Precision Medicine: Lecture 16
Value Function Inference

Michael R. Kosorok,
Nikki L. B. Freeman, Owen E. Leete, and Xiaotong Jiang

Department of Biostatistics
Gillings School of Global Public Health
University of North Carolina at Chapel Hill

Fall, 2021

Outline

Generalization Error for Decision Problems

Estimating Value Function Using Jackknife

Introduction

- The generalization error of a predictive model is a measure of its performance when applied to a population of interest
- In classification problems, a commonly used measure of generalization error can be expressed as a weighted expected number of mistakes
- In this context, the term ‘generalization error’ may refer to one of several functionals
  - The performance of the optimal predictive model within a pre-specified class
  - The conditional expected performance of a fitted model given the observed data
  - Unconditional expected performance of a fitted model averaged over the observed data

Notation and Setup

- We assume that the observed data are $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$
  - The covariate vector $\mathbf{X}$, takes values in $\mathbb{R}^p$
  - The label, Y, is binary and coded to take values in $\{-1, 1\}$
- A classification rule is map, $c : \text{dom}\mathbf{X} \mapsto \text{dom}Y$ such that the predicted label at input $\mathbf{X} = \mathbf{x}$ is $c(\mathbf{x})$
- Let P denote joint distribution of $(\mathbf{X}, Y)$
- The generalization error of a rule c is

$$M(c) \equiv P\mathbb{1}\{Y \neq c(\mathbf{X})\} = P\mathbb{1}\{Yc(\mathbf{X}) < 0\}$$

- $M(c)$ captures the probability that c will incorrectly label a random pair $(\mathbf{X}, Y) \sim P$

Fixed Classification Rule

- Let $\mathbb{P}_n$ denote the empirical distribution
- The plug-in estimator of $M(c)$ is $\widehat{M}_n(c) \equiv \mathbb{P}_n \mathbb{1}\{Yc(\mathbf{X}) < 0\}$
- For a fixed classification rule, c, it follows from the central limit theorem that

$$\sqrt{n}\left\{\widehat{M}_n(c) - M(c)\right\} \rightsquigarrow N[0, M(c)\{1 - M(c)\}]$$

- Estimation of and inference for the generalization error of a fixed classification rule is straightforward
- When the classification rule depends on the problem domain, quantifying uncertainty about the generalization error is considerably more complex

$M(\widehat{c}_n)$

- Let Ω denote the space of distributions over $\mathrm{dom}(\mathbf{X} \times Y)$ and define a classification algorithm for the class $\mathcal{C}$ to be a map $\Gamma : \Omega \mapsto \mathcal{C}$
- Under Γ, the estimated classifier given the observed data $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ is $\widehat{c}_n = \Gamma(\mathbb{P}_n)$
- The generalization error of $\widehat{c}_n$ is $M(\widehat{c}_n)$
  - The misclassification rate associated with applying $\widehat{c}_n$ to a new input-label pair $(\mathbf{X}, Y)$ drawn from P
- This generalization error is useful for quantifying the value of applying the estimated classification rule, $\widehat{c}_n$, to make decisions in the domain of interest

$M(c^{\textbf{opt}})$

- Given a classification algorithm Γ, define $c^{\text{opt}} = \Gamma(P)$ to be the classification rule relative to Γ for the populations P

- c^{opt} need not be the optimal classifier over the space of all measurable maps from dom$\mathbf{X}$ into dom$Y$ (i.e. the Bayes classifier)

- Inference for $M(c^{\text{opt}})$ may be of interest in the context of evaluating a data-driven classification rule relative to some existing classifier c_o (e.g. test $H_0 : M(c_0) \leq M(c^{\text{opt}})$)

- $M(c^{\text{opt}})$ does not measure the quality of a classification rule estimated from a finite data-set which is often of greater interest

$M_n(\Gamma)$

- An alternative measure of performance is the expected generalization error of a learning algorithm defined as $M_n(\Gamma) = \mathbb{E}M(\widehat{c}_n)$
- $M_n(\Gamma)$ is the average performance of the classification algorithm Γ across i.i.d. samples of size n drawn from P, where the randomness is from $\widehat{c}_n$ as a function of the sample.
- The curve, $n \to M_n(\Gamma)$ (learning curve of Γ), is a measure of how efficiently the algorithm learns from data on average
- $M_n(\Gamma)$ is most useful as a means to compare algorithms in a given domain across a range of data set sizes

Asymptotic Behavior of the Generalization Error

- The three generalization errors $M(\widehat{c}_n)$, $M(c^{\text{opt}})$, and $M_n(\Gamma)$ represent three different performance metrics

- These three generalization errors need not converge to each other even as n diverges to ∞

- For illustration purposes, we consider linear classification rules

- Define $\widehat{\beta}_n = \text{argmin}_{\beta \in \mathbb{R}^p} \mathbb{P}_n (Y - \mathbf{X}^\mathsf{T}\beta)^2$ and $\widehat{c}_n(\mathbf{x}) = \text{sign}(\mathbf{x}^\mathsf{T}\widehat{\beta}_n)$

- Let $\beta^* = \text{argmin}_{\beta \in \mathbb{R}^p} P(Y - X^\mathsf{T}\beta)^2$ denote the population analog of $\widehat{\beta}_n$ and define $c^{\text{opt}}(\mathbf{x}) = \text{sign}(\mathbf{x}^\mathsf{T}\beta^*)$

Asymptotic Behavior of the Generalization Error

- Under the previous model, the generalizations are
  - $M(\widehat{c}_n) = P\mathbb{1}\{Y\mathbf{X}^\mathsf{T}\widehat{\beta}_n < 0\} = \int \mathbb{1}\{y\mathbf{x}^\mathsf{T}\widehat{\beta}_n < 0\}dP(\mathbf{x}, y)$
  - $M(c^{\text{opt}}) = P\mathbb{1}\{Y\mathbf{X}^\mathsf{T}\beta^* < 0\}$
  - $M_n(\Gamma) = \mathbb{E}M(\widehat{c}_n)$
- Under mild moment conditions, $\sqrt{n}(\widehat{\beta}_n - \beta^*) \rightsquigarrow N\{0, \Sigma(\beta^*)\}$
- Thus, it follows that

$$\begin{aligned} M(\widehat{c}_n) &= P\mathbb{1}\left\{Y\mathbf{X}^\mathsf{T}\widehat{\beta}_n < 0\right\} \\ &= P\mathbb{1}\left\{Y\mathbf{X}^\mathsf{T}\sqrt{n}(\widehat{\beta}_n - \beta^*) < 0\right\}\mathbb{1}\left\{\mathbf{X}^\mathsf{T}\beta^* = 0\right\} \\ &\quad + P\mathbb{1}\left\{Y\mathbf{X}^\mathsf{T}\beta^* < 0\right\}\mathbb{1}\{\mathbf{X}^\mathsf{T}\beta^* \neq 0\} + o_P(1) \\ &\rightsquigarrow P\mathbb{1}\left\{Y\mathbf{X}^\mathsf{T}\mathbb{Z} < 0\right\}\mathbb{1}\left\{\mathbf{X}^\mathsf{T}\beta^* = 0\right\} + P\mathbb{1}\left\{Y\mathbf{X}^\mathsf{T}\beta^* < 0\right\} \end{aligned}$$

where $\mathbb{Z} \sim N\{0, \Sigma(\beta^*)$

Differences in Asymptotic Behavior

- Using the previous expressions, it follows that

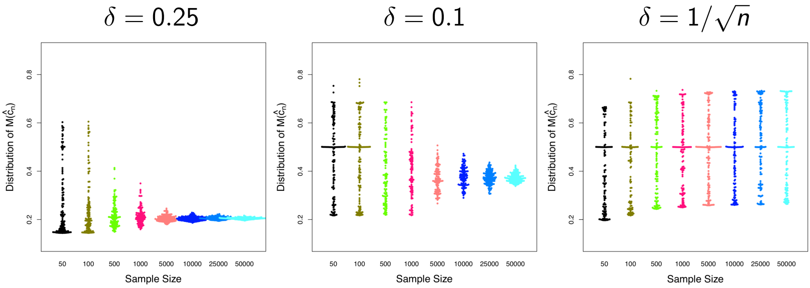
  $$M_n(\Gamma) = \mathbb{E}M(\widehat{c}_n) \to P\mathbb{1}\{\mathbf{X}^\mathsf{T}\beta^* = 0\}/2 + P\mathbb{1}\{Y\mathbf{X}^\mathsf{T}\beta^* < 0\}$$

  as $n \to \infty$
- Thus, the three types of generalization error need not coincide even asymptotically.
- The differences depend on the amassing of data points on the boundary $\mathbf{x}^\mathsf{T}\beta^* = 0$
  - if $P(\mathbf{X}^\mathsf{T}\beta^* = 0) = 0$ then the three definitions converge to the same limit
- Standard asymptotic approximations can perform poorly in finite samples

Simulation Studies

- To illustrate the differences, consider the following generative model
  - $Y \sim \text{Uniform}\{-1, 1\}$
  - $U \sim \text{Uniform}[0, 1]$
  - $X|Y = 1 \sim N(2 - 4 \times 1\{U \leq 1/2 - \delta\}, \sigma^2)$,
  - $X|Y = -1 \sim N(0, 0.5^2)$
- Consider linear decision rules of the form $c(x) = \text{sign}(\beta_0 + \beta_1 x)$ indexed by $\beta = (\beta_0, \beta_1)^\mathsf{T}$
  - If $\delta = 0$ then $\beta^* = \text{argmin}_\beta P(Y - \beta_0 - \beta_1 X)^2$ is identically zero
  - Otherwise, β^* is nonzero
- This generative model illustrates that $M(\widehat{c}_n)$ can be unstable even in large samples due to the presence of the non-smooth indicator function

Simulation Study Results



$\delta = 0.25$

$\delta = 0.1$

$\delta = 1/\sqrt{n}$

- Letting $\delta = O(1/\sqrt{n})$ retained the small-sample instability even as n diverged
- Moving-parameter asymptotic analysis is commonly used for nonregular quantities like $M(\widehat{c}_n)$ because they retain salient small sample behaviors even in infinite samples

Marginal Mean Outcome in Decision Making

- Consider a one-stage decision problem with observed data $\{\mathbf{X}_i, A_i, Y_i\}_{i=1}^{n}$
- The one-stage decision problem is closely related to the classification problem studied previously
- In a decision problem, the outcome Y provides indirect feedback about the quality of decision A in context X
- The marginal mean outcome under a given decision rule d can be expressed as

$$V(d) = P\left[\frac{Y\mathbb{1}\{d(\mathbf{X}) = A\}}{P(A|\mathbf{X})}\right] = P\left[\frac{Y\mathbb{1}\{d(\mathbf{X}) \neq -A\}}{P(A|\mathbf{X})}\right],$$

which can be viewed as a weighted misclassification rate for d with labels $-A$ and weights $Y/P(A|\mathbf{X})$

Value Function Inference

- The expression for $V(d)$ can be used to directly extend the inferential methods for the generalization error in classification to the marginal mean outcome in the single-stage decision setting
- The three generalization errors $M(\widehat{c}_n)$, $M(c^{\text{opt}})$, and $M_n(\Gamma)$ have analogous quantities in the single-stage decision setting
- Valid finite sample and asymptotic inference about the value function needs to account for the different behavior of these quantities

Outline

Generalization Error for Decision Problems

Estimating Value Function Using Jackknife

The Setup

- Assume we have n i.i.d copies of triplets $\{\mathbf{X}_i, A_i, Y_i\}$, where
    - $\mathbf{X}_i \in \mathcal{X} \subseteq \mathcal{R}^p$ is a $p \times 1$ vector of covariates for patient i
    - $A_i \in \mathcal{A} = \{-1, 1\}$ is treatment label for patient i
    - $Y \in \mathcal{R}$ is the outcome/response variable for patient i
    - WLOG, higher values of Y represent more favorable clinical outcomes
    - $\mathbf{X}$ and A could be independent or dependent
- The ITR is defined as a function $d : \mathcal{X} \rightarrow \mathcal{A}$ with $a_i = d(\mathbf{x}_i)$
- The Jackknife is leave-one out cross validation, where the training fold leaves one individual out at a time as the testing set.
- The decision rule trained from all n but without the ith subject is denoted as $\hat{d}_n^{(-i)}$.

The Jackknife Estimator

- Recall

$$V_0(d) = E\left[\frac{Y1\{A = d(\mathbf{X})\}}{P(A|\mathbf{X})}\right]$$

$$\widehat{V}(d) = \frac{\sum_{i=1}^{n} y_i 1\{a_i = d(\mathbf{x}_i)/P(a_i|\mathbf{x}_i)}{\sum_{i=1}^{n} 1\{a_i = d(\mathbf{x}_i)/P(a_i|\mathbf{x}_i)}$$

- We propose the jackknife estimator of the value function:

$$\widehat{V}^{jk}\left(\hat{d}_n\right) = \frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} w_i}$$

with $u_i = y_i \frac{1\{a_i = \hat{d}_n^{(-i)}(\mathbf{x}_i)\}}{P(a_i|\mathbf{x}_i)}$ and $w_i = \frac{1\{a_i = \hat{d}_n^{(-i)}(\mathbf{x}_i)\}}{P(a_i|\mathbf{x}_i)}$

The Jackknife Estimator

- The estimated variance of $\widehat{V}^{jk}(\hat{d}_n)$ is

$$\widehat{\text{Var}}\left[\widehat{V}^{jk}\left(\hat{d}_n\right)\right] = \frac{1}{n(n-1)}\sum_{i=1}^{n} r_i^2,$$

  where $r_i = \frac{1}{\bar{w}_n}u_i - \frac{\bar{u}_n}{\bar{w}_n^2}w_i$ is a bias-corrected, influence function-inspired form of the value function

- r_i is derived from the difference between $\widehat{V}(d)$ and $V_0(d)$, which is asymptotically linear with the summation of influence functions

Derivation of R_i

Assume $Y = O_p(1)$ and $E[W] \in (\epsilon, 1-\epsilon)$ for $0 < \epsilon < 0.5$.

$$
\begin{aligned}
& \hat{V}(d) - V_0(d) \\
= \;& \frac{n^{-1}\sum_{i=1}^n U_i}{n^{-1}\sum_{i=1}^n W_i} - \frac{E[U]}{E[W]} \\
= \;& \frac{n^{-1}\sum_{i=1}^n U_i}{n^{-1}\sum_{i=1}^n W_i} - \frac{E[U]}{n^{-1}\sum_{i=1}^n W_i} + \frac{E[U]}{n^{-1}\sum_{i=1}^n W_i} - \frac{E[U]}{E[W]} \\
= \;& \frac{n^{-1}\sum_{i=1}^n (U_i - E[U])}{n^{-1}\sum_{i=1}^n W_i} - \frac{n^{-1}E[U]\cdot(\sum_{i=1}^n W_i - E[W])}{(n^{-1}\sum_{i=1}^n W_i)\,E[W]} \\
= \;& \frac{n^{-1}\sum_{i=1}^n (U_i - E[U])}{E[W]} - \frac{n^{-1}E[U]\cdot(\sum_{i=1}^n W_i - E[W])}{(E[W])^2} + o_P(1)
\end{aligned}
$$

Derivation of R_i

- According to Theorem 18.7 in Kosorok (2008)

$$\sqrt{n}(\hat{V}(d) - V_0(d)) = \sqrt{n}\sum_{i=1}^{n} \check{\psi}_i + o_p(1)$$

for a fixed d

- The influence function and its estimator are

$$\begin{aligned}
\check{\psi}_i &= \frac{(U_i - E[U])}{E[W]} - \frac{E[U](W_i - E[W])}{(E[W])^2} \\
&= \frac{1}{E[W]}U_i - \frac{E[U]}{(E[W])^2}W_i \\
\ddot{\psi}_i &= \frac{1}{\bar{W}}U_i - \frac{\bar{U}}{\bar{W}^2}W_i
\end{aligned}$$

Why Jackknife?

- Weak assumptions (i.i.d, unrestricted probability distribution)
- Approximately unbiased for the true prediction error
- Avoids randomly selecting folds and repeating the splitting
- Consistency

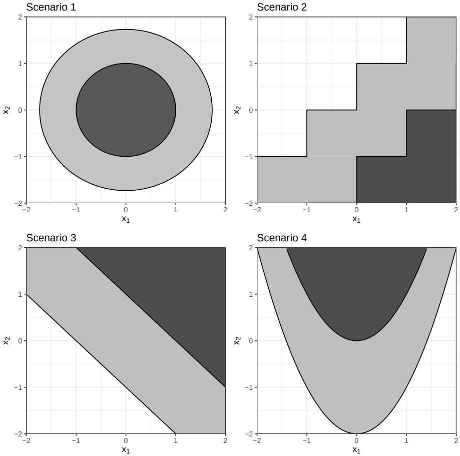
Theorem (Consistency)

Assume $E[P_{\mathbf{X}}(\hat{d}_n(\mathbf{X}) \neq \hat{d}_{n-1}(\mathbf{X}))] \to 0$ *and* $E\left[\frac{Y^2}{P(A|\mathbf{X})} + \frac{1}{P(A|\mathbf{X})}\right] < \infty$. *Then,*

$$\frac{\sum_{i=1}^{n} \frac{Y_i 1\{A_i = \hat{d}_n^{(-i)}(\mathbf{X}_i)\}}{P(A_i|\mathbf{X}_i)}}{\sum_{i=1}^{n} \frac{1\{A_i = \hat{d}_n^{(-i)}(\mathbf{X}_i)\}}{P(A_i|\mathbf{X}_i)}} - E[Y|A = \hat{d}_n(\mathbf{X})] \underset{p}{\to} 0$$

- Asymptotic normality was shown via simulations

Simulation Settings



- $n = 50, 100, 200, 400$
- $\mathbf{X} = \{X_1, X_2, X_3\} \sim U(-2, 2)$
- $A = \{0, 1, 2\} \sim \text{MN}\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$
- $Y \sim N(Q_0, 1)$ where $Q_0 = X_1 + X_2 + \delta_0(X_1, X_2, A)$
- True decision boundaries of 4 simulation scenarios (left)
- Number of simulations: 500

Simulation Results

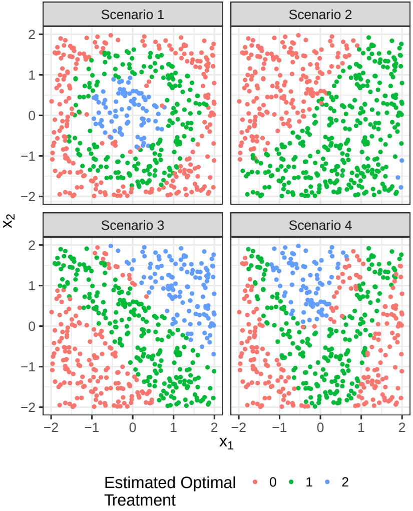


Figure 1: An example of jackknife estimated decision boundaries for a simulated dataset of size $n = 500$, trained by kernel ridge regression (KRR)

List of Models

Model	Parameters	#
Penalized regression	Lasso, $\alpha = 1$	1
	Ridge, $\alpha = 0$	2
	Elastic Net, $\alpha = 0.5$	3
Kernel ridge regression (KRR)	Gaussian kernel	4
Random forests (RF)	Rules based on each individual outcome Number of trees = 500	5
Reinforcement learning trees (RLT)	Number of trees = 50	6
List-based dynamic treatment regime (DTR)	Q-functions estimated by KRR Number of nodes = 2, 3, 5, 10	7,8, 9,10
	Q-functions estimated by RF Number of nodes = 2, 3, 5, 10	11,12, 13,14
	Q-functions estimated by Super Learning Number of nodes = 2, 3, 5, 10	15,16, 17,18
	Q-functions estimated by elastic net Number of nodes = 10	19
Residual weighted learning (RWL)	Linear kernel	20
	Polynomial kernel with 2nd order	21
	Polynomial kernel with 3rd order	22
Bayesian Additive Regression Tree (BART)	Number of trees = 500 Number of draws = 5500 (including 500 burnins)	23
Zero-order (ZOM)	Always assign to 0	24
	Always assign to 1	25
	Always assign to 2	26

Comparing PMMs and ZOMs

- ZOM is the **zero-order model** where all subjects receive the same single treatment
- PMM is the **precision medicine model** where subjects receive individualized treatment regimes
- We fit 3 ZOMs and 23 PMMs and select the optimal ZOM and PMM
- The **optimal model** is the one model from all candidate models that has the highest estimated value function while taking into account their standard errors

Asymptotic Normality

- The test statistic for jackknife estimators is

$$T_0^{sim} = \frac{[\hat{V}^{jk}(\hat{d}_{\text{PMM}}) - \hat{V}^{jk}(\hat{d}_{\text{ZOM}})] - [V_0(\hat{d}_{\text{PMM}}) - V_0(\hat{d}_{\text{ZOM}})]}{\sqrt{\frac{\sum_{i=1}^{n}(R_{\text{PMM},i} - R_{\text{ZOM},i})^2}{n(n-1)}}}$$

$$H_0 : V_0(\hat{d}_{PMM}) - V_0(\hat{d}_{ZOM}) = 0$$

- We tested the normality of T_0^{sim} using the Shapiro-Wilk test (p-values below)

Sample Size	50	100	200	400
Scenario 1	0.20	0.37	0.15	0.18
Scenario 2	0.85	0.92	0.41	0.79
Scenario 3	<0.01	0.99	0.13	0.61
Scenario 4	<0.01	0.67	0.81	0.84

Asymptotic Normality

Only two sample sizes 50 and 400 were shown for cleaner plots.

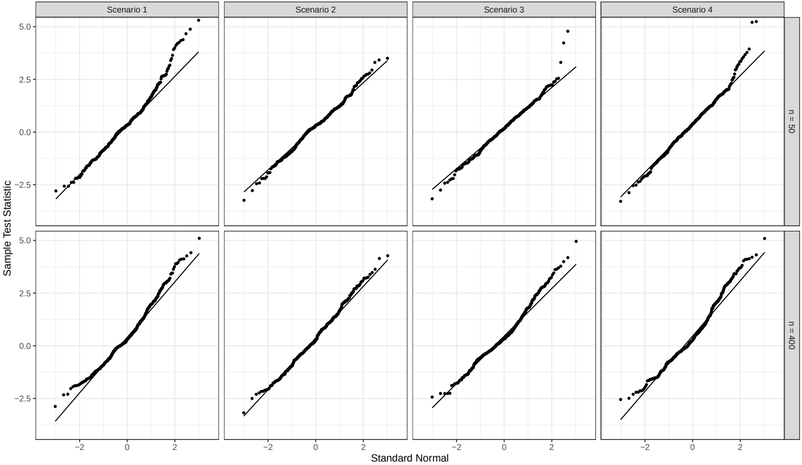


Figure 2: Q-Q plots of the distribution of jackknife T_0^{sim} across 100 simulations versus the standard normal distribution

Clinical Application

- Knee osteoarthritis (OA)
  - One of the most common forms of arthritis worldwide
  - Where the cartilage in the knee joint wears away
  - Symptoms: pain, stiffness, limited range of motion, etc
- IDEA (Intensive Diet and Exercise for Arthritis)
  - A 18-month single-center randomized clinical trial
  - 454 overweight and obese adults with symptomatic knee OA
  - 3 interventions: exercise E, diet D, diet and exercise D+E
- Goal of IDEA
  - To determine whether reduction in body weight induced by D or D+E would improve mechanistic and clinical outcomes more than E alone.

Clinical Application

- Our Goals
  - To confirm the results found by the original IDEA trial $\Longrightarrow$ that there may be subgroups who would achieve more benefits from a specific intervention.
  - To come up with a simple, data-driven, precision medicine-based treatment recommendation for clinical practice in knee OA
- Total outcomes: 7
  - weight loss, pain/function/stiffness scores, PCS, IL-6, force
- Preprocessed input data 1
  - $n = 343$, $p = 15$ baseline covariates (dimension reduction applied)

Pipeline

- Identify candidate ZOMs and candidate PMMs that suit the data
- Fit the models on each outcome and select the optimal ZOM and optimal PMM
- Compare the optimal ZOM and PMM with a two-sample Z-test

$$H_o : V_0(\hat{d}_{\text{PMM}}) = V_0(\hat{d}_{\text{ZOM}})$$

$$T^{jk}(\hat{d}_{\text{PMM}}, \hat{d}_{\text{ZOM}}) = \frac{\widehat{V}^{jk}(\hat{d}_{\text{PMM}}) - \widehat{V}^{jk}(\hat{d}_{\text{ZOM}})}{\sqrt{\frac{\sum_{i=1}^{n}(R^{jk}_{PMM} - R^{jk}_{ZOM})^2}{n(n-1)}}}$$

- For outcomes with significant results, estimate the decision rule from the optimal PMM
- This decision rule is the data-driven treatment recommendation for clinicians to consider

Results

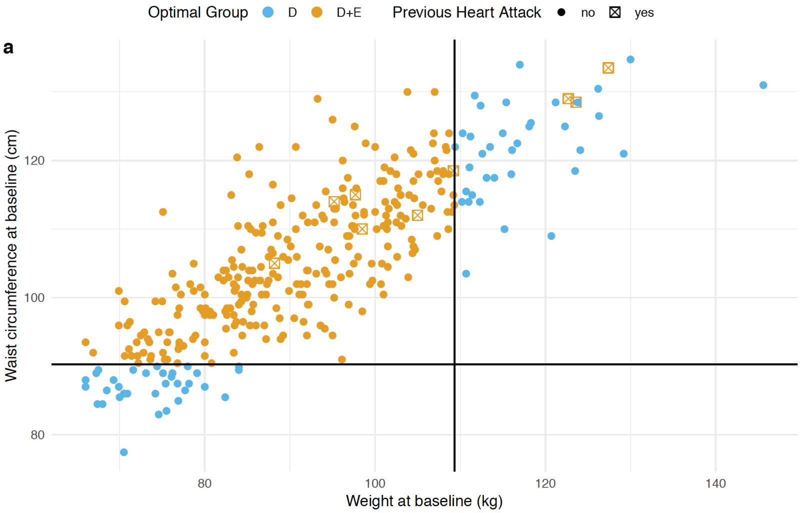


Figure 3: Visualization of the estimated ITRs for the outcome: weight loss at 18th month since baseline

Multiple Outcomes

- To take into account potential correlations among outcomes, we took a **weighted sum** across three outcomes (weight loss, pain, function)

- The weights were derived from a **minimax** algorithm
    - Apply RF models to look for the weight combination that maximized the worst jackknife value function estimates (as a percentiles within outcome) across the three outcomes
    - Use a coarse-to-fine grid search to reduce computation time
    - Create a compositie outcome using the selected minimax weights to train a RF model and estimate the optimal treatment rule $\Longrightarrow$ call this multiple-outcome PMM

- Note that we did not scale the outcomes, but allowed the weights to adjust for different scales in the outcomes.

Results

- For weight loss, we observed significant effect size of our multiple-outcome PMM vs. optimal ZOM (D+E), p=0.05
- For other outcomes (although not statistically significant), the multiple-outcome PMM
  - outperformed single-outcome PMMs for outcomes correlated to the three outcomes (e.g., pain/function/stiffness scores),
  - but did not outperform for outcomes uncorrelated to the three outcomes (e.g., compressive force and IL-6)