose not change if Y is replaced to $Y + c$ for any constant c.

Challenge 1: missing data

- Ideal solution: compare $E[Y|X, d_1]$ to $E[Y|X, d_2]$.
- Reality: only ONE treatment can be applied to each patient. Half of the data is **missing**.



	Placebo	Treatment
Patient 1	NA	23.123
Patient 2	12.250	NA
Patient 3	28.9	NA
Patient 4	NA	32.810
Patient 5	9.128	NA
Patient 6	18.901	NA
Patient 7	NA	25.635

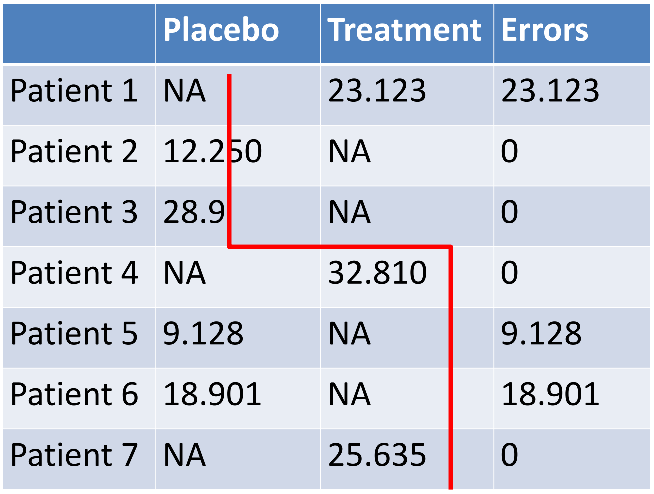
Empirical objective function

- Using the observed data, the empirical error was defined by:

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}\frac{I(A_i \neq D(\mathbf{x}_i))}{P(A_i)}\mathbf{y}_i \\ = & n^{-1}\sum_{i=1}^{n}\frac{\mathbf{y}_i}{P(A_i)}I(A_i \neq D(\mathbf{x}_i)). \end{aligned} \tag{4}$$

  where A_i is the true treatment assignment for the ith patient.
- Such an empirical error has been commonly used in practices (e.g., Outcome Weighted Learning [15] and Modified Covariate [11]).

Issue



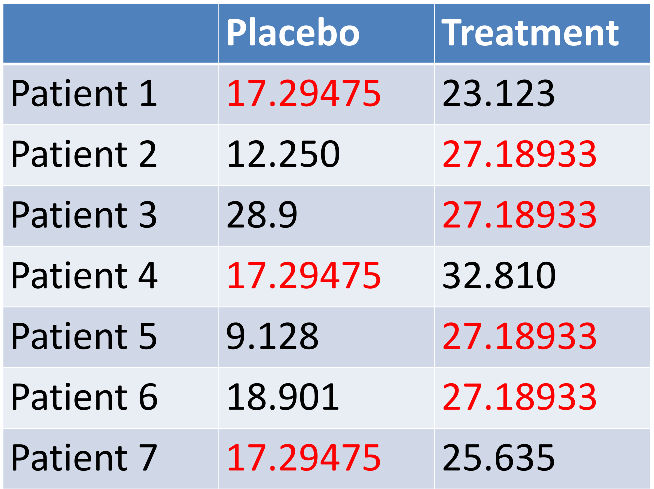
	Placebo	Treatment	Errors
Patient 1	NA	23.123	23.123
Patient 2	12.250	NA	0
Patient 3	28.9	NA	0
Patient 4	NA	32.810	0
Patient 5	9.128	NA	9.128
Patient 6	18.901	NA	18.901
Patient 7	NA	25.635	0

$$n^{-1}\sum_{i=1}^{n}\frac{\mathbf{y}_i}{P(A_i)}I(A_i \neq D(\mathbf{x}_i)).$$

- For subjects with the same treatment assignment, the empirical error only adds 0 in the summation, rather than the outcome from the alternative assignment.
- The solution will change if we add a constant c to each y_i.
- Biased estimation.

Augmented data

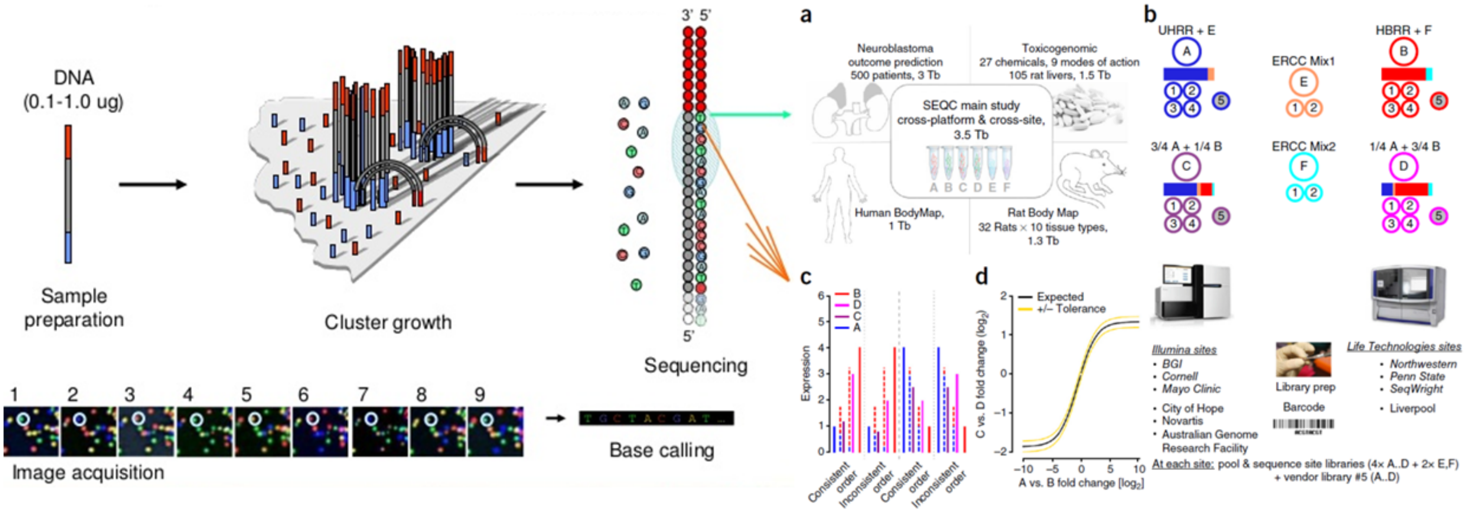
- Data augmentation: For each patient, estimate the clinical outcome for the alternative treatment that was not assigned to this patient. Add these estimated outcomes to the original data.
- In the augmented data, each patient has clinical outcomes from both placebo and treatment.
- Much better performance in simulations.



	Placebo	Treatment
Patient 1	17.29475	23.123
Patient 2	12.250	27.18933
Patient 3	28.9	27.18933
Patient 4	17.29475	32.810
Patient 5	9.128	27.18933
Patient 6	18.901	27.18933
Patient 7	17.29475	25.635

Challenge 2: reproducibility

- Data quality control of complex assays or newer technology.
- Data processing methods.
- Independent validation.

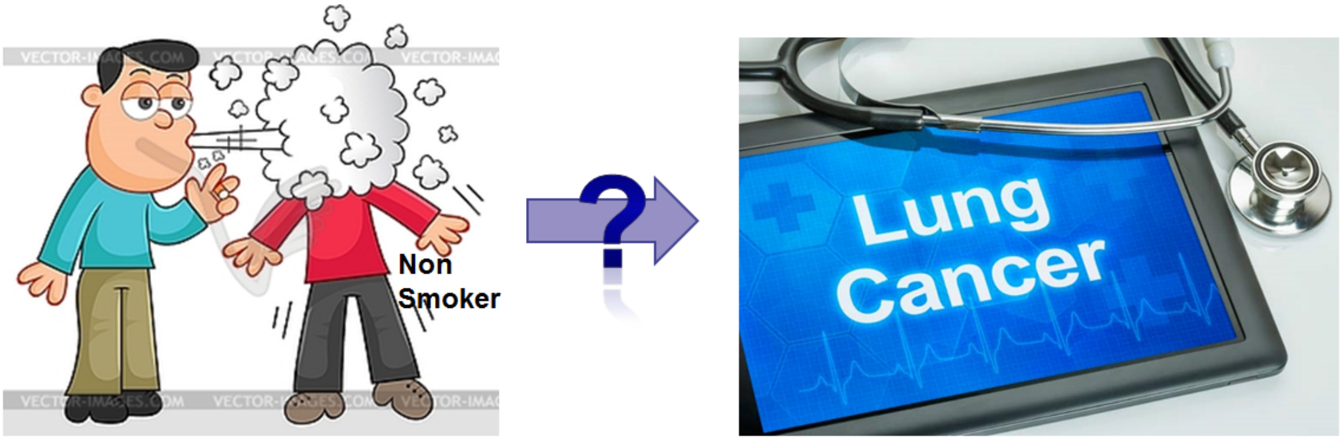


A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium. *Nature Biotechnology*, 2014

Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biology, 16:133,* 2015

Challenge 3: population diversity

- The non-smoker cancer signature is distinct to smoker
- Asian smoker cancer signature is similar to European smoker
- Asian non-smoker signature is distinct to European non-smoker
- High incidence of lung cancer in Asian never-smokers is NOT due to second hand smoke.



Whole Genome Sequencing of Asian Lung Cancers Reveals Asian Never-Smokers have Distinct Molecular Signature from Smokers. *Cancer Research*, 74(21):6071-6081, 2014

Summary

A good exporatory/confirmatory biomarker study must have

- A solid scientific background with uncomplicated biomarker hypothesis.
- Good sample size.
- Good instrument for biological measurement.
- **Good analytical/statistical methods.**

Methods

- Tree-based methods:
  - Perform partitioning to the covariate space to establish a tree structure and mutually exclusive subgroups;
  - Deliver a patient partition using binary conditions in each step and the resulted structure is relatively easy for interpretation.
- Model-based methods:
  - Are often applied in the optimal treatment regime studies under a penalized regression framework;
  - Classify patients with “black-box” mechanisms which may be difficult to interpret.
- Bayesian subgroup methods:
  - Utilize model selection ideas;
  - Allow incorporation of prior information.

Methods in Comparison

- Interaction Tree (IT) procedure [10];
- Qualitative Interaction Trees (QUINT) procedure [4];
- Virtual Twins (VT) procedure [5];
- Generalized Unbiased Interaction Detection and Estimation (GUIDE) procedure [6].

Interaction Tree (IT) Procedure

- Adopts a CART-based recursive partitioning algorithm, guided by the strength of interaction between the treatment and a potential subgroup.
- Enumerates and examines all possible splits at each step.
- Grows a large initial tree and perform pruning procedure using v-fold cross validation or bootstrapping methods.
- Provides an optional terminal node amalgamation procedure and a variable importance ranking algorithm.

Qualitative Interaction Trees (QUINT) Procedure

- Only focuses on the detection of qualitative interaction, where the treatment effect in one subgroup has a different sign comparing to another subgroup.
- Performs recursive partitioning guided by a criterion, which considers both the scaled treatment effect size in each child group and the resulted group size.
- Performs bootstrap-based pruning procedure after growing the initial tree.

Virtual Twins (VT) Procedure

- Aims to identify a covariate subspace A, classify patients to A and A^c.
- Fits a random forest, and constructs a "virtual twin" by making prediction with reversed treatment assignment.
- Denotes the treatment difference between the"twins" as a new variable Z.
- VT(R) Approach: constructs a regression tree with Z as the response, and classifies the patients into A and A^c based on the prediction of Z in terminal nodes.
- VT(C) Approach: classifies the patients into $0-1$ groups based on Z and constructs a classification tree to finally classify the patients into A and A^c.
- We apply VT(R) procedure without the final classification step and denote it as VT in the following comparison.

GUIDE Procedure

- Performs recursive partitioning guided by the test statistic of χ^2 tests on the sign of residuals across different treatment assignments.
- Mainly consists of 2 algorithms: Gi ('i' for interaction) and Gs ('s' for sum).
- GUIDE is an unbiased selection procedure while IT, QUINT and VT are not.
- GUIDE is able to handle missing data naturally without imputation.
- GUIDE is able to fit multi-arm trial data.
- GUIDE offers bootstrap confidence intervals for the treatment differences at each terminal node.

Evaluation Criteria

Evaluate the empirical performance via three aspects:

- Hypothesis testing: define type I error rate (TIE).
- Power analysis: define receiver operating characteristic (ROC) curve.
- Structure recovery: define a set of novel measure named T-AIC/T-BIC.

Simulation Setup

- Balanced treatment allocation: $trt \sim Bernoulli(0.5)$.
- Binary covariates: $X_1 \sim Bernoulli(0.5)$, $X_3 \sim Bernoulli(0.7)$.
- Normal covariates: $X_2 \sim N(0.5, 2^2)$, $X_i \sim N(0, 1)$ for $3 < i \leq p$.
- Total sample size n can take 100, 300, 500.
- Total number of candidate variables p is tested at 10, 50 separately.
- The noise level is tested at $\sigma = 1, 2$ separately.

Simulation Setup

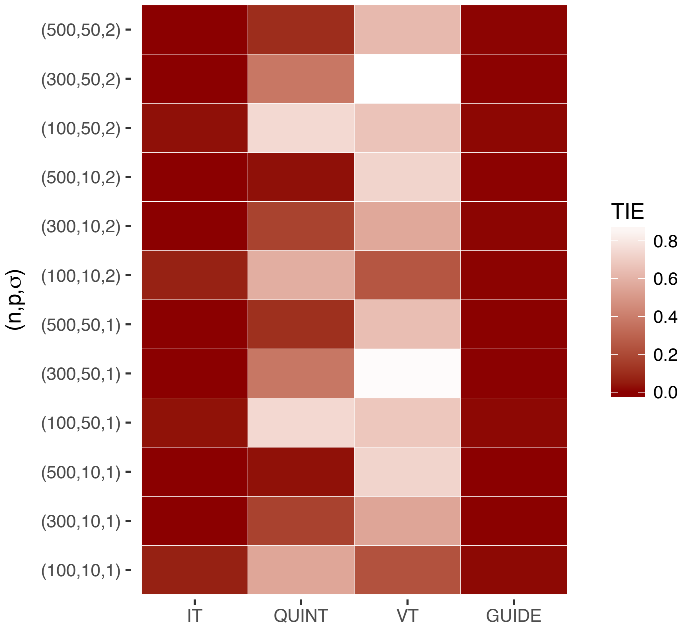
- We present the results for 2 scenarios under each of the type I error study and the power study.
- Default tuning parameter choices are used in implementation.
- In each simulation, the method(s) who produced the smallest T-AIC/T-BIC value will be “selected”.
- Denote the method selection percentage (MSP) by $\frac{\text{\# of selections}}{\text{\# of simulations}}$. It is normalized to adjust for ties.
- In type I error study, $b(X) \equiv 0$. The T-AIC/T-BIC comparison will reduce to the comparison of number of terminal nodes.

Type I Error Study Scenario 1.1

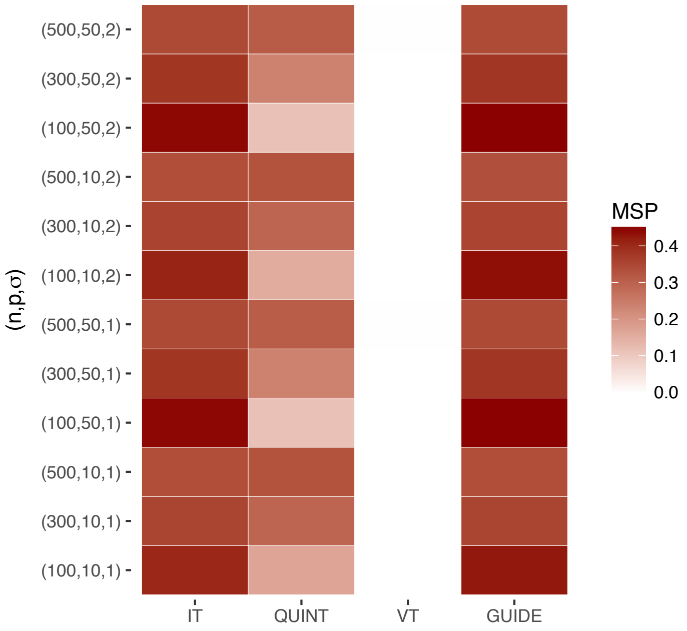
True model: $y = 2 + \epsilon$. Simulation results are averaged over 1000 repetitions.

Heatmaps from left to right: TIE andMCP by T-BIC.

TIE results for type I error study scenario 1.1



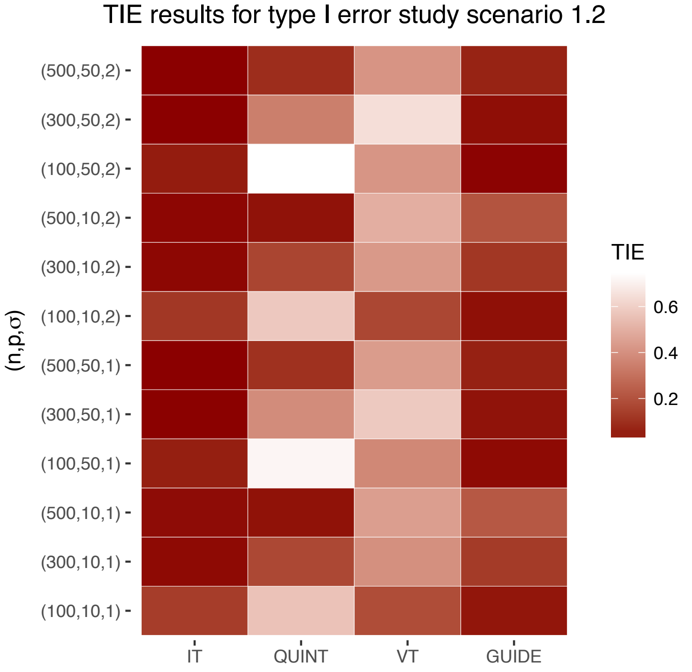
MSP results by T–BIC for type I error study scenario 1.1

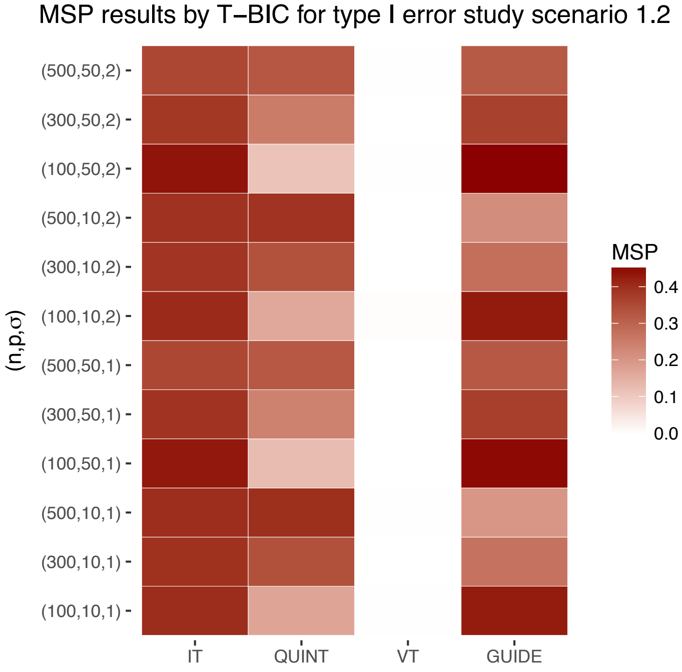


Type I Error Study Scenario 1.2

True model: $y = 2 + 2[I(X_1 = 0) + I(X_2 > 0) + \exp(X_4) + (X_5 + X_6)^2] + \epsilon$. Simulation results are averaged over 1000 repetitions.

Heatmaps from left to right: TIE and MCP by T-BIC.

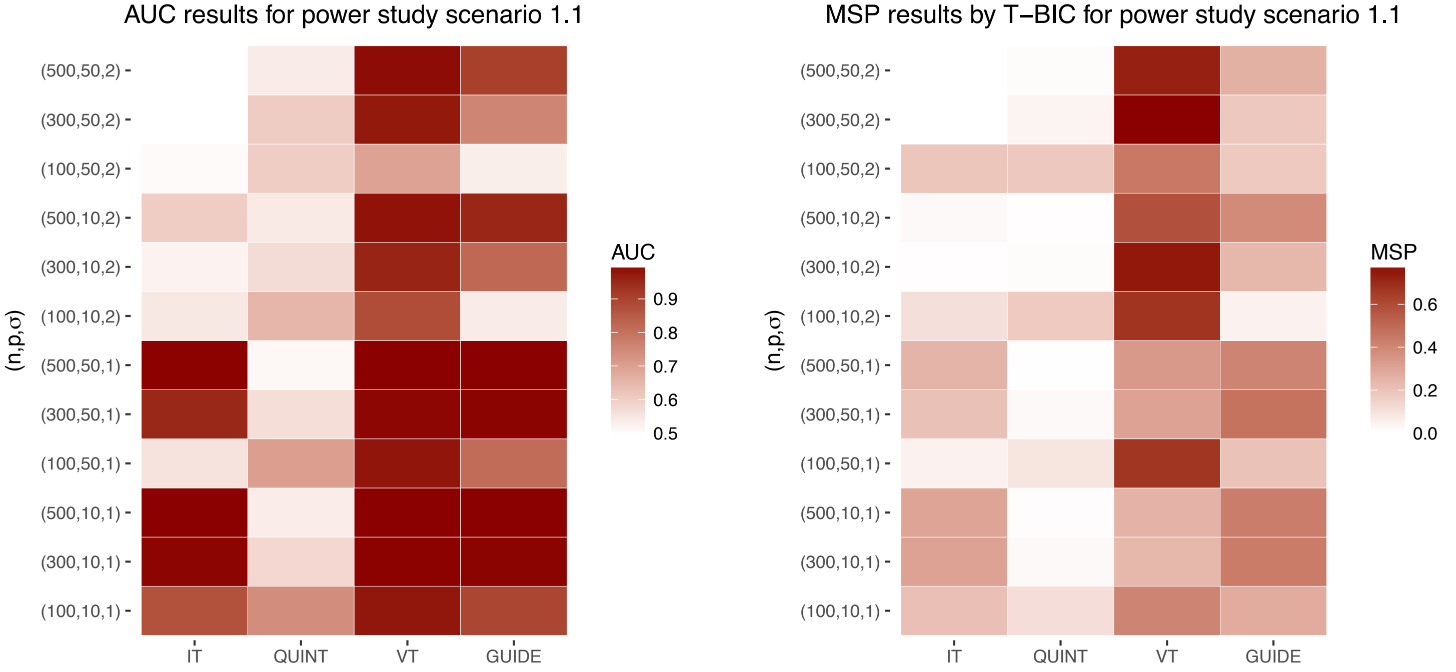

TIE results for type I error study scenario 1.2


MSP results by T–BIC for type I error study scenario 1.2

Power Study Scenario 1.1

True model: $y = 2 + 2I(X_4 > 0)trt + \epsilon$. Simulation results are averaged over 200 repetitions.

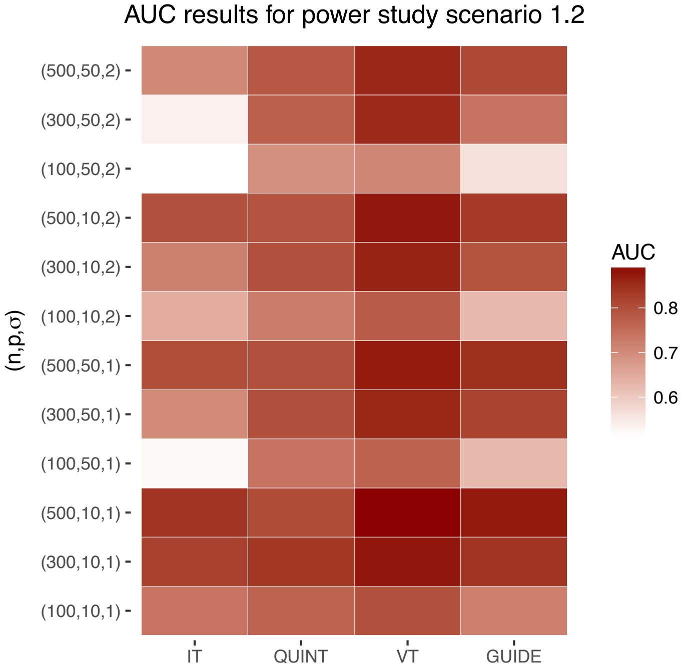
Heatmaps from left to right: AUC and MCP by T-BIC.

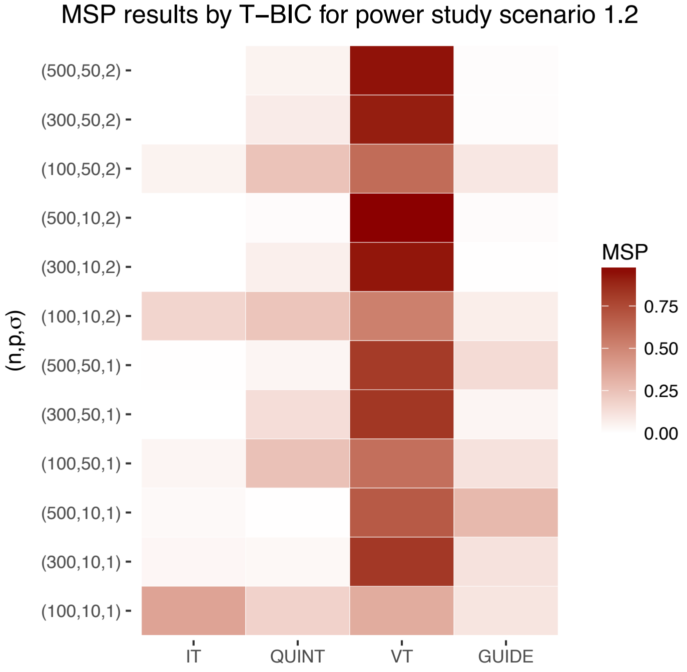


Power Study Scenario 1.2

True model: $y = 2 + 2I(X_1 = 0) + 2I(X_2 > 0) + 2(X_5 + X_6)trt + \epsilon$. Simulation results are averaged over 200 repetitions.

Heatmaps from left to right: AUC and MCP by T-BIC.





Summary

The top picks for each case below:

1. Type I Error Study:
   - Based on the TIE measure: IT, QUINT and **GUIDE**.
   - Based on the T-AIC/T-BIC measure: IT, QUINT and **GUIDE**.
2. Power Study:
   - Based on the AUC measure: VT and **GUIDE**.
   - Based on the T-AIC/T-BIC measure: VT and **GUIDE**.

Dose ranging study

- In a phase II dose-ranging study, patients are randomized into T dosing groups with doses from a dose space $\mathcal{A} = \{d_1, d_2, ..., d_T\}$, where $d_1 < d_2 < \cdots < d_T$. Note that $d_1 = 0$ denotes the placebo group.
- We assume that the mean clinical outcome in the overall population is monotonic in dose.

Assumption

For any biomarker $\mathbf{x}$*, one of the following conditions holds:*

$$
\begin{aligned}
&\textit{1.}\ E(Y|\mathbf{x}, d_1) \le E(Y|\mathbf{x}, d_2) \le \cdots \le E(Y|\mathbf{x}, d_T),\\
&\textit{2.}\ E(Y|\mathbf{x}, d_1) \ge E(Y|\mathbf{x}, d_2) \ge \cdots \ge E(Y|\mathbf{x}, d_T).
\end{aligned}
$$

Personalized dose and dose-dependent optimal subgroup

- For a given efficacy margin $\delta > 0$ and biomarkers $\mathbf{x}$, the personalized effective dose (PED) is the minimum dose at which the expected clinical outcome exceeds that of the placebo by an efficacy margin δ. More precisely, given a margin $\delta > 0$, the PED is

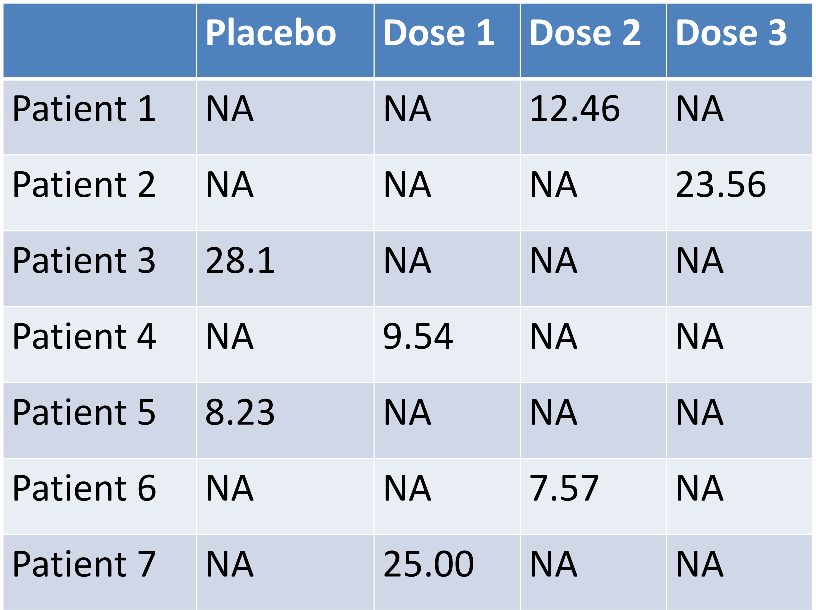
$$D^*(\mathbf{x}, \delta) = \min\{d_i \in \mathcal{A} : E(Y|\mathbf{x}, d_i) - E(Y|\mathbf{x}, d_1) \geq \delta\}.$$

- The Dose-Dependent Optimal Subgroup (DDO-Subgroup) for dose d_i and efficacy δ is defined as

$$\mathcal{S}(d_i, \delta) = \{\mathbf{x} \in \mathcal{X} : E(Y|\mathbf{x}, d_i) - E(Y|\mathbf{x}, d_1) \geq \delta\}.$$

  which is the biomarker subspace on which the efficacy margin is at least δ at dose d_i.

More missing data



	Placebo	Dose 1	Dose 2	Dose 3
Patient 1	NA	NA	12.46	NA
Patient 2	NA	NA	NA	23.56
Patient 3	28.1	NA	NA	NA
Patient 4	NA	9.54	NA	NA
Patient 5	8.23	NA	NA	NA
Patient 6	NA	NA	7.57	NA
Patient 7	NA	25.00	NA	NA

Our solution: Iterative dose-dependent Nonparametric regression with Isotonic Adjustment (INIA) [7].

Data and notation

- To simplify the notation, we denote the mean response function as

  $$g(\mathbf{x}, d) = E(Y|\mathbf{x}, d).$$

  For continuous outcomes, we have

  $$Y = g(\mathbf{x}, d) + e, \text{ where } e \sim \mathrm{N}(0, \sigma^2), \tag{5}$$

  and for binary outcomes, we have

  $$\mathrm{P}(Y = 1|\mathbf{x}, d) = g(\mathbf{x}, d). \tag{6}$$

- We assume that n_i patients are treated at the dose d_i, and denote the total sample size as n. Let $\mathbf{x}_{ij} = (x_{ij}^1, x_{ij}^2, ..., x_{ij}^p)$ denote the vector of biomarkers for the j^{th} patient at the dose d_i and y_{ij} denote the patient's clinical outcome, $i = 1, ..., T, j = 1, ..., n_i$.

Iterative dose-dependent Nonparametric regression with Isotonic Adjustment (INIA)

1 Initial fitting: Fit the regression function $\hat{g}(\mathbf{x}, d_i)$ to the data at dose d_i with a nonparametric regression method such as smoothing splines [13, 12], gradient boosting [2], or random forest [1] *etc*.

Iterative dose-dependent Nonparametric regression with Isotonic Adjustment (INIA)

2 Isotonic adjustment: For each patient $\mathbf{x}_{ij}$, obtain its predicted outcomes at all doses $\hat{g}(\mathbf{x}_{ij}, d_k)$ from step 1, $k = 1, ..., T$. The isotonic adjustment is then applied to $\hat{g}(\mathbf{x}_{ij}, d_k)$'s at $\mathbf{x}_{ij}$ assuming either increasing or decreasing dose-response by optimizing the following

$$\begin{aligned} &\min_{a_1,\ldots,a_T} \textstyle\sum_{k=1}^{T} \left[a_k - \hat{g}(\mathbf{x}_{ij}, d_k)\right]^2, \\ &s.t. \quad a_1 \leq a_2 \leq \cdots \leq a_T \end{aligned} \tag{7}$$

or

$$\begin{aligned} &\min_{a_1,\ldots,a_T} \textstyle\sum_{k=1}^{T} \left[a_k - \hat{g}(\mathbf{x}_{ij}, d_k)\right]^2. \\ &s.t. \quad a_1 \geq a_2 \geq \cdots \geq a_T \end{aligned} \tag{8}$$

The Pool-Adjacent-Violators algorithm [3] is used to obtain the solutions.

Iterative dose-dependent Nonparametric regression with Isotonic Adjustment (INIA)

3 Data augmentation: we compare the residual sum of squares from the two models (7) and (8) and choose the model with smaller errors. Denote the solution from the chosen model as $\hat{a}_k, k = 1, ..., T$. The predicted value at $\mathbf{x}_{ij}$ and dose d_k is then:

$$\hat{y}_{ij}^{(k)} = \hat{a}_k. \tag{9}$$

Now we obtain the augmented data for $\mathbf{x}_{ij}$ such that clinical outcome is available for all doses.

Iterative dose-dependent Nonparametric regression with Isotonic Adjustment (INIA)

4 Refitting the augmented data: Update the estimated mean response function $\hat{g}(\mathbf{x}, d_k)$ by fitting the augmented data. For binary outcomes, augmented data is the predicted probabilities and our refitting procedure is similar to the quasi maximum likelihood estimate for fractional response data[8].

5 Final model: iterate between step 2-4 until it converges.

Iterative dose-dependent Nonparametric regression with Isotonic Adjustment (INIA)

4 Refitting the augmented data: Update the estimated mean response function $\hat{g}(\mathbf{x}, d_k)$ by fitting the augmented data. For binary outcomes, augmented data is the predicted probabilities and our refitting procedure is similar to the quasi maximum likelihood estimate for fractional response data[8].

5 Final model: iterate between step 2-4 until it converges.

Estimated PED and the DDO-subgroup

- Estimated PED:

$$\hat{D}^*(x, \delta) = \min\{d_i : \hat{g}(\mathbf{x}, d_i) - \hat{g}(\mathbf{x}, d_i) \geq \delta\}.$$

- Estimated DDO-subgroup:

$$\hat{\mathcal{S}}(d_i, \delta) = \{\mathbf{x} : \hat{g}(\mathbf{x}, d_i) - \hat{g}(\mathbf{x}, d_i) \geq \delta\}.$$

Confidence intervals for the mean response functions and confidence regions for DDO-subgroups are constructed using bootstrapping.

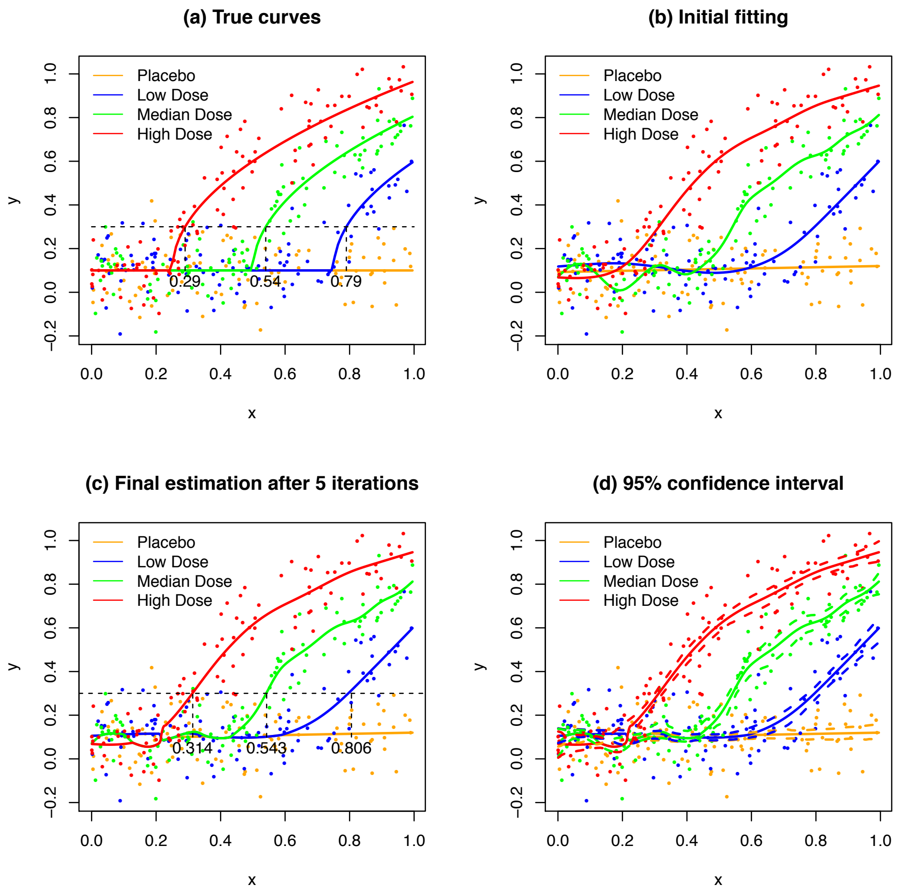
Single marker examples

- The first example is simulated from the mean response function:

$$g(x,d) = \begin{cases} 0.1, & \text{if } x \leq \gamma(d) \\ \sqrt{x-\gamma(d)} + 0.1, & \text{if } \gamma(d) < x \end{cases}$$

- x is uniformly distributed over $[0,1]$
- $\gamma(d) = 0.75, 0.5$ and 0.25 for the low, median and high dose group respectively.
- Sample size is 100 for each dose group.
- A thousand simulations are performed for both continuous and binary outcomes.
- In each simulation, the efficacy margin δ is uniformly drawn from the interval $[0.1, 0.4]$.

Single marker examples: continuous outcome

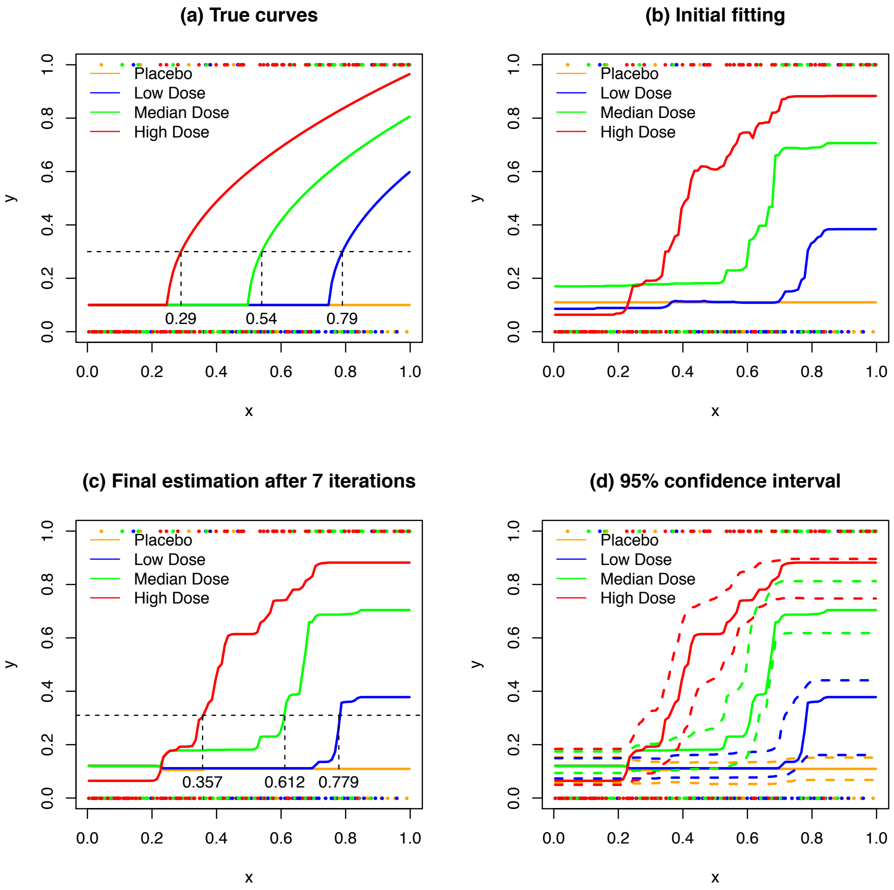


Single marker examples: continuous outcome

		Only-High			Group-All			Our Method		
	Stat	Low	Median	High	Low	Median	High	Low	Median	High
PED	SEN	0	0	0.95				0.91	0.94	0.93
	SPE	1	1	0.42				0.99	0.97	0.98
	PPV	0	0	0.36				0.97	0.93	0.94
	NPV	0.82	0.75	0.97				0.98	0.98	0.98
	MR	0.18	0.25	0.45				0.02	0.04	0.03
Subgroup	SEN			0.98	1	0.95	0.66	0.91	0.98	0.99
	SPE			0.99	0.66	0.92	1	0.99	0.98	0.97
	PPV			0.99	0.4	0.93	1	0.97	0.98	0.99
	NPV			0.97	1	0.97	0.6	0.98	0.99	0.98
	MR			0.02	0.27	0.059	0.23	0.02	0.02	0.02

Table: Comparison of the PED and DDO-subgroup estimates for the single marker example with continuous outcomes. Note that PED can't be estimated by the "Group-All" method and DDO-subgroup can't be estimated by the "Only-High" method.

Single marker examples: binary outcome



(a) True curves

(b) Initial fitting

(c) Final estimation after 7 iterations

(d) 95% confidence interval

Single marker examples: binary outcome

		Only-High			Group-All			Our Method		
	Stat	Low	Median	High	Low	Median	High	Low	Median	High
PED	SEN	0	0	0.82				0.92	0.71	0.78
	SPE	1	1	0.40				0.96	0.95	0.92
	PPV	0	0	0.31				0.86	0.88	0.79
	NPV	0.82	0.75	0.89				0.98	0.91	0.93
	MR	0.18	0.25	0.49				0.05	0.11	0.12
Subgroup	SEN			0.93	1	0.91	0.63	0.92	0.90	0.95
	SPE			0.94	0.68	0.92	0.99	0.95	0.96	0.93
	PPV			0.98	0.43	0.93	1	0.86	0.96	0.97
	NPV			0.89	1	0.95	0.59	0.98	0.94	0.92
	MR			0.07	0.26	0.077	0.25	0.05	0.06	0.06

Table: Comparison of the PED and DDO-subgroup estimates for the single marker example with binary outcomes. Note that PED can't be estimated by the "Group-All" method and DDO-subgroup can't be estimated by the "Only-High" method.

Multi-marker examples

- We simulate a dose-ranging study with genotyping markers to emulate a real Phase-II trial.
- The continuous outcome is simulated to represent the outcome of the lung function test, and the binary outcome is simulated to represent the event of exacerbation.
- A hundred single nucleotide polymorphisms (SNPs) in a target region are simulated under the Hardy-Weinberg equilibrium with minor allele frequency ranging between 0.01 and 0.5.
- The first 10 SNPs are prognostic markers independent of the treatment and the next 10 SNPs are dose-dependent predictive markers.

Multi-marker examples

- Data are generated from the following mean response function:

$$g(x,d) = \frac{1}{c}\left(\sum_{i=1}^{10} x_i + d\sum_{i=11}^{20} x_i\right). \tag{10}$$

  where $x_i = 0, 1$ or 2 is the number of minor alleles for the ith SNP and d is the dose.
- We use d =0, 5, 10 and 20, for the placebo, low, median and high dose respectively, with the scaling parameter $c = 1$ for the continuous outcome and $c = 150$ for the binary outcome.
- In each simulation, we generate a training dataset and a test dataset, both having 400 samples (100 samples for each group).
- For each simulation, the efficacy margin δ is drawn uniformly from the interval $[1, 50]$ for the continuous outcome, and from the interval $[0.05, 0.5]$ for the binary outcome.

Multi-marker examples: continuous outcome

		Only-High			Group-All			Our Method		
	Stat	Low	Median	High	Low	Median	High	Low	Median	High
PED	SEN	0	0	0.79				0.70	0.47	0.81
	SPE	0.95	1	0.03				0.59	0.65	0.88
	PPV	0	0	0.13				0.54	0.42	0.33
	NPV	0.52	0.65	0.45				0.57	0.77	0.90
	MR	0.48	0.35	0.86				0.13	0.27	0.16
Subgroup	SEN			0.99	0.82	0.96	0.93	0.89	0.98	0.99
	SPE			0.27	0.096	0.19	0.31	0.59	0.49	0.35
	PPV			0.98	0.49	0.58	0.78	0.64	0.89	0.88
	NPV			0.39	0.53	0.40	0.16	0.57	0.48	0.39
	MR			0.03	0.44	0.13	0.077	0.13	0.09	0.03

Table: Comparison of the PED and DDO-subgroup estimates for the multi-marker example with continuous outcomes. Note that PED can't be estimated by the "Group-All" method and DDO-subgroup can't be estimated by the "Only-High" method.

Multi-marker examples: binary outcome

		Only-High			Group-All			Our Method		
	Stat	Low	Median	High	Low	Median	High	Low	Median	High
PED	SEN	0	0	0.00				0.26	0.21	0.18
	SPE	0.56	1	0.05				0.42	0.62	0.67
	PPV	0	0	0.00				0.35	0.20	0.12
	NPV	0.0042	1	0.25				0.49	0.80	0.80
	MR	0.95	0	0.95				0.65	0.38	0.33
Subgroup	SEN			0.95	0.43	0.58	0.82	0.46	0.64	0.97
	SPE			0.11	0.18	0.14	0.12	0.42	0.31	0.13
	PPV			0.99	0.65	0.75	0.76	0.75	0.72	0.99
	NPV			0.09	0.004	0.003	0.003	0.12	0.11	0.09
	MR			0.04	0.28	0.30	0.36	0.20	0.21	0.03

Table: Comparison of the PED and DDO-subgroup estimates for the multi-marker example with binary outcomes. Note that PED can't be estimated by the "Group-All" method and DDO-subgroup can't be estimated by the "Only-High" method.

Discussion

- Personalized treatment is very difficult task.
- Requirement: good design, good sample size, good assays and **good statistical methods**.
- Review of existing methods means a lot of work.
- Need to derive noval methods for practical issues.
- Industry-Academia collaboration is a shortcut.

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32.

[2] Leo Breiman. Bias, variance, and arcing classifiers. Technical report, 1996.

[3] Jan de Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009.

[4] Elise Dusseldorp and Iven Van Mechelen. Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in medicine*, 33(2):219–237, 2014.

[5] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

[6] Wei-Yin Loh, Xu He, and Michael Man. A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in medicine*, 34(11):1818–1833, 2015.

[7] Zheng W. Ma, X. and Y. Lu. Personalized effective dose selection in dose ranging studies. *Statistical Applications from Clinical Trials and*

Personalized Medicine to Finance and Business Analytics, pages 91–104, 2016.

[8] Leslie E. Papke and Jeffrey M. Wooldridge. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11:619–632, 1996.

[9] Biran B Spear, Margo Heath-Chiozzi, and Jeffrey Huff. Clinical application of pharmacogenetics. *Trends in Molecular Medicine*, 7:201–204, 2001.

[10] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158, 2009.

[11] L. Tian, A. A. Alizadeh, A. J. Gentles, and R. Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109:1517–1532, 2014.

[12] Grace Wahba and P. Craven. Smoothing noisy data with spline functions. *Numerische Mathematik*, (31):337-403, 1979.

[13] Yuedong Wang. *Smoothing splines: methods and applications*. Monographs on Statistics and Applied Probability 121. CRC Press, Boca Raton, 2011.

[14] Y. Q. Zhao, D. Zeng, E. B. Laber, R. Song, M. Yuan, and M. R. Kosorok. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102:151–168, 2015.

[15] Y. Q. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107:1106–1118, 2012.

Major Challenges

- Most tree-based methods do not have decision making criteria to label subgroups as “effective” or “ineffective” in terms of treatment effect.
- It is not straightforward to apply the conventional type I error and type II error concepts in tree models.
- Tree structures may be equivalent even if their splitting processes are not identical.
- How to properly compare non-nested trees.

Definition of Type I Error Rate (TIE)

- Get the p-values p_t for the two-sided t-test of $H_0 : \beta_{t1} = 0$ in model $y = \beta_{t0} + \beta_{t1} trt + \eta_t$, which is fitted at each terminal node $t = 1, \ldots, N_T$, where N_T is the total number of terminal nodes.
- In the s-th simulation repetition, define TIE_s as

$$TIE_s = \begin{cases} 1, & \text{If } N_T \geq 2 \text{ and } \exists t,\ s.t.\ p_t < \alpha; \\ 0, & \text{Otherwise.} \end{cases} \tag{11}$$

- Define $TIE = \frac{\sum TIE_s}{\#\text{ of simulations}}$.
- $\alpha = 0.05$ is chosen in our simulations.

Procedure to Obtain ROC Curve

- Define G_1 as the subgroup with non-negative treatment effect, i.e. for which the treatment is beneficial, and G_0 is the rest.
- Pre-assign a series of d values, e.g. $\mathcal{D} = \{0, 0.05, 0.1, \dots, 4\}$. A proper choice of the range of d can be decided by pilot studies.
- Fit $y = \beta_{t0} + \beta_{t1} trt + \eta_t$ to each node t. Assign $\hat{d} = \hat{\beta}_{t1}$ to all patients from that node.
- Classify a patient into $\hat{G}_{1j}$ if $\hat{d}$ is no less than d_j for each $d_j \in \mathcal{D}$.
- Sensitivity (SS): $\frac{\#\{\hat{G}_1 \bigcap G_1\}}{\# G_1}$, and specificity (SP): $\frac{\#\{\hat{G}_0 \bigcap G_0\}}{\# G_0}$.
- Average (SS_{js}, SP_{js}) over all simulations.

Definition of T-AIC/T-BIC Measure

- Denote Z as the $n \times N_T$ terminal node assignment matrix, where N_T is the total number of terminal nodes of a tree.
- Fit linear regression $b(X) \sim Z$, obtain its residual sum of squares RSS.
- Define T-AIC as $T\text{-}AIC = n \log[max(RSS, dt)] + 2N_T$.
- Define T-BIC as $T\text{-}BIC = n \log[max(RSS, dt)] + \log(n)N_T$.
- dt is the differential threshold parameter, which prespecifies the “desired” minimum difference of “goodness of fit”.

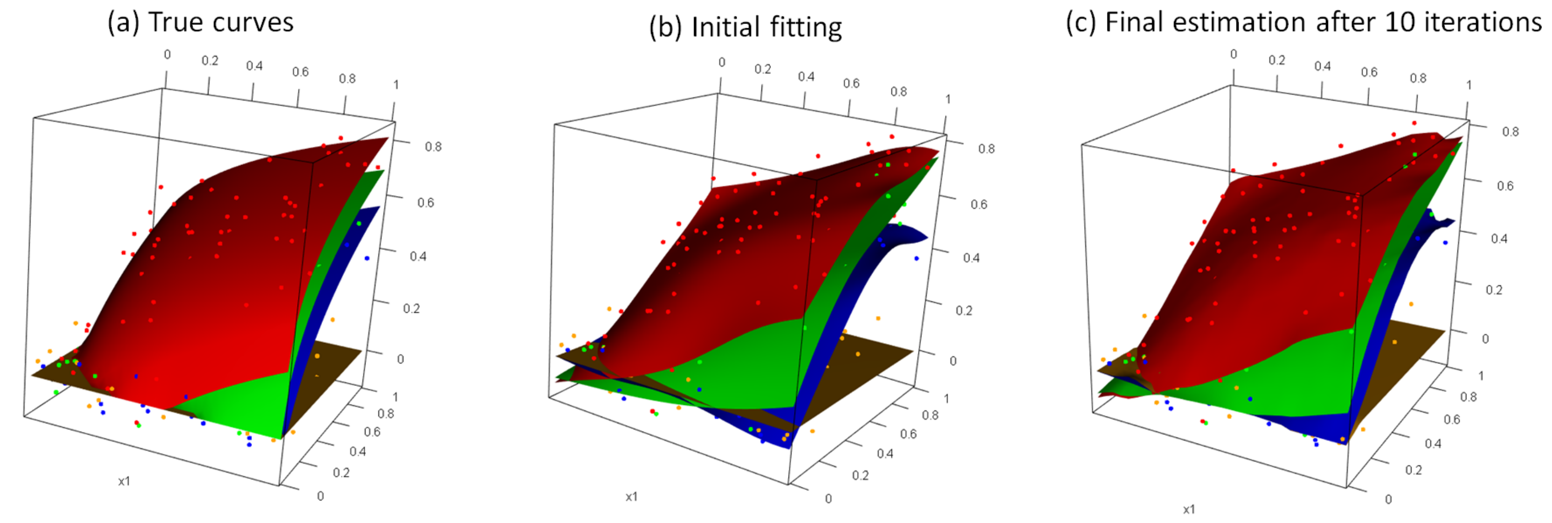
A two-marker example

- We simulate data from the following mean response function:

$$g(x, d) = \begin{cases} 0, & \text{if } \sqrt{x_1} + \sqrt{x_2} \leq \gamma(d) \\ \log(\sqrt{x_1} + \sqrt{x_2} - \gamma(d) + 1), & \text{if } \sqrt{x_1} + \sqrt{x_2} > \gamma(d) \end{cases}$$

- x_1 and x_2 are independently drawn from Uniform$[0, 1]$
- the cutoff point $\gamma(d) = 2, 1.25, 1, 0.75$ is for the placebo, low, median and high dose respectively.
- Sample size is 100 for each dose group.
- A thousand simulations are performed for both continuous and binary outcomes.
- In each simulation, the efficacy margin δ is uniformly drawn from the interval $[0.1, 0.4]$.

A two-marker example


(a) True curves

(b) Initial fitting

(c) Final estimation after 10 iterations

A two-marker example

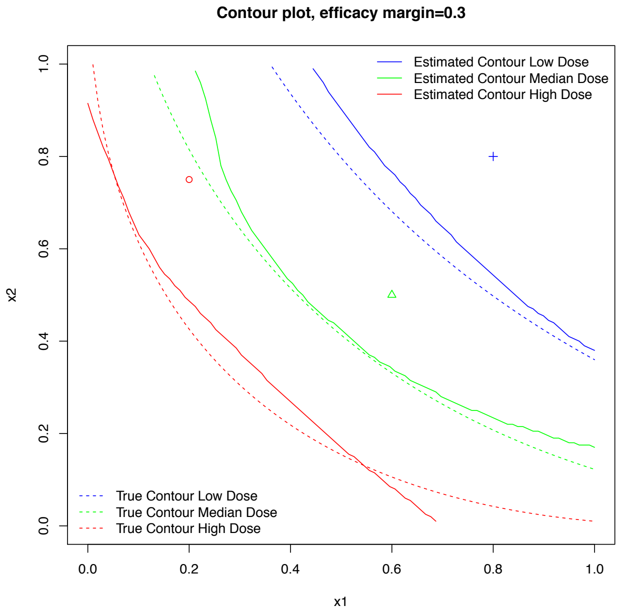


Figure: An example with two-marker simulation with continuous outcomes for estimating DDO-subgroups. The blue, green and red solid line defines the boundary of the true DDO-subgroup for the low, median and high dose respectively with the efficacy margin $\delta = 0.3$; the dotted lines are their estimated counterparts.

A two-marker example

		Only-High			Group-All			Our Method		
	Stat	Low	Median	High	Low	Median	High	Low	Median	High
PED	SEN	0	0	0.90				0.90	0.87	0.86
	SPE	1	1	0.27				0.99	0.96	0.94
	PPV	0	0	0.27				0.98	0.88	0.82
	NPV	0.7	0.74	0.91				0.97	0.95	0.96
	MR	0.3	0.26	0.59				0.03	0.07	0.08
Subgroup	SEN			0.97	1	0.94	0.69	0.90	0.94	0.97
	SPE			0.93	0.63	0.94	1	0.99	0.97	0.91
	PPV			0.98	0.53	0.97	1	0.98	0.98	0.98
	NPV			0.91	1	0.94	0.48	0.97	0.94	0.91
	MR			0.04	0.25	0.049	0.24	0.03	0.04	0.04

Table: Comparison of the PED and DDO-subgroup estimates for the two-marker example with continuous outcomes. Note that PED can't be estimated by the "Group-All" method and DDO-subgroup can't be estimated by the "Only-High" method.