Subgroup Selection in Adaptive Signature Designs of Confirmatory Clinical Trials

Zhiwei Zhang

Department of Statistics
University of California, Riverside
zhiwei.zhang@ucr.edu

Joint work with Meijuan Li (FDA), Min Lin (FDA), Guoxing Soon (FDA), Tom Greene (U. Utah), and Changyu Shen (Harvard); no endorsement by FDA

Outline

- Introduction
- Formulation
- Subgroup selection
- Treatment effect estimation
- Examples
- Summary

Introduction

- Treatment effects are frequently heterogeneous.
- A clinically meaningful treatment benefit is often limited to a subpopulation of patients (e.g., Simon, 2008, 2010).
- If a promising subpopulation is known at the design stage, this knowledge can be used to
  - plan a subgroup analysis in a broad eligibility trial, or
  - restrict enrollment in a targeted trial.
- If unknown at the design stage, a target subpopulation can be developed in a data-driven manner
  - at the end of a broad eligibility trial (adaptive signature design; ASD), or
  - at an interim analysis (adaptive enrichment design).

- Our focus is on subgroup selection in ASDs (Freidlin and Simon, 2005; Freidlin, Jiang and Simon, 2010), and we also consider treatment effect estimation for the selected subgroup.
- Our objective is to maximize the power for detecting a positive treatment effect in the selected subgroup or, more generally, the expected gain based on a specified utility function.
- The latter formulation allows investigators to take into account the size of the selected subgroup as well as the clinical value of demonstrating treatment efficacy for the selected subgroup.
- Our objective is not to find an optimal treatment regime that maximizes the expected outcome for the entire population.

- Our approach is based on a simple and general characterization of the optimal subgroup.
  - For a binary outcome, this characterization takes the form of a halfspace in terms of covariate-specific response rates (in both treatment groups) and utility (if specified).
- Motivated by this characterization, we propose a subgroup selection procedure which consists of the following three steps:
  1. Estimate the covariate-specific response rate in each treatment group;
  2. Estimate the expected gain for each candidate halfspace defined by a vector of coefficients and the estimates from Step 1;
  3. Choose the halfspace with the largest estimate of the expected gain.
- A cross-validation approach can be used to estimate the treatment effect for the chosen subgroup.
- A bootstrap procedure can be used to make inference about the treatment effect for the chosen subgroup.

Formulation

- An ASD is just an "all-comer" clinical trial with a prospective plan for testing treatment efficacy for a data-driven subgroup of patients.
- Unlike other adaptive designs, an ASD does not (necessarily) involve an interim analysis; it is adaptive in the selection of patients for a possible subgroup analysis.
- Usually, an ASD includes a test of overall treatment efficacy (at level α_1) as well as a test of treatment efficacy for a data-driven subgroup (at level α_2), where α_1 and α_2 are chosen to control the familywise error rate.
- Here we restrict attention to ASDs without a test of overall treatment efficacy.

- Consider an "all-comer" clinical trial with randomized treatment T (1 experimental; 0 control), primary outcome Y, and baseline covariates X (for subgroup selection).
- The question is how to choose a subgroup $A \subset \mathcal{X}$ and test the associated hypothesis on the basis of $\{(X_i, T_i, Y_i) : i = 1, \ldots, n\}$.
- If the investigator is only concerned about power, then an optimal choice of A is a maximizer of $\text{pow}(A)$.
- There may, however, be additional considerations concerning the size and content of A:
  - It is more desirable to demonstrate treatment efficacy for a large subpopulation (say 90%) than for a small one (say 10%).
  - Successful demonstration of treatment efficacy is more important for a subgroup of patients with no alternative treatments than for a subgroup with many treatment options.

- Such considerations can be accommodated using a utility function $u : \mathcal{X} \to [0, \infty)$.
  - If a subset A tests significant, the realized gain is $u(A) = \int_A u(x) \mathrm{d}F(x)$, where F is the distribution of X.
  - For a specified utility function, an optimal choice of A is a maximizer of the expected gain $\gamma(A) = u(A) \operatorname{pow}(A)$.
- Because the exact power may be difficult to calculate, we work with an approximation based on asymptotic normality:

$$g(A) = u(A)\widetilde{\operatorname{pow}}(A).$$

- We will attempt to find a subset A that maximizes $g(A)$ and estimate the treatment effect in the chosen subpopulation, which could be the entire population.
- Since $u(A)$ is already taken into account, there is no need for a separate test of overall treatment efficacy.

Subgroup Selection

- To fix ideas, suppose Y is binary (1 success; 0 failure), and write

$$p_t(x) = \mathrm{P}(Y = 1 | T = t, X = x) \qquad t = 0, 1.$$

- We assume that X has at least one continuous component and that the functions (u, p_0, p_1) are continuous in the continuous components of X.
- Under appropriate conditions, a subset A_{opt} that maximizes $g(A)$ consists of $x \in \mathcal{X}$ such that

$$(1, u(x), p_0(x), p_1(x))c(A_{\text{opt}}) \leq 0,$$

  where $c(A_{\text{opt}})$ is a 4-vector that depends on A_{opt} but not on x.
- If the utility function is constant or if there is no utility function (so the objective function is simply $\widetilde{\text{pow}}(A)$), the same characterization of A_{opt} applies after removing $u(x)$ from the above expression.

- Motivated by this characterization, we consider the class of subsets $\mathcal{A} = \{A(c) : \|c\| = 1\}$, where

$$A(c) = \{x \in \mathcal{X} : (1, u(x), \widehat{p}_0(x), \widehat{p}_1(x))c \leq 0\},$$

  $\widehat{p}_t(x)$ is an estimate of $p_t(x)$ based on a model for $\mathrm{P}(Y = 1|X, T)$, and the unit norm constraint on c is for uniqueness.
- This approach is insensitive to any collinearity in X.
- For a given c, an estimate of $g(A(c))$ can be obtained by substituting estimates of $(F(A), u(A), p_0(A), p_1(A))$, where $A = A(c)$ is considered fixed and $p_t(A) = \mathrm{P}(Y = 1|T = t, X \in A)$, $t = 0, 1$.
  - $\widehat{F}(A) = \frac{1}{n}\sum_{i=1}^n I(X_i \in A)$, $\widehat{u}(A) = \frac{1}{n}\sum_{i=1}^n I(X_i \in A)u(X_i)$
  - $\widehat{p}_t^{\mathrm{emp}}(A) = \frac{\sum_{i=1}^n I(X_i \in A, T_i = t, Y_i = 1)}{\sum_{i=1}^n I(X_i \in A, T_i = t)}$
  - $\widehat{p}_t^{\mathrm{aug}}(A) = \widehat{p}_t^{\mathrm{emp}}(A) - \frac{\sum_{i=1}^n I(X_i \in A)\{I(T_i = t) - \widehat{\varpi}_t(A)\}\widehat{p}_t(X_i)}{\sum_{i=1}^n I(X_i \in A, T_i = t)}$, where $\widehat{\varpi}_t(A)$ is the proportion of subjects in A that receive treatment t.

- Our proposal is to estimate A_{opt} by $\widehat{A}_{\text{opt}} = A(\widehat{c}_{\text{opt}})$, where $\widehat{c}_{\text{opt}}$ maximizes the estimate of $g(A(c))$ over the unit sphere.
- This is a low-dimensional maximization problem, which can be solved using standard techniques (e.g., grid search).
- If $\widehat{p}_t(x)$ estimates $p_t(x)$ consistently, then we can expect $\widehat{c}_{\text{opt}}$ and $\widehat{A}_{\text{opt}}$ to approach $c(A_{\text{opt}})$ and A_{opt} respectively in large samples.
- If $\widehat{p}_t(x)$ is inconsistent for $p_t(x)$ (e.g., due to model misspecification), then $\widehat{A}_{\text{opt}}$ estimates a local optimum (within the class $\mathcal{A}$).
- A severe departure of $\widehat{A}_{\text{opt}}$ from A_{opt} could be detected by comparing $\widehat{c}_{\text{opt}}$ with $c(\widehat{A}_{\text{opt}})$.

Treatment Effect Estimation

- Under appropriate conditions, $\widehat{p}_t(\widehat{A}_{\text{opt}}) - p_t(\widehat{A}_{\text{opt}}) = o_p(1)$, and $\sqrt{n}\{\widehat{p}_t(\widehat{A}_{\text{opt}}) - p_t(\widehat{A}_{\text{opt}})\}$ is asymptotically normal.
- Thus, for asymptotic inference, one can largely ignore the fact that $\widehat{A}_{\text{opt}}$ and $\{\widehat{p}_t(A) : A \in \mathcal{A}\}$ are obtained from the same set of data.
- In finite samples, however, a selection bias can arise when the same sample is used to develop $\widehat{A}_{\text{opt}}$ and estimate the treatment effect in this subgroup.
- A cross-validation approach can be used to remove or reduce the selection bias.

- K-fold cross-validation:
  - Partition the study cohort randomly into a specified number, say K, of subsamples that are roughly equal in size.
  - For each $k \in \{1, \ldots, K\}$, we use the kth subsample as the validation sample and combine the other subsamples into a training sample.
  - From the training sample we obtain $\widehat{A}_{\text{opt}}^{(-k)} = \text{argmax}_{A \in \mathcal{A}^{(-k)}} \widehat{g}^{(-k)}(A)$ using the exact same method for obtaining $\widehat{A}_{\text{opt}}$.
  - Next, we apply $\widehat{A}_{\text{opt}}^{(-k)}$ to the validation sample and obtain $\widehat{p}_t^{(k)}(\widehat{A}_{\text{opt}}^{(-k)})$, where $\widehat{p}_t^{(k)}(\cdot)$ is based on the validation sample alone.
  - The final cross-validated estimator of $p_t(\widehat{A}_{\text{opt}})$ is given by

$$\widehat{p}_t^{\text{cv}}(\widehat{A}_{\text{opt}}) = \frac{1}{K} \sum_{k=1}^{K} \widehat{p}_t^{(k)}(\widehat{A}_{\text{opt}}^{(-k)}).$$

- Inference on $p_1(\widehat{A}_{\text{opt}}) - p_0(\widehat{A}_{\text{opt}})$ can be based on nonparametric bootstrap standard errors and confidence intervals.

Example 1 (MAGIC)

- The Magnesium in Coronaries (MAGIC) study is a randomized clinical trial that investigated, in high risk patients with ST-elevation myocardial infarction, the effect of supplemental administration of intravenous magnesium on short-term mortality.
- The MAGIC study enrolled 6213 patients, who were randomized $1:1$ to magnesium sulphate or matching placebo.
- The primary endpoint was all-cause mortality within 30 days of randomization.
- The observed mortality rate was 15.3% in the magnesium group and 15.2% in the placebo group, with an odds ratio of 1.0 (95% CI: 0.9–1.2).
- No benefit or harm of magnesium was observed in 8 pre-specified and 15 exploratory subgroup analyses.

- In our retrospective analysis, the baseline covariate vector consists of age, gender, systolic blood pressure, heart rate, a simple risk index, a modified TIMI score, and nine other covariates.
- Three subjects with missing covariate data are excluded from our analysis.
- We estimate $p_t(x)$ under a logistic regression model and estimate $p_t(A)$ with the augmented estimator $\widehat{p}_t^{\text{aug}}(A)$.
- Our analysis is based on superiority hypotheses ($H_0 : p_1(A) \leq p_0(A)$ vs $H_1 : p_1(A) > p_0(A)$), one-sided $\alpha = 0.025$, and a constant utility function, and involves grid search and 20-fold cross-validation.

Quantity of Interest	Pt. Est. (%)		Std. Error (%)	
	Naive	CV	Naive	CV
$F(\widehat{A}_{\text{opt}})$	63.1	63.8	13.8	12.1
$\widetilde{\text{pow}}(\widehat{A}_{\text{opt}})$	73.8	0.7	12.2	23.2
$g(\widehat{A}_{\text{opt}})$	46.5	0.5	14.9	20.1
$p_0(\widehat{A}_{\text{opt}})$	86.7	87.5	1.9	1.6
$p_1(\widehat{A}_{\text{opt}})$	89.4	86.9	1.8	1.3
$\delta_p(\widehat{A}_{\text{opt}})$	2.7	-0.5	0.6	0.9

Example 2 (HEMO)

- The HEMO study is a randomized clinical trial that evaluated the effects of the dose of dialysis and the level of flux of the dialyzer membrane on mortality and morbidity among patients undergoing maintenance hemodialysis.
- The HEMO study enrolled 1846 patients undergoing thrice-weekly dialysis and randomized them to a standard or high dose of dialysis $(1:1)$ and to a low-flux or high-flux dialyzer $(1:1)$ under a two-by-two factorial design.
- The primary endpoint was time to death from any cause, which was not significantly influenced by the dose or flux assignment:
  - the hazard ratio for high versus standard dose was estimated to be 0.96 (95% CI: 0.84–1.10; $p = 0.53$);
  - the hazard ratio for high versus low flux was estimated to be 0.92 (95% CI: 0.81–1.05; $p = 0.23$).
- However, possible interactions were identified between dose and sex (unadjusted $p = 0.01$) and between flux and prior years of dialysis (≤ 3.7 years versus > 3.7 years; unadjusted $p = 0.005$).

- The corresponding subgroup analyses suggested that women might benefit from a high dose of dialysis and that patients with longer history of dialysis might benefit from high flux.
- Although definitive answers to these questions would require pre-planned analyses, we present a retrospective analysis here mainly to illustrate the proposed methodology.
- The treatment of interest is the level of flux (1 high; 0 low), and the baseline covariate vector consists of the same seven covariates pre-specified for subgroup analyses and also included in the primary (Cox regression) analysis.
- We work with survival status (1 alive; 0 dead) at 3 years post-randomization as the outcome variable, and restrict attention to the 1414 subjects who were randomized at least 3 years prior to the administrative end of the study.
- Our analysis of this example is similar to the previous one except for the use of one-sided $\alpha = 0.05$ and 10-fold cross-validation.

Quantity of Interest	Pt. Est. (%)		Std. Error (%)	
	Naive	CV	Naive	CV
$F(\widehat{A}_{\text{opt}})$	76.2	79.8	12.5	10.3
$\widetilde{\text{pow}}(\widehat{A}_{\text{opt}})$	63.9	33.7	9.6	21.9
$g(\widehat{A}_{\text{opt}})$	48.7	26.9	13.6	21.1
$p_0(\widehat{A}_{\text{opt}})$	65.9	65.8	3.5	2.6
$p_1(\widehat{A}_{\text{opt}})$	71.6	69.2	3.4	2.4
$\delta_p(\widehat{A}_{\text{opt}})$	5.6	3.4	1.2	2.0

Summary

- This work provides new insights and methods for ASDs:
  - A simple characterization of the optimal subgroup
  - A three-step procedure for subgroup selection
  - A cross-validation procedure for treatment effect estimation
- Advantages of the proposed methodology:
  - No need to perform two separate tests and split alpha
  - Insensitivity to collinearity in X
  - Use of AIPW to incorporate covariate information and improve precision
- The main ideas generalize easily to other types of outcome variables.