# TWO-STAGE CLINICAL TRIAL DESIGNS WITH SURVIVAL OUTCOMES AND ADJUSTMENT FOR MISCLASSIFICATION IN PREDICTIVE BIOMARKERS.

**By**

**YANPING CHEN**

**A Dissertation submitted to the**

**School of Public Health**

**and the**

**Graduate School – New Brunswick**

**Rutgers, The State University of New Jersey**

**in partial fulfillment of the requirements**

**for the degree of**

**Doctor of Philosophy**

**Written under the direction of**

**Professors Yong Lin and Shou-En Lu**

**and Approved by**

_____

_____

_____

_____

**New Brunswick, New Jersey**

**May, 2020**

## ABSTRACT OF THE DISSERTATION

# Two-stage Clinical Trial Designs with Survival Outcomes and Adjustment for Misclassification in Predictive Biomarkers.

**by Yanping Chen**

**Dissertation Directors: Professors Yong Lin and Shou-En Lu**

Oncology indispensably leads us to personalized medicine, which allows an individual approach to be taken with each subject. Personalized oncology is based on pharmacogenomics and the effect of genetic differences in individuals. Biomarkers detected using molecular biology tools allow the molecular characterization of cancer signatures and provide information relevant for personalized treatment. The key to success of these targeted therapy is to identify a "predictive biomarker" and validate the "predictive biomarker" through randomized clinical trials. In this dissertation, we focus on biomarker based two-stage clinical trial designs with *survival outcomes*.

In Part I of this dissertation, we assume that there is no misclassification of biomarker and we design a two-stage adaptive enrichment clinical trial, based on a binary "predictive"biomarker. At the interim analysis, based on the statistics observed from the biomarker negative strata, a decision is made to either continue enrolling both biomarker positive and biomarker negative subjects or enrich the remaining number of subjects only to biomarker positive subjects.

In Part II, we address the issue of misclassification of biomarker which is common in determining the predictive biomarker status. A two-stage stratified study design is proposed and evaluated. We use the information obtained from both marker appeared-positive strata

and marker appeared-negative strata, to solve the adjusted log rank statistics for true marker positive and true marker negative group. No additional distributional assumption is needed for this stratified designs.

In Part III, we extend the biomarker misclassification adjustment method to the two-stage enrichment designs proposed in Part I. With some additional distributional assumption (exponential distribution assumption for survival times), we can use the information obtained from interim analysis, to help obtain the adjusted log rank statistics for the true marker positive group, even though the marker appeared-negative group was discontinued after interim analysis and no marker appeared-negative subjects are enrolled in Stage II.

Family-wise type I error control is achieved by considering correlation of log rank statistics from the same and/or different stages. R-code is developed to calculate critical values, to achieve specified global power, or specified marginal power, and to calculate sample size as well.

# Acknowledgements

I would like to thank my thesis advisors, Dr. Yong Lin and Dr. Shou-En Lu for their wonderful academic guidance, and for being patient and very understanding through my dissertation research. I also thank Dr. Weichung Joe Shih and Dr. Hui Quan, for their comments/suggestions while serving in my thesis committee.

# Dedication

To my wife Serina, my daughters Sunny and Elsa, and my parents,

for their love, patience, and support.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Cancer is one of the leading causes of death in the United States, second only to heart disease. The conventional cancer treatment has been chemotherapy. Chemotherapeutic drugs are designed to target all rapidly dividing cells, including cancer cells and certain normal cells. There has been growing interest in biomarker-driven personalized cancer therapy, also known as precision medicine, or targeted therapy. Like conventional chemotherapy, targeted cancer therapies use pharmacological agents that increase cell death and restrict the spread of cancer. Targeted therapy stops the action of molecules that are key to the growth of cancer cells. By acting on specific oncogenic proteins, rather than interfering with all rapidly dividing cells, these targeted therapies hold promise for improved patient outcomes.

There are two main types of targeted therapy: small molecule drugs and monoclonal antibodies. Small molecule drugs enter cells and monoclonal antibodies are too large to enter cells. Instead, monoclonal antibodies affect targets outside of cells or targets on cells' surface. They have diverse mechanisms of action.

One of the first breakthrough of molecular target biology was imatinib, used for the treatment of chronic myeloid leukemia (CML). Philadelphia chromosome, a unique characteristic of CML, is related to BCR-Abl tyrosine kinase overexpression, which does not occur in normal cells. Therefore, this selective BCR-Abl tyrosine kinase inhibitor, imatinib, could suppress the growth of Philadelphia chromosome-positive CML with less harm to normal cells.

In terms of targeted therapy, it is difficult to have a single therapy for all cancers, not even for a single type of cancer. Therefore, the concept of personalized medicine becomes relevant and points to the need to evaluate every patient according to his/her unique tumor

phenotype. Consequently, the next step of targeted cancer therapy is the identification of new specific targets. The identified target molecules will then be used for the identification of the specific sub-population of patients who have the receptor of the identified target molecule and therefore could benefit from the treatment. This is the major aim of "personalized medicine". The key to success of targeted therapy is to identify a "predictive biomarker" and validate the "predictive biomarker" through clinical trials.

In medicine, a biological marker, or "biomarker" is, in the broadest sense, anything that can be used as an indicator of a particular disease state or some other biological state of an organism. Classically, the term referred to a basic laboratory parameter used to help physicians diagnose a disease and select a course of treatment. For example, the detection of the carcinoembryonic antigen (CEA) in blood samples has been an important diagnostic, progression or recurrence marker especially for cancers of the gastrointestinal tract.

A variety of factors influence a patient's clinical outcome, including intrinsic characteristics of the patient, disease, or medical condition, and the effects of any treatments that the patient receives. Some of the intrinsic characteristics may be reflected as prognostic biomarkers, i.e., biomarkers used to identify likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest, and others as predictive biomarkers, i.e., biomarkers used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent. Prognostic biomarkers and predictive biomarkers cannot generally be distinguished when only patients who have received a particular therapy are studied. Some biomarkers are both prognostic and predictive. Prognostic biomarkers are often identified from observational data and are regularly used to identify patients more likely to have a particular outcome.

Even through there are circumstances in which preclinical and early clinical data provide such compelling evidence that a new treatment will not work in patients without the biomarker, predictive biomarkers can only be properly validated in a prospectively designed randomized controlled clinical trials: compare a treatment to a control in patients with and without the biomarker.

In the past decades, a number of biomarker-based design solution have been proposed

to study treatments within possibly heterogeneous patient subpopulations (ie, genomic biomarker is predictive of treatment effect, and may or may not be prognostic of disease response) (Shih and Lin, 2017; Renfro et al., 2016). These includes: 1) Targeted design (also called enriched design): randomize only biomarker positive patients to treatment groups. Using a genomic classifier to exclude patients from eligibility of a study requires a substantial level of confidence in the classifier, and a reproducible assay with a high level of sensitivity and specificity; 2) Stratified design: for all-comes untargeted design, treatment groups may not be balanced with respect to the biomarker-defined cohorts. A stratified design randomize patients to treatment groups within marker-defined subgroup (treatment groups are more likely balanced with respect to biomarker status). This design can detect a statistical significant biomarker-by-treatment interaction effect, thereby statistically confirming the predictive ability of the biomarker. This design often requires a relatively large sample size, as its structure resembles multiple randomized trials conducted in parallel; 3) Precision medicine design (also named "marker-based strategy design): patients are randomized into two arms: marker-dependent arm and marker-independent arm. For marker-independent arm, patients are randomized to treatment groups (active treatment T or control treatment C) without considering the biomarker status. In marker-dependent arm, marker-positive patients receive the active treatment T and marker-negative subjects receive control treatment C.

In lieu of the fixed designs that provides no flexibility, movements toward adaptive biomarker based trial designs emerged. "Adaptive" refer to designs utilizing data accumulated from patients early in the trial to prospectively shift accrual, eligibility, or objective later on in the trial. Those adaptive design features an unselected randomized design with sequential hypothesis testing in the overall and marker-positive subpopulations. A predictive signature is developed in the first set of patients. When the overall test based on first set of patients is negative, the subset treatment effect is evaluated in an independent second set of patients.

Adaptive designs are intended to add efficiency to drug development process, allow better use of available resources, enhance decision making, shorten development programs, and more quickly arrive at correct decisions about the therapeutic value of a treatment for

particular group of patients. They are the focus of this dissertation.

# Chapter 2

# Literature review

## 2.1 Adaptive designs on time-to-event outcomes

### 2.1.1 Brannath's confirmatory adaptive designs with Bayesian decision tools

**Primary objective:**

The primary objectives of Brannath et al. (2009) study is (a) to confirm or disregard a sub-population S, which is identified in a separate exploratory study, and (b) to confirm the treatment effect of the novel therapy in the selected target population (that is, either in the sub-population S or in the full population F). The primary endpoint is progression-free survival (PFS), a time to event outcome.

**Trial design:**

The specific design comprises three stages defined by two interim analyses. The aim of the first interim analysis is to decide whether to continue recruiting from F or to continue recruiting patients only from S. The final analysis would consist of testing efficacy in both F and S or only in S, respectively. The second interim analysis allows for possible early stopping, without further adaption of the trial. This adaptive design is an adaptive phase II/III seamless design with population selection at the first interim analysis and the possibility to stop early for futility or early success at the second interim analysis.

**Elucidating example:**

The first interim analysis of the study, for the decision on trial adaption, is based on 170 events for F (i.e. 18.5% of the maximum of 918 events - total information). It is expected

to take place 12 months after first patient first visit, when approximately 600 (50%) of the patients are enrolled. The second interim analysis provides an opportunity of early stopping for superiority. It takes place at approximately 60% of maximum event number, whether the study continues in F (i.e. at 551 events out of a targeted 918 events for the final analysis in F) or in S (i.e. at 384 events out of a targeted 640 events for the final analysis in S), at which time all the patients are expected to be enrolled.

**Multiplicity control within a stage:**

The adaptive closed testing methodology is applied to obtain a valid multiple test procedure for the hypotheses $H_0^{\{F\}}$ and $H_0^{\{S\}}$. This requires combination tests for $H_0^{\{F\}}$, $H_0^{\{S\}}$ and the intersection hypothesis $H_0^{\{S,F\}} = H_0^{\{S\}} \cap H_0^{\{F\}}$. The stage-wise p-value for $p_i^{\{F\}}$ for $H_0^{\{F\}}$, $i = 1, 2, 3$, are defined as above for the stratified logrank tests with strata for $S$ and $S^c$; the $p$-value $p_i^{\{S\}}$ for $H_0^{\{S\}}$ are similar but based on unstratified logrank tests for patient in $S$. The first stage $p$-value $p_1^{\{S,F\}}$ for $H_0^{\{S,F\}}$ is chosen to be the multiplicity adjusted p-value according to Simes' procedure

$$p_1^{\{S,F\}} = \min\{2\min(p_1^{\{S\}}, p_1^{\{F\}}), \max(p_1^{\{S\}}, p_1^{\{F\}})\}.$$

$$p_i^{\{S,F\}} = \min\{2\min(p_i^{\{S\}}, p_i^{\{F\}}), \max(p_i^{\{S\}}, p_i^{\{F\}})\}, i = 2, 3.$$

**Statistical method to analyze data from different stages:**

For simplicity, the p-value combination test approach is followed to analyze data from different stage. In the three stage combination tests with inverse normal combination function, the null $H_0$ is rejected

1) at the first stage if

$$C_1(p_1) = \Phi^{-1}(1 - p_1) \geq c_1,$$

2) at the second stage if

$$C_2(p_1, p_2) = \{w_1\Phi^{-1}(1 - p_1) + w_2\Phi^{-1}(1 - p_2)\}/\sqrt{w_1^2 + w_2^2} \geq c_2,$$

3) and at the last stage if

$$C_3(p_1, p_2, p_3) = w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2) + w_3\Phi^{-1}(1-p_3) \geq c_3,$$

where $w_i > 0$ are weights satisfying $\sum_{i=1}^{3} w_i^2 = 1$ and $\Phi^{-1}$ is the inverse of the standard normal distribution function. The type of combination function, the weights $w_i$ and the boundaries $c_i$ must be predefined in the protocol.

The designs allow possibility of stopping the trial with the acceptance of $H_0$, which will deflate the type I error rate, but does not account for this option in the determination of $c_i$. This has the advantage of not being bound to a specific futility rule at the time of study start.

With a time to event endpoint, the stage-wise p-values could be based on increments of the logrank scores $U_i$ from the right-censored event times at stage $i$, $p_1 = \Phi\{U_1/\sqrt{I_1}\}$ and $p_i = \Phi\{(U_i - U_{i-1})/\sqrt{I_i - I_{i-1}}\}$ for $i = 2, 3$ where $I_i$ is the variance estimate of $U_i$, i.e. the observed Fisher's information at stage $i$. If the increments $U_i - U_{i-1}$ are independent and normally distributed with mean 0, and the variance of $U_i - U_{i-1}$ is consistently estimated by $I_i - I_{i-1}$.

| Patient selection decision | Hypothesis testing strategies |
|---|---|
| Continue with sub-population, $H_0^{\{S\}}$ of primary interest | Reject $H_0^{\{S\}}$ if $C_2(p_1^{\{S,F\}}, p_2^{\{S\}}) \geq c_2$ and $C_2(p_1^{\{S\}}, p_2^{\{S\}}) \geq c_2$ at 2nd interim or if $C_3(p_1^{\{S,F\}}, p_2^{\{S\}}, p_3^{\{S\}}) \geq c_3$ and $C_2(p_1^{\{S\}}, p_2^{\{S\}}, p_3^{\{S\}}) \geq c_3$ |
| Continue with full population, $H_0^{\{F\}}$ and $H_0^{\{S\}}$ of primary interest | Reject $H_0^{\{F\}}$if $C_2(p_1^{\{S,F\}}, p_2^{\{S,F\}}) \geq c_2$ and $C_2(p_1^{\{F\}}, p_2^{\{F\}}) \geq c_2$ at 2nd interim |

The trial can be terminated at the first interim analysis with the rejection of $H_0^{\{S\}}$and $H_0^{\{F\}}$ if $p_1^{\{S\}}$and $p_1^{\{F\}}$are both less or equal to $1 - \Phi(c_1)$.

**Patient selection rule at the first interim analysis:**

Bayesian decision tools are applied in patient selection at the first interim analysis point.

Given the interim data, predictive probabilities will be used to estimate how likely the various null hypothesis are to be rejected, where posterior probabilities can be used to estimate the treatment effect in different populations.

Three predictive probabilities are calculated at the interim analysis, the probability to reject $H_0^{\{S\}}$ when continuing with sub-population $S$ only (denoted by $PP^{\{S\}}$), the probability to reject $H_0^{\{F\}}$ or $H_0^{\{S\}}$ when continuing with F (denoted by $PP^{\{F\}}$). The posterior probability $P^{\{S^c\}}$ that $\theta^{\{S^c\}} \leq \log(\delta)$ where $\delta$ is the predefined desired clinical effect, i.e the posterior probability that the treatment does achieve the desired efficacy in population $S^c$ .

The decision rules at interim are:

1) Stop the study for futility if $PP^{\{F\}} \leq \pi^{\{F\}}$ and $PP^{\{S\}} \leq \pi^{\{S\}}$, where $\pi^{\{F\}}$ and $\pi^{\{S\}}$ are pre-defined.

2) Continue to stage 2 with $S$ when $PP^{\{S\}} > \pi^{\{S\}}$ and $PP^{\{F\}} \leq \pi^{\{F\}}$.

3) Continue to stage 2 with $S$ when $PP^{\{F\}} > \pi^{\{F\}}$ , $PP^{\{S\}} > \pi^{\{S\}}$ and the posterior probability $P^{\{S^c\}}$ is below $\pi^{\{S^c\}}$.

4) Continue to stage 2 with $F$ when $PP^{\{F\}} > \pi^{\{F\}}$ , $PP^{\{S\}} > \pi^{\{S\}}$ and $P^{\{S^c\}} \geq \pi^{\{S^c\}}$.

5) Continue to stage 2 with $F$ when $PPP^{\{F\}} > \pi^{\{F\}}$ and $PP^{\{S\}} \leq \pi^{\{S\}}$.

For details to calculate the posterior and predictive probabilities, see Brannath et al. (2009).

The rejection boundaries were set according to an O'Brien-Fleming $\alpha$-spending function at one-sided level 0.025 with the rejection boundaries $c_1 = 4.97, c_2 = 2.644$ and $c_3 = 1.984$. Note that $C_1(p_1) \geq c_1$ is equivalent to $p_1 \leq 1.32 \times 10^{-6} = 1 - \Phi(c_1)$.

### 2.1.2 Jenkins' adaptive seamless phase II/III design with subpopulation selection using correlated survival endpoints

**Primary objectives:**

The primary objectives of Jenkins et al. (2001) study are (a) to confirm or disregard a sub-population $S^c$; and (b) to confirm increased efficacy with the new treatment in the

selected target population, both the full population ($F$) and subgroup ($S$). The subgroup $S$ is clearly defined at the start of the trial.

**Study design:**

The study is a randomized, parallel group clinical trial with two arms, experimental and control. There will be two distinct stages, an interim analysis takes place based on stage 1 subjects only, where the final analysis is based on all subjects. At the interim analysis, which considers a short-term intermediate time-to-event endpoint, the trial can either:

1) continue in co-primary population F and S, or

2) continue in subgroup S only, or

3) continue in the full population F without analysis in S, or

4) stop for futility

**Elucidating example:**

The interim analysis occurs after 200 PFS events from the 300 patients recruited to the stage 1. A further 800 patients are then recruited to the stage 2 if the trial continues in the full population or co-primary case, with 400 recruited if only the subgroup $S$ is continued in the stage 2. The final analysis based on overall survival (OS) is performed when 250 deaths have occurred in stage 1 subjects and stage 2 subjects have produced 500 deaths in the full or co-primary case or 250 deaths in the subgroup $S$ only case.

**Multiplicity control at final analysis:**

Within each population, there is a single null hypothesis of no difference between arms (denoted as $H_0^F$ and $H_0^S$ ). The alternative hypothesis is that the new treatment demonstrates increased efficacy over the comparator in terms of prolonging OS (denoted as $H_1^F$ and $H_1^S$).

The closure principle is used to control the family-wise error rate at a nominal level $\alpha$. The closure principle consider all possible intersection hypotheses $\cap H_0^j$, where the $H_0^j$ are in the set of original null hypotheses $\{H_0^F, H_0^S\}$. This produces three hypotheses, $H_0^F, H_0^S$,

and also $H_0^{FS}$, which specifies that there is no survival difference in either F, or S. A null hypothesis $H_0^j$ is rejected overall if all intersection hypotheses that imply $H_0^j$ are also rejected. For example, $H_0^F$ can only be rejected overall if individual tests reject both $H_0^F$ and $H_0^{FS}$ at level $\alpha$.

For subjects recruited in stage $i \in \{1, 2\}$ the p-values for testing $H_0^F$ and $H_0^S$ will be denoted $p_i^F$ and $p_i^S$ , respectively. Specifically, $p_1^F$ and $p_1^S$ are based on the OS data for subjects recruited in stage 1 using their OS through stages 1 and 2, while $p_2^F$ and $p_2^S$ are calculated from OS for stage 2 subjects only. The stage $i$ p-values corresponding to $H_0^{FS}$, $p_i^{FS}$ is a function of $p_i^F$ and $p_i^S$ correcting for multiplicity. A Hochberg correction with equal weighting of $H_0^F$ and $H_0^S$ gives $p_i^{FS} = \min\{2\min(p_i^{\{S\}}, p_i^{\{F\}}), \max(p_i^{\{S\}}, p_i^{\{F\}})\}$ .

**Statistical method to analyze data from different stages:**

The final analysis on all subjects use an inverse-normal combination test, which controls the type-I error rate, regardless of the decision at the interim analysis. Weights $w_1$ and $w_2$, with $w_1 = \sqrt{N_1/(N_1+N_2)}$, $w_2 = \sqrt{N_2/(N_1+N_2)}$, are specified to combine the p-values from each stage and the null hypothesis is rejected if $C(p_1, p_2) = \{w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2)\} \geq c$. For a one-side significance level of 0.025, c=1.96.

The weights and p-values to be used in combination tests are shown below.

- Co-primary case - when considering both $H_0^F$ and $H_0^S$
  - Testing $H_0^F$ : $w_1\Phi^{-1}(1-p_1^F) + w_2\Phi^{-1}(1-p_2^F)$
  - Testing $H_0^S$ : $w_1\Phi^{-1}(1-p_1^S) + w_2\Phi^{-1}(1-p_2^S)$
  - Testing $H_0^{FS}$ : $w_1\Phi^{-1}(1-p_1^{FS}) + w_2\Phi^{-1}(1-p_2^{FS})$
- F only case - when considering $H_0^F$ only
  - Testing $H_0^F$ : $w_1\Phi^{-1}(1-p_1^F) + w_2\Phi^{-1}(1-p_2^F)$
  - Testing $H_0^{FS}$ : $w_1\Phi^{-1}(1-p_1^{FS}) + w_2\Phi^{-1}(1-p_2^{FS})$
- S only case - when considering $H_0^F$ only
  - Testing $H_0^S$ : $w_1\Phi^{-1}(1-p_1^S) + w_2\Phi^{-1}(1-p_2^S)$

– Testing $H_0^{FS} : w_1\Phi^{-1}(1-p_1^{FS}) + w_2\Phi^{-1}(1-p_2^{FS})$

**Patient selection rule at the interim analysis:**

The decisions about which hypothesis to proceed to test will be made at the interim analysis. Interim decision rules are based on the estimated hazard ratios for PFS within the full population and the subgroup S. Target values are set and the trial only continues in those subgroups for which the estimated hazard ratio exceeds the target.

| PFS hazard ratio estimated at the interim analysis | $\widehat{HR}^F < 0.8$ | $\widehat{HR}^F \geq 0.8$ |
|---|---|---|
| $\widehat{HR}^S < 0.6$ | Continue co-primary | Continue subgroup S only |
| $\widehat{HR}^S \geq 0.6$ | Continue population F only | Stop for futility |

### 2.1.3 Friede's conditional error function approach for subgroup selection in adaptive clinical trials

**Primary objective:**

The primary objectives of Friede et al. (2012) study are (a) to confirm or disregard a subpopulation $S^c$; and (b) to confirm increased efficacy with the new treatment in the selected target population, both the full population ($F$) and subgroup ($S$). The subgroup $S$ is clearly defined at the start of the trial.

**Study design:**

The trial consists of two stages: interim analysis and final analysis. A decision is made at the interim analysis whether to continue with comparing the experimental and control treatments in both the subgroup and the full population, in the full population alone or in the subgroup alone, this decision being taken on the basis of either a short-term endpoint, or the long-term endpoint (same as the final analysis) observed at the interim analysis. The final analysis is based on the long-term endpoint, i.e., overall survival (OS).

**Multiplicity control within a stage:**

A method proposed by Spiessens and Debois (2010) is used to control the FWER of the multiple tests in subgroup and full population within a stage. In particular, $H_0^{\{S\}}$ is tested at the nominal level $\alpha_1$ with the use of $Z^{\{S\}}$. $H_0^{\{F\}}$ is tested at the nominal level $\alpha_2$ with the use of $Z^{\{F\}}$. Reject $H_0^{\{F,S\}}$ if either of these tests reject, that is, if $Z^{\{S\}} > z_{\alpha_1}$ or $Z^{\{F\}} > z_{\alpha_2}$. To control the test of the intersection hypothesis at level $\alpha$ , it is thus required that

$$P\left(Z^{\{S\}} > z_{\alpha_1} \text{ or } Z^{\{F\}} > z_{\alpha_2} | H_0^{\{F,S\}}\right) = \alpha,$$

this implies

$$P\left(Z^{\{S\}} \leq z_{\alpha_1} \text{ or } Z^{\{F\}} > z_{\alpha_2} | H_0^{\{F,S\}}\right) = \alpha - \alpha_1.$$

Since $Z^{\{S\}}$ is based on a subset of the data used to calculate $Z^{\{F\}}$ , their joint distribution under $H_0^{\{F,S\}}$ is, analogy to that obtained in the group-sequential setting, given by

$$\begin{pmatrix} Z^{\{S\}} \\ Z^{\{F\}} \end{pmatrix} \sim MN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{\tau} \\ \sqrt{\tau} & 1 \end{pmatrix}\right),$$

where $\tau$ is the fraction of information in the subsample with respect to that in the full population, usually approximately equal to the proportion of observations in the subsample. Given $\alpha_1$, we can obtain the value of $\alpha_2$ to achieve overall specified error rate control for the intersection hypothesis, $H_0^{\{F,S\}}$, with the use of numerical integration as following:

$$\int_{-\infty}^{z_{\alpha_1}} \Phi\left(\frac{z_{\alpha_2} - \sqrt{\tau} z^{\{F\}}}{\sqrt{1-\tau}}\right) \Phi(z^{\{F\}}) dz^{\{F\}} = 1 - \alpha.$$

**Statistical method to analyze data from different stages:**

*Conditional error function appproach is applied to analyze data.*

Let $Z_1^{\{F\}}$ and $Z_1^{\{S\}}$ denote standardized test statistics for the full population and the subpopulation based on observed stage 1 data. As explained previously, these are based on the long-term endpoint data, so these may not be available at the time of interim analysis for subgroup selection. Furthermore, Let $Z_2^{\{F\}}$ and $Z_2^{\{S\}}$ denote standardized test statistics for

the full population and the subpopulation based on observed new stage 2 data. The second stage test statistics are independent of the first stage statistics. let $w_1$ and $w_2$ be weights with $w_1^2 + w_2^2 = 1$ and $w_i^2$ proportional to the stage wise sample size at stage $i$ for $i = 1, 2$. Let $S_1^{\{F\}} = w_1 Z_1^{\{F\}}$, $S_1^{\{S\}} = w_1 Z_1^{\{S\}}$, $S_2^{\{F\}} = w_1 Z_1^{\{F\}} + w_2 Z_2^{\{F\}}$ and $S_2^{\{S\}} = w_1 Z_1^{\{S\}} + w_2 Z_2^{\{S\}}$. Then under the intersection hypothesis $H_0^{\{F,S\}}$, we have

$$\begin{pmatrix} S_1^{\{F\}} \\ S_1^{\{S\}} \\ S_2^{\{F\}} \\ S_2^{\{S\}} \end{pmatrix} \sim MN \left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} w_1^2 & w_1^2\sqrt{\tau} & w_1^2 & w_1^2\sqrt{\tau} \\ w_1^2\sqrt{\tau} & w_1^2 & w_1^2\sqrt{\tau} & w_1^2 \\ w_1^2 & w_1^2\sqrt{\tau} & 1 & \sqrt{\tau} \\ w_1^2\sqrt{\tau} & w_1^2 & \sqrt{\tau} & 1 \end{pmatrix} \right).$$

**Patient selection rule at interim analysis:**

*Interim selection $\epsilon$-rules.*

In terms of subgroup selection, continue with both full population and subpopulation if the corresponding $Z$-statistics are sufficiently close to each other and otherwise to continue only with the population with the maximal test statistics, that is, to continue in the second stage with populations $i \in \{F, S\}$ for which $Z_1^{\{i\}} \geq \max(Z_1^{\{F\}}, Z_1^{\{S\}}\} - \epsilon$ with $\epsilon \geq 0$. The rule with $\epsilon = 0$ means that only the tests regarding the population with the maximal test statistic is taken forward. If the adaption is based on an early outcome rather than the final outcome, then $Z_1^{\{F\}}$ and $Z_1^{\{S\}}$ are replaced by the respective statistics $Z_1^{\{F\}*}$ and $Z_1^{\{S\}*}$ obtained on the basis of the early outcome data.

### 2.1.4 Mehta's biomarker driven population enrichment for adaptive oncology trials

**Primary objective:**

The primary objectives of Mehta et al. (2014) study are (a) to confirm or disregard a subpopulation $\overline{S}$; and (b) to confirm increased efficacy with the new treatment in the selected target population, both the full population ($F$) and subgroup ($S$). The subgroup $S$ is clearly defined at the start of the trial.

**Study design:**

This trial is two-stage design, in which an experimental arm (E) is compared to a control arm (C) with respect to a time-to-event endpoint, say survival. Patients arriving in staggered fashion from some population F are stratified on the basis of a binary biomarker into subgroup $S$ or subgroup $\overline{S}$ and then randomized to one of the two treatment arms. Let $\theta_S$ and $\theta_{\overline{S}}$ denote the negative log hazard ratio of E relative to C in subgroup $S$ and $\overline{S}$, respectively. Testing null hypothesis $H^S : \theta_S \leq 0$ and $H^{\overline{S}}$: $\theta_{\overline{S}} \leq 0$ against one-sided alternatives, with strong control of the family-wise error rate (FWER). At the interim analysis, the data are unblinded, and a decision is taken to either continue with $F$ for the reminder of the trial, drop $\overline{S}$ and continue with $S$ for the reminder of the trial, or terminate the trial for futility. This can maximize the power to reject $H^S$ by enriching the reminder of the trial with subgroup $S$ patient only.

**Non-small cell lung cancer trial example:**

A numerical example is presented: the trial itself is hypothetical. But the design inputs are realistic, being based on real trial data. The hypothetical trial is to compare an experimental drug (Treatment E) to standard of care (Treatment C) in patients with metastatic non-small cell lung cancer. The primary endpoint is the same at interim and final analysis, i.e., progression free survival (PFS). The experiment drug is targeted at an biomarker which partitioned the population into subgroup $S$ and subgroup $\overline{S}$. The prior belief is that the hazard ratio for the treatment E versus treatment C is between 0.5 and 0.6 in subgroup $S$, where in subgroup $\overline{S}$, it is not expected to be any lower than 0.8. The total sample size is 160. In the first stage, 80 patients are recruited (including 40 patients from subgroup $S$ and 40 patients from subgroup $\overline{S}$). The interim analysis is performed when the required 80 patients are enrolled in the study. One of three decisions is taken based on the results of the interim analysis:

1) Recruit the remaining 80 patients in equal numbers for each subgroup so that 40 patients are enrolled from subgroup $S$ and 40 patients from subgroup $\overline{S}$.

2) Drop subgroup $\overline{S}$. and recruit the remaining 80 patients from subgroup $S$ only.

3) Terminate the trial for futility.

**Multiplicity control within a stage:**

Closed testing procedure with the conditional error rate approach is applied to adjust for multiplicity. This implies that the three hypotheses, $H^S$, $H^{\overline{S}}$, and $H^{\{S,\overline{S}\}} = H^S \cap H^{\overline{S}}$ must each to be controlled at level $\alpha$. It is convenient to formulate this requirement in terms of the decision of the decision function $\psi^S, \psi^{\overline{S}}$, and $\psi^{\{S,\overline{S}\}}$ such as the following: $\psi^S = 1$ if $H^S$ is rejected and 0 otherwise; $\psi^{\overline{S}} = 1$ if $H^{\overline{S}}$ is rejected and 0 otherwise; $\psi^{\{S,\overline{S}\}} = 1$ if $H^{\{S,\overline{S}\}}$ is rejected and 0 otherwise. We require $E_0(\psi^S) = E_0(\psi^{\overline{S}}) = E_0(\psi^{\{S,\overline{S}\}}) = \alpha$ where $E_0(.)$ denotes expectation under the appropriate null hypothesis.

Let $T^S_{k^S}$ ($T^{\overline{S}}_{k^{\overline{S}}}$) be the logrank score for testing the null hypothesis $H^S$ ($H^{\overline{S}}$) after observing $k^S(k^{\overline{S}})$ deaths in subgroup $S$ ($\overline{S}$). Then the decision function $\psi^S$and $\psi^{\overline{S}}$ are indicator variables $\psi^S = I(T^S_{k^S} > c^S)$ and $\psi^{\overline{S}} = I(T^{\overline{S}}_{k^{\overline{s}}} > c^{\overline{S}})$ for suitable critical boundaries $c^S$ and $c^{\overline{S}}$, respectively, that satisfy the level requirement. The decision function for the intersection hypothesis is the indicator variable $\psi^{\{S,\overline{S}\}} = I((T^S_{k^S}, T^{\overline{S}}_{k^S}) \in R)$ where R is a rejection region of the form

$$R = \{(t^S, t^{\overline{S}})|((t^S > d^S) \cup (t^{\overline{S}} > d^{\overline{S}})\}.$$

**Statistical method for adaptive feature at interim:**

The type I error of the modified design will be protected if the conditional error rates of the test of $H^S$ and $H^{\{S,\overline{S}\}}$ in the modified design are bounded by the corresponding conditional error rates of the original design. To be specific, if it is decided to drop subgroup $\overline{S}$ at the interim analysis, and possibly increase the number of events for the subgroup $S$ from $k^S$ to $\widetilde{k}^S$, we must define a new final decision function $\Psi^S$ for testing $H^S$ and $H^{\{S,\overline{S}\}}$ that preserves the conditional rejection probabilities

$$E_0(\Psi^S|X) \leq E_0(\psi^S|X) \text{ and } E_0(\Psi^S|X) \leq E_0(\psi^{\{S,\overline{S}\}}|X),$$

where $X$ is the set of all interim information on patients in $S$ and $\overline{S}$ used for the decision on the design modification. It may be impossible to explicitly specify the vector $X$, which

includes observed times-to-event as well as preliminary information correlated with time-to-event from patients who have not yet reached the endpoint. However, it is sufficient to condition on a random vector $Y$ for which you can compute the conditional expectations $E_0(\psi^S|Y)$, $E_0(\psi^{\{S,\overline{S}\}}|Y)$, and $E_0(\Psi^S|Y)$, and which has the property that $X$ is stochastically independent of the decisions functions $\psi^S$, $\psi^{\{S,\overline{S}\}}$, and $\Psi^S$ given $Y$. The conditional rejection probability (CRP) principle requires the new decision function to satisfy

$$E_0(\Psi^S|Y) \leq E_0(\psi^S|Y) \text{ and } E_0(\Psi^S|Y) \leq E_0(\psi^{\{S,\overline{S}\}}|Y).$$

The new decision function is an indicator variable of the form $\Psi^S = I(T^S_{\widetilde{k}^S} > \widetilde{c}^S)$. If subgroup $\overline{S}$ is not dropped at the interim analysis, then of course $\Psi^S$ will not be computed and $H^S$ will be rejected by a closed test in accordance with the decision function $\psi^S, \psi^{\overline{S}}$,and $\psi^{\{S,\overline{S}\}}$.

At the calendar time of the interim analysis, a subset $S' \subseteq S$ of patients has been randomized, of whom a subset of patients $S'_{dead}$ has already died, while its complement $S'_{risk}$ consists of patients in $S'$still at risk.The subsets $\overline{S}', \overline{S}'_{dead}, \overline{S}'_{risk}$ are defined similarly. Our method permits the use of all available information in $S'$ and $\overline{S}'$, including even the information about early outcomes like PFS or tumor regression in $S'_{risk}$ and $\overline{S}'_{risk}$, for the interim decision making.

It is necessary to specify the following quantities prior to unblinding the interim data:

1) Specify $k^S$ and $k^{\overline{S}}$ , the total number of events to be obtained from subgroups $S$ and $\overline{S}$, respectively, at the time of the final analysis under the original design.

2) Specify $k^{\overline{S}'}$ , the contribution from the subset $S'$ to $k^{\overline{S}}$. This specification is needed to ensure that the conditioning event Y will be properly defined even if recruitment to $\overline{S}$ is stopped after the interim analysis.Ideally, $k^{\overline{S}'}$ should be chosen so that the arrival of the last of the $k^{\overline{S}'}$ events in $\overline{S}'$is closely aligned in calendar time with the arrival of the last of the $k^{\overline{S}}$events in $\overline{S}$.

The conditional events are defined as following:

a) The conditional events from $S'$: The conditioning event is $T^{S'}_{k^S}$, the logrank statistics

calculated from patients belonging to subset $S^{'}$ at the time of the arrival of the $k^S$th event from subgroup S. This implies that the conditioning events is not observed at the time of the interim analysis but rather at the time of the pre-planned final analysis for $H^S$ under the original design. Let $S^{''} = S \backslash S^{'}$ denote the subset of patients in $S$ that are enrolled after the interim analysis. Let $k^{S^{'}}$ be the contribution from patients in subset $S^{'}$ to the $k^S$ events required from subgroup $S$.

b) The conditional events from $\overline{S}^{'}$: Let $\overline{S}^{''} = \overline{S} \backslash \overline{S}^{'}$ denote the subset of patients in $\overline{S}$ that are enrolled after the interim analysis under the original design. Pre-specifying the total number of events recruited from $\overline{S}$ is $k^{\overline{S}}$ with the first $k^{\overline{S}^{'}}$ of these events to be contributed from subset $\overline{S}^{'}$. Therefore, the number of events to be contributed from subset $\overline{S}^{''}$ must be $k^{\overline{S}^{''}} = k^{\overline{S}} - k^{\overline{S}^{'}}$. Not that $k^{\overline{S}}$ and $k^{\overline{S}^{'}}$ are pre-specified, $k^{\overline{S}^{''}}$ is well defined even if recruitment to subgroup $\overline{S}$ is stopped after the interim analysis. The conditioning event is $T^{\overline{S}^{'}}_{(k^{\overline{S}^{'}}, k^{\overline{S}^{''}})}$. A logrank statistics computed from patients belonging to subset $\overline{S}^{'} \subseteq \overline{S}$ as follows:

   i) If $\overline{S}$ is dropped at the interim analysis, $T^{\overline{S}^{'}}_{(k^{\overline{S}^{'}}, k^{\overline{S}^{''}})}$ is computed at the calendar time that $k^{\overline{S}^{'}}$ events have arrived from $\overline{S}^{'} \subseteq \overline{S}$.

   ii) If $\overline{S}$ is not dropped at the interim analysis, $T^{\overline{S}^{'}}_{(k^{\overline{S}^{'}}, k^{\overline{S}^{''}})}$ is computed at the later of the two calendar times when either $k^{\overline{S}^{'}}$ events have arrived from $\overline{S}^{'}$ or $k^{\overline{S}}$ events have arrived from $\overline{S}$ , with only the first $k^{\overline{S}^{'}}$ events have arrived from $\overline{S}^{'}$ contributing to the calculation of the statistic.

c) The conditioning event $Y$ is thus the pair of logrank statistics $\left(T^{S^{'}}_{k^s}, T^{\overline{S}^{'}}_{(k^{\overline{S}^{'}}, k^{\overline{S}^{''}})}\right)$. Suppose the trial is modified at the interim analysis, by discontinuing enrollment to subgroup $\overline{S}$, and possibly increasing the number of events in $S$ from $k^S$ to $\widetilde{k}^S$ for the final analysis. In order to preserve the type I error, the new critical value $\widetilde{c}^S$ must satisfy the following CRP condition:

$$P_0\left(T^S_{\widetilde{k}^S} > \widetilde{c}^S | T^{S^{'}}_{\widetilde{k}^S}\right) \leq \min\left\{P_0\left(T^S_{k^S} > c^S | T^{S^{'}}_{k^S}\right), P_0\left((T^S_{k^s}, T^{\overline{S}}_{k^{\overline{S}}}) \in R | T^{S^{'}}_{k^S}, T^{\overline{S}^{'}}_{(k^{\overline{S}^{'}}, k^{\overline{S}^{''}})}\right)\right\}$$

where $P_0(.)$ denotes probability under the appropriate null hypothesis and $T^S_{\widetilde{k}^S}$ is the

logrank statistics computed at the time of the final analysis, when $\widetilde{k}^S$ events have arrived from patients in subgroup $S$.

Under the null hypothesis $H^S$, asymptotically, $T^{S'}_{k^S}$ and $T^{S}_{k^S} - T^{S'}_{k^S}$ are independent at information time $k^S$; Similarly, $T^{\overline{S}'}_{k^S}$ and $T^{\overline{S}}_{k^S} - T^{\overline{S}'}_{k^S}$ are independent at information time $k^{\overline{S}}$ . It is this stochastic independence that permits interim decision to be based on all available stage 1 data without influencing the final number of events to be realized from the stage 1 recruits.

**The interim decision rules:**

Simple decision rules, based on conditional power, are utilized for interim decision. Let $CP_S$ and $CP_{\overline{S}}$ denote the conditional power, under the original design, to reject $H^S$ and $H^{\overline{S}}$. Then

1) if $\widehat{HR}_S$ , the estimate of the hazard ratio for treatment versus control in subgroup $S$, is less than $A$, terminate the trial for futility;

2) if conditional power $CP_S > B$ and $CP_{\overline{S}} < C$, stop further enrollment to subgroup $\overline{S}$ and enroll all remaining patients to subgroup $S$; and otherwise;

3) continue to the end of the trial with both subgroups.

## 2.2 Adaptive enrichment designs on continuous outcomes

### 2.2.1 Wang et al. (2007)'s adaptive stratified enrichment design

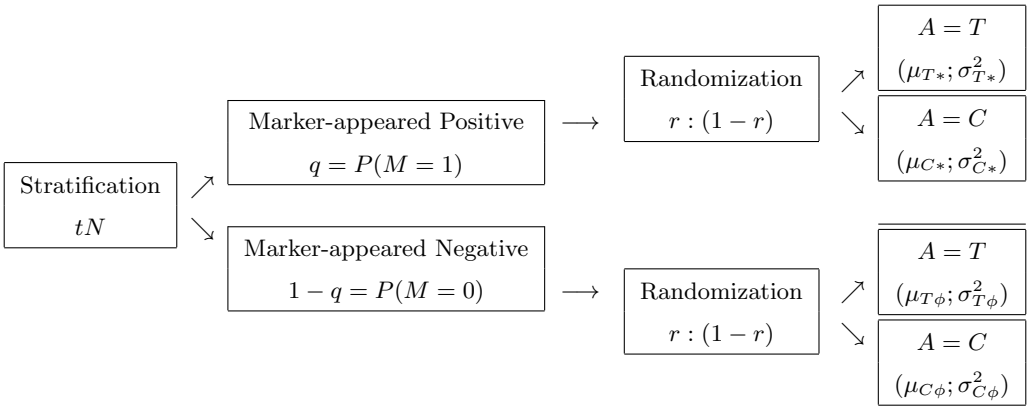
Wang et al. (2007) introduced one of the first biomarker-based clinical trial designs which allowed mid-trial adaption based on the results of interim analyses. The design is a two-stage adaptive design. The first stage is a stratified design based on biomarker status of the patients. At an interim analysis, if the experiment treatment effects reach a futility threshold in the marker-negative group, accrual of marker-negative patients is terminated. The remaining sample size is re-allocated to marker-positive patients. In that case, the primary hypothesis tested at the trial's conclusion is the treatment effect in marker-positive subgroup. On the other hand, if the futility is not reached in the marker-negative group at

the interim analysis, the trial continues as stratified design. In this case, the design performs both overall and subgroup-specific tests of treatment effects at the final analysis timepoint with trial-wise type I error control. Wang et al. (2007)'s approach applies for continuous endpoints (or asymptotic normally distributed endpoints). More recent advancement in adaptive enrichment designs includes the development approaches suitable for time-to-event endpoints.

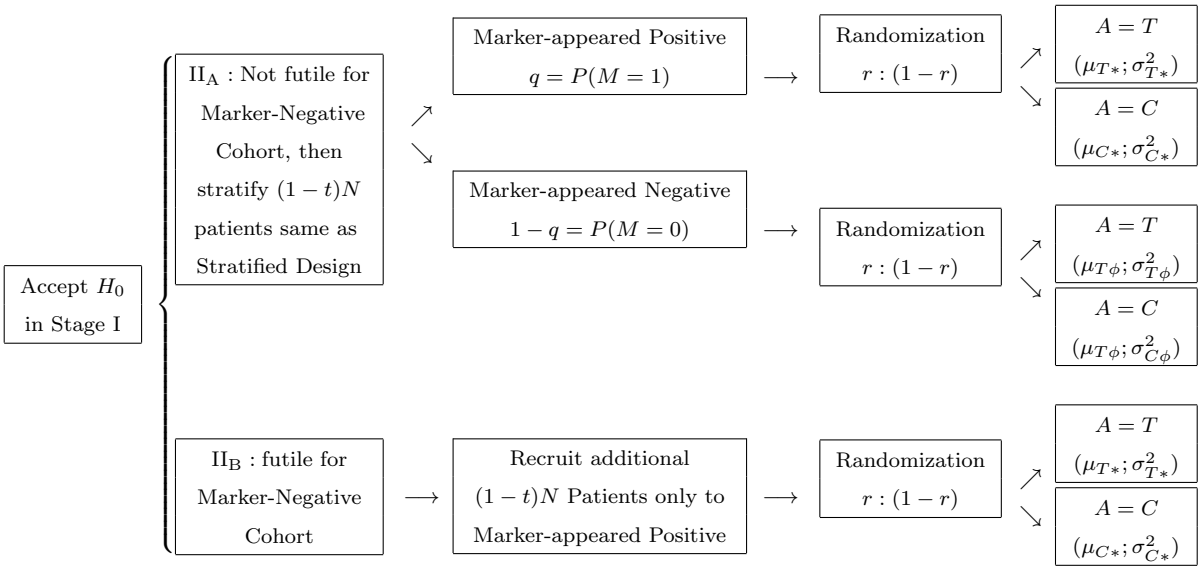
### 2.2.2 Lin et al.'s two-stage enrichment design with adjustment for misclassification in predictive biomarkers

Unlike the traditional stratified designs using patient's baseline characteristics such as gender, age, race, or ECOG performance status, the biomarker status and classification rule are prone to errors due to imperfect assays and/or classification rules. In 2019, Lin, Shih, and Lu introduced a two-stage enrichment design with adjustment for misclassification in predictive biomarkers. The design scheme is similar to Wang et al. (2007)'s adaptive stratified enrichment design, but *with classification errors*. The design was intended for normally or asymptotic normally distributed outcomes. The following is the diagram of the stratified two-stage enrichment design with misclassification in predictive biomarkers.

The diagram of the stratified two-stage enrichment design:



**Stage II**




where $N$ is the total sample size in the study, $t$ is the information assigned for Stage I, $q$ is the prevalence rate of biomarker-appeared positive, $r$ is the randomization proportion to receive the active treatment, and $\mu$, $\sigma^2$ are the corresponding mean and variance for designated treatment group.

# Chapter 3

# Research questions and objectives

## 3.1 Study populations and outcomes

We want to design adaptive clinical trials with *survival outcomes* in the following setting:

1. A diagnostic test is available to partition patients into predictive biomarker positive $S+$ and negative $S-$;

2. A predictive biomarker with biological basis is identified with the assumption of considerable treatment efficacy in $S+$ but little in $S-$.

As we know, even though there is a strong biological basis for the above assumption, the ultimate proof that a biomarker is predictive can only come from a randomized clinical trial.

## 3.2 Hypothesis testing

For the clinical study based on the predictive biomarker, we are interested in the treatment effect on the overall patient population as well as for the marker-positive cohort. With survival outcomes, the treatment effect will be based on either difference in survival functions or the hazard functions.

In terms of survival function, we are interested in testing hypothesis of the treatment difference for the overall patient population, that is

$$H_{0a} : S_T(t) = S_C(t) \text{ vs. } H_{1a} : S_T(t) \neq S_C(t),$$

and in testing hypothesis of the treatment difference for the marker-positive cohort,

$$H_{0+} : S_{+T}(t) = S_{+C}(t) \text{ vs. } H_{1+} : S_{+T}(t) \neq S_{+C}(t),$$

where $S_T(t)$ is the survival function for the overall population who receive the active treatment, and $S_C(t)$ is the survival function for the overall population who receive the control treatment, where $S_{+T}(t)$ is the survival function for the biomarker positive subjects who receive the active treatment, and $S_{+C}(t)$ is the survival function for the biomarker positive subjects who receive the control treatment.

In terms of hazard functions, the hypotheses can be written as follows. For the overall patient population, testing

$$H_{0a} : \lambda_T(t) = \lambda_C(t) \text{ vs. } H_{1a} : \lambda_T(t) \neq \lambda_C(t),$$

and for the marker-positive cohort, testing

$$H_{0+} : \lambda_{+T}(t) = \lambda_{+C}(t) \text{ vs. } H_{1+} : \lambda_{+T}(t) \neq \lambda_{+C}(t),$$

where $\lambda_T(t)$ is the hazard function for the overall population who receive the active treatment, and $\lambda_C(t)$ is the survival function for the overall population who receive the control treatment, where $\lambda_{+T}(t)$ is the hazard function for the biomarker positive subjects who receive the active treatment, and $\lambda_{+C}(t)$ is the hazard function for the biomarker positive subjects who receive the control treatment.

Since we are pursuing either treatment effect on the overall population or only the marker-positive cohort, we are testing the compositive hypotheses as shown below.

In terms of survival functions,

$$H_0 : S_T(t) = S_C(t) \ \& \ S_{+T}(t) = S_{+C}(t) \text{ vs.}$$

$$H_1 : S_T(t) \neq S_C(t) \text{ or } S_{+T}(t) \neq S_{+C}(t).$$

In terms of hazard,

$$H_0 : \lambda_T(t) = \lambda_C(t) \ \& \ \lambda_{+T}(t) = \lambda_{+C}(t) \text{ vs.}$$

$$H_1 : \lambda_T(t) \neq \lambda_C(t) \text{ or } \lambda_{+T}(t) \neq \lambda_{+C}(t).$$

Since these two hypotheses are equivalent, in the later chapters, we will discuss only in terms of hazard functions difference between the active treatment versus the control in respective subpopulations. The test will be based on log rank statistics.

## 3.3 Research objectives

In Part I, our goal is to construct a two-stage enrichment design to compare the treatment arm ($T$) to a control arm ($C$) with respect to a time-to-event endpoint, assuming no misclassification of biomarker status. That is to say, our biomarker classification is perfect, and our stratified design is based on true biomarker status. At interim analysis, after certain number of events are observed based on the Stage I subjects (i.e., $tN$ to enroll), a decision is made whether to recruit the Stage II subjects either 1) from full population $F$ (enroll the rest $(1-t)N$ planned subjects from full population); or 2) from subgroup $S+$ only (enrich and enroll the rest $(1-t)N$ planned subjects from subgroup $S+$ only). The final analysis is performed when a pre-specified number of events from Stage I subjects and a pre-sepcified number of events from Stage II subjects are observed. Both interim analysis and final analysis are based on the same time-to-event endpoints.

In Part II, our goal is to construct a two-stage stratified design to compare the treatment arm ($T$) to a control arm ($C$) with respect to a time-to-event endpoint, when the biomarker classification is subjected to errors (there is misclassification for biomarker). In this case, our stratified design is based on the biomarker-appeared status. This study is of importance because the biomarker misclassification is very common. For instance, in phase I KEYNOTE-001 trial, the method for biomarker classification is imperfect, with $\lambda_{sen} = \lambda_{spec} \approx 0.80$ (see Garon et al., 2015 and Herbst et al., 2016). A two-stage group sequential trial is designed in this setting.

In Part III, we extend the misclassification adjustments to the two-stage enrichment design discussed in Part I.

Here is how the dissertation is organized:

**Part I.** Two-stage enrichment design without misclassification of predictive biomarker.

**Part II.** Two-stage stratified design with misclassification of predictive biomarker.

**Part III.** Two-stage enrichment design with misclassification of predictive biomarker.

# Part I

# Two-Stage Enrichment Design Assuming No Misclassification

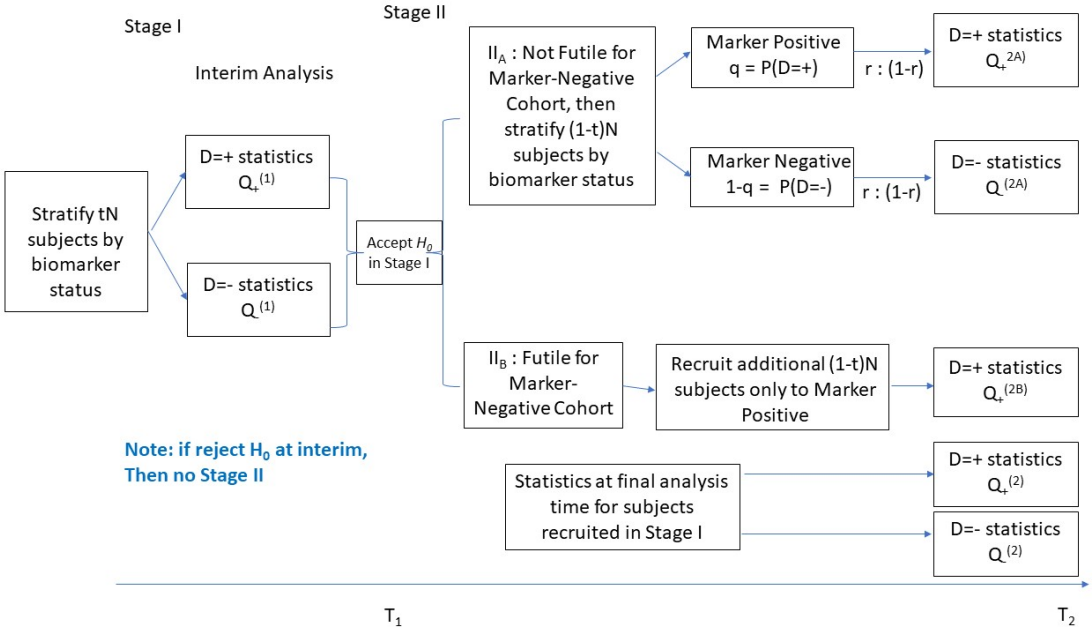
# Chapter 4

# Statistical methods for two-stage enrichment design

In this chapter, we develop a two-stage enrichment design with survival outcomes. Log rank statistics are derived for testing the treatment effects. To simplify the problem and set the frame work in the later chapters, we assume there are no biomarker misclassifications in this chapter. Therefore, our two-arm randomized clinical trial is stratified by the true biomarker status in this chapter instead of by biomarker-appeared status (with misclassification) for the later chapters. Sequential tests and methods to control family-wise type I error rate are detailed. Power and sample size calculations are shown.

## 4.1 Two-stage adaptive enrichment design for survival outcome

The following shows the diagram of the two-stage enrichment design we will discuss in this chapter.



where $D$ denotes the true marker status, $t$ represents the fraction of the total sample size $N$ that is allocated to Stage I, and $q$ represents the prevalence rate of the biomarker positive. The statistics $Q_i^j$ is defined in Section 4.4, $i = +, -$ (marker positive, negative) and $j = 1, 2, 2A, 2B$ for analysis time/scenarios in the subsequent sections.

## 4.2 Stratified randomization based on true marker status

Stratified randomization is based on true marker status, assuming that the biomarker classification is perfect, with randomization ratio of $r$ to $1 - r$ (i.e., assign $r$ percent of patients to active treatment and $1 - r$ percent of patients to control treatment) where $0 < r < 1$.

At $k^{th}$ event time, let $n_{ik}$ be the total number of subjects at risk, $d_{ik}$ be the total number of events for ith marker group, where $i = +, -$. Let $d_+^{(1)}$ and $d_+^{(2)}$ be the total number of events for marker positive group from event time 0 to specified interim analysis time $T_1$ and final analysis time $T_2$ (for both active treatment and control group) within marker positive strata, respectively.

| Death | At risk at $k^{th}$ event time |
|---|---|
| $d_{ijk}$ | $n_{ijk}; i = +, -; j = T, C$ |
| $d_{ik}$ | $n_{ik} = n_{iTk} + n_{iCk}; i = +, -; j = T, C$ |

Note: $i$ indexes for marker status: $i = +$ for marker positive; $i = -$ for marker negative; $j$ indexes for treatment group: $j = T$ for active treatment, $j = C$ for control treatment. Let $Q$ denote log rank statistic; $Q_i^{(j)}$ be the log rank statistic for $i = +, -$ (marker positive, negative) and $j$ =1, 2, 2A, 2B for analysis time/scenarios in the subsequent sections.

## 4.3 Difference between observed events and expected events in marker positive group

Let total number of deaths at $k^{th}$ event time be $d_{+k} = d_{+Tk} + d_{+Ck}$ for marker positive group. Consider the difference between observed number of deaths and the expected number of deaths in active treated group at $k^{th}$ event time, we have

$$d_{+Tk} - \frac{d_{+k} n_{+Tk}}{n_{+k}} = -d_{+Ck} + \frac{d_{+k} n_{+Ck}}{n_{+k}},$$

and the variance of $d_{+Tk}$

$$\widehat{Var}(d_{+Tk}|n_{+k}) = \frac{n_{+Tk}n_{+Ck}d_{+k}(n_{+k}-d_{+k})}{n_{+k}^2(n_{+k}-1)}.$$

Let

$$Q_+ = \sum_{k=1}^{d_+}(d_{+Tk} - \frac{d_{+k}n_{+Tk}}{n_{+k}}).$$

We have

$$\widehat{Var}(Q_+) = \sum_{k=1}^{d_+} \frac{n_{+Tk}n_{+Ck}d_{+k}(n_{+k}-d_{+k})}{n_{+k}^2(n_{+k}-1)}$$

and asymptotically

$$\frac{Q_+^2}{\widehat{Var}(Q_+)} \xrightarrow{D} \chi_1^2.$$

This test statistic based on the difference between observed number of events and expected number of events is closely related to the hazard rate difference in two groups, when the difference is expressed using counting process (FlemingHarrington, 1991).

## 4.4 Statistical method for two-stage adaptive enrichment design for survival outcome

We are considering a two-stage design with one interim analysis at time $T_1$ and final analysis time at time $T_2$. In total, $N$ subjects are planned for the two stages, with $tN$ subjects enrolled in Stage I (from start time $T_0$ to interim analysis time $T_1$). Among subjects enrolled in Stage I, let $n_+^{(1)}$ be expected total number of subjects (enrolled in Stage I) in marker positive group and $n_-^{(1)}$ be expected total number of subjects (enrolled in Stage I) in marker negative group; $n_+^{(1)} + n_-^{(1)} = tN$, and $n_+^{(1)} = ptN$.

### Asymptotic distribution of log-rank statistics

Let $Q_+^{(1)}$ and $Q_-^{(1)}$ be log rank statistics at interim analysis time $T_1$, for marker positive strata and negative strata, respectively.

$$E(Q_i^{(1)}) = E\left(\sum_{k|t_k \le T_1}(d_{iTk} - \frac{d_{i+k}n_{iTk}}{n_{i+k}})\right),$$

where $i = +$ for marker positive and $i = -$ for marker negative group.

At the interim analysis $T_1$, let $d_+^{(1)}$ be the expected number of deaths in marker positive group and $Q_+^{(1)} = \sum_{k=1}^{d_+^{(1)}} (d_{+Tk} - \frac{d_{+k} n_{+Tk}}{n_{+k}})$, and $\mu_+^{(1)}$ be the asymptotic mean of $Q_+^{(1)}$:

$$\mu_+^{(1)} = r(1-r) d_+^{(1)} \log \theta_+,$$

where $\theta_+$ is the hazard ratio between the active treatment and the control for marker positive group (Schoenfeld, 1981).

The asymptotic variance $(\sigma_+^{(1)})^2$ of $Q_+^{(1)}$ is $r(1-r)d_+^{(1)}$. The standardized log-rank statistic for marker positive group is

$$Z_+^{(1)} = (Q_+^{(1)} - \mu_+^{(1)})/\sigma_+^{(1)} \xrightarrow{D} N(0,1).$$

Similarly let $Q_-^{(1)}$ denote the log rank statistic at interim analysis time $T_1$ for the marker negative group. Let $d_-^{(1)}$ be expected number of death in marker negative group and $Q_-^{(1)} = \sum_{k=1}^{d_-^{(1)}} (d_{-Tk} - \frac{d_- n_{-Tk}}{n_-})$, and $\mu_-^{(1)}$be the asymptotic mean of $Q_-^{(1)}$:

$$\mu_-^{(1)} = r(1-r) d_-^{(1)} \log \theta_-,$$

where $\theta_-$ is the hazard ratio between active treatment and control for marker negative group.

The asymptotic variance $(\sigma_-^{(1)})^2$ of $Q_-^{(1)}$ is $r(1-r)d_-^{(1)}$. The standardized log rank statistic for marker negative group is:

$$Z_-^{(1)} = (Q_-^{(1)} - \mu_-^{(1)})/\sigma_- \xrightarrow{D} N(0,1).$$

At the interim analysis, when there is no misclassification, $Q_+^{(1)}$ and $Q_-^{(1)}$ are independent.

The test statistic $Q^{(1)}$ for overall treatment effect at interim analysis time $T_1$ can be defined as a weighted sum of treatment effects for marker positive strata (with weight $w_+$) and marker negative strata (with weight $w_-$). Note: $w_+$ and $w_-$ can be any positive number

and the sum of them don't have to be 1. There are two weighting scenarios reported in Shih and Lin (2017): when $w_+ = w_- = 1$, the absolute treatment effect can be obtained; when $w_+ = p$, and $w_- = 1 - p$, treatment utility is constructed. We use the latter weights in this and later chapters to construct the treatment utility.

Let

$$Q^{(1)} = pQ_+^{(1)}/\sigma_+^{(1)} + (1-p)Q_-^{(1)}/\sigma_-^{(1)}.$$

The variance of $Q^{(1)}$ is

$$Var(Q^{(1)}) = (\sigma^{(1)})^2 = p^2 + (1-p)^2.$$

Under $H_0$

$$Z^{(1)} = \frac{Q^{(1)}}{\sqrt{Var(Q^{(1)})}} = \frac{Q^{(1)}}{\sigma^{(1)}} \xrightarrow{d} N(0,1).$$

At interim analysis time $T_1$, a futility criterion ($Q_-^{(1)} \geq c_0$) for marker-negative subjects will be used, to decide the enrollment pattern for Stage II, between two mutually exclusive scenarios $II_A$ and $II_B$:

$II_A$: Enroll full population (both marker positive and marker negative subjects) for the remaining $(1-t)N$, when $Q_-^{(1)} < c_0$;

$II_B$: Stop enrolling marker-negative subjects and only enroll $(1-t)N$ marker positive subjects for Stage II, when $Q_-^{(1)} \geq c_0$.

Going to Stage II, under scenario $II_A$ , we continue enrolling $(1-t)N$ marker-unselected subjects, the corresponding test statistics at the final analysis time $T_2$ are as follows.

Let $Q_+^{(2)}$ and $Q_-^{(2)}$ be log rank statistics, and $d_+^{(2)}$, $d_-^{(2)}$ be the expected number of deaths at time $T_2$ for subjects enrolled in Stage I, while $Q_+^{(2A)}$ and $Q_-^{(2A)}$ the log rank statistics, and $d_+^{(2A)}$ , $d_-^{(2A)}$ the expected number of death at time $T_2$ for subjects enrolled in Stage II. Then the standardized log rank statistic at final analysis time $T_2$ for marker positive strata is

$$Z_+ = \left(Q_+^{(2)} + Q_+^{(2A)}\right) / \sqrt{\left(Var(Q_+^{(2)}) + Var(Q_+^{(2A)})\right)},$$

where $Q_+^{(2)}$and $Q_+^{(2A)}$ are independent, $Q_+^{(2)}$ contains data for patients enrolled in Stage I (from $T_0$ to $T_2$), and $Q_+^{(2A)}$ contains data from patients enrolled in Stage II (from $T_1$ to $T_2$). We have

$$E(Q_+^{(2)}) = E\left(\sum_{k|t_k \le T_2}(d_{+Tk} - \frac{d_{+k}n_{+Tk}}{n_{+k}})\right) = r(1-r)d_+^{(2)}\log\theta_+,$$
$$E(Q_+^{(2A)}) = E\left(\sum_{k|T_1 < t_k \le T_2}(d_{+Tk} - \frac{d_{+k}n_{+Tk}}{n_{+k}})\right) = r(1-r)d_+^{(2A)}\log\theta_+,$$

and

$$Var(Q_+^{(2)}) = (\sigma_+^{(2)})^2 = r(1-r)d_+^{(2)},$$
$$Var(Q_+^{(2A)}) = (\sigma_+^{(2A)})^2 = r(1-r)d_+^{(2A)}.$$

Similarly, the standardized log rank statistic for marker negative group at time $T_2$ is:

$$Z_- = \left(Q_-^{(2)} + Q_-^{(2A)}\right) / \sqrt{\left(Var(Q_-^{(2)}) + Var(Q_-^{(2A)})\right)},$$

where $Q_-^{(2)}$and $Q_-^{(2A)}$ are independent. We have

$$E(Q_-^{(2)}) = E\left(\sum_{k|t_k \le T_2}(d_{-Tk} - \frac{d_{-k}n_{-Tk}}{n_{-+k}})\right) = r(1-r)d_-^{(2)}\log\theta_0,$$
$$E(Q_-^{(2A)}) = E\left(\sum_{k|T_1 < t_k \le T_2}(d_{-Tk} - \frac{d_{-k}n_{-Tk}}{n_{-k}})\right) = r(1-r)d_-^{(2A)}\log\theta_-,$$

and

$$Var(Q_-^{(2)}) = (\sigma_-^{(2)})^2 = r(1-r)d_-^{(2)},$$
$$Var(Q_-^{(2A)}) = (\sigma_-^{(2A)})^2 = r(1-r)d_-^{(2A)}.$$

The test statistic for overall treatment effect at time $T_2$ is

$$Q = pQ_+/\sigma_+ + (1-p)Q_-/\sigma_-,$$

and the variance of $Q$ is

$$Var(Q) = \sigma^2 = p^2 + (1-p)^2.$$

Under $H_0$,

$$Z = \frac{Q}{\sqrt{Var(Q)}} = \frac{Q}{\sigma} \xrightarrow{d} N(0,1).$$

Under scenario $II_B$ , we enroll $(1-t)N$ marker positive subjects, the corresponding test statistics at time $T_2$ are shown as following.

Let $Q_+^{(2B)}$ be the log rank statistic at the time $T_2$, for subjects enrolled in Stage II. Then the log rank statistic $\overline{Q}_+ = Q_+^{(2)} + Q_+^{(2B)}$ at final analysis time $T_2$ is

$$\overline{Z}_+ = \left(Q_+^{(2)} + Q_+^{(2B)}\right) / \sqrt{\left(Var(Q_+^{(2)}) + Var(Q_+^{(2B)})\right)},$$

where $Q_+^{(2)}$and $Q_+^{(2B)}$ are independent. $Q_+^{(2)}$ contains data for patients enrolled in Stage I at time from $T_0$ to $T_2$, and $Q_+^{(2B)}$ contains data for patients enrolled in Stage II at time from $T_1$ to $T_2$.

$$E(Q_+^{(2B)}) = E\left(\sum_{k|T_1<t_k\leq T_2} (d_{+Tk} - \frac{d_{+k}n_{+Tk}}{n_{+k}})\right) = r(1-r)d_+^{(2B)}\log\theta_+,$$

and

$$Var(Q_+^{(2B)}) = (\sigma_+^{(2B)})^2 = r(1-r)d_+^{(2B)}.$$

Under $H_0$,

$$\overline{Z}_+ = \frac{\overline{Q}_+}{\sqrt{Var(\overline{Q}_+)}} = \frac{\overline{Q}_+}{\overline{\sigma}_+} \xrightarrow{d} N(0,1).$$

### 4.4.1 Correlation of the test statistics

To control the type I error rate and to estimate the power, we need to calculate the correlation matrix of $\left(Z^{(+)}, Z_+^{(1)}, Z_-^{(1)}, Z, Z_+\right)$ with respect to scenario $II_A$ and that of $\left(Z^{(1)}, Z_+^{(1)}, Z_-^{(1)}, \overline{Z}_+\right)$ with respect to scenario $II_B$, and use it in the asymptotic joint multivariate normal distribution.

1) For scenario $II_A$, between the Z-statistics of Stage I, we have

$$Cov(Z^{(1)}, Z_+^{(1)}) = \frac{p}{\sigma^{(1)}} = \frac{p}{\sqrt{p^2+(1-p)^2}},$$
$$Cov(Z^{(1)}, Z_-^{(1)}) = \frac{(1-p)}{\sigma^{(1)}} = \frac{1-p}{\sqrt{p^2+(1-p)^2}}.$$

Between the Z-statistics of Stage I and Stage II,

$$Cov(Z, Z_+^{(1)}) = \frac{p\sigma_+^{(1)}}{\sqrt{p^2+(1-p)^2}\sigma_+} = \frac{p\sigma_+^{(1)}}{\sigma\sigma_+},$$
$$Cov(Z, Z_-^{(1)}) = \frac{(1-p)\sigma_-^{(1)}}{\sqrt{p^2+(1-p)^2}\sigma_-} = \frac{(1-p)\sigma_-^{(1)}}{\sigma\sigma_-},$$

$$Cov(Z, Z^{(1)}) = \frac{p^2\sigma_+^{(1)}/\sigma_+ + (1-p)^2\sigma_-^{(1)}/\sigma_-}{p^2+(1-p)^2} = \frac{p^2\sigma_1^{(1)}/\sigma_+ + (1-p)^2\sigma_-^{(1)}/\sigma_-}{\sigma^2},$$

$$Cov(Z_+, Z^{(1)}) = \frac{p\sigma_+^{(1)}}{\sqrt{p^2+(1-p)^2}\sigma_+} = \frac{p\sigma_+^{(1)}}{\sigma\sigma_+},$$

$$Cov(Z_+, Z_+^{(1)}) = \frac{\sigma_+^{(1)}}{\sigma_+},$$

$$Cov(Z_+, Z) = \frac{p}{\sqrt{p^2+(1-p)^2}} = \frac{p}{\sigma}.$$

In summary, the covariance matrix of the $Z$-statistics under scenario $II_A$ is:

$$Cov\left(\left(Z^{(1)}, Z_+^{(1)}, Z_-^{(1)}, Z, Z_+\right)^T\right)$$
$$= \begin{bmatrix} 1 & \frac{p}{\sigma^{(1)}} & \frac{(1-p)}{\sigma^{(1)}} & \frac{p^2\sigma_+^{(1)}/\sigma_+ + (1-p)^2\sigma_-^{(1)}/\sigma_-}{\sigma^2} & \frac{p\sigma_+^{(1)}}{\sigma\sigma_+} \\ & 1 & 0 & \frac{p\sigma_+^{(1)}}{\sigma\sigma_+} & \frac{\sigma_+^{(1)}}{\sigma_+} \\ & & 1 & \frac{(1-p)\sigma_-^{(1)}}{\sigma\sigma_-} & 0 \\ & & & 1 & \frac{p}{\sigma} \\ & & & & 1 \end{bmatrix}$$

2) For scenario $II_B$, between the $Z$-statistics of Stage I, we have

$$Cov(Z^{(1)}, Z_+^{(1)}) = \frac{p}{\sqrt{p^2+(1-p)^2}},$$

$$Cov(Z^{(1)}, Z_-^{(1)}) = \frac{1-p}{\sqrt{p^2+(1-p)^2}}.$$

Between the $Z$-statistics from Stage I and Stage II,

$$Cov(\overline{Z}_+, Z_+^{(1)}) = \frac{\sigma_+^{(1)}}{\overline{\sigma}_1},$$

$$Cov(\overline{Z}_+, Z^{(1)}) = \frac{p\sigma_+^{(1)}}{\sqrt{p^2+(1-p)^2}\overline{\sigma}_1}.$$

In summary, the covariance matrix of the Z-statistics under scenario $II_B$ is

$$Cov\left(\left(Z^{(1)}, Z_+^{(1)}, Z_-^{(1)}, \overline{Z}_+\right)^T\right)$$
$$= \begin{bmatrix} 1 & \frac{p}{\sqrt{p^2+(1-p)^2}} & \frac{1-p}{\sqrt{p^2+(1-p)^2}} & \frac{p\sigma_+^{(1)}}{\sqrt{p^2+(1-p)^2}\overline{\sigma}_+} \\ & 1 & 0 & \frac{\sigma_+^{(1)}}{\overline{\sigma}_+} \\ & & 1 & 0 \\ & & & 1 \end{bmatrix}.$$

Summarize all, the correlation matrix for standardized log-rank statistics is

$$Cov\left(\left(Z^{(1)}, Z_+^{(1)}, Z_-^{(1)}, Z, Z_+, \overline{Z}_+\right)^T\right)$$
$$= \begin{bmatrix} 1 & \frac{p}{\sigma^{(1)}} & \frac{(1-p)}{\sigma^{(1)}} & \frac{p^2\sigma_+^{(1)}/\sigma_+ + (1-p)^2\sigma_-^{(1)}/\sigma_-}{\sigma^2} & \frac{p\sigma_+^{(1)}}{\sigma\sigma_+} & \frac{p\sigma_+^{(1)}}{\sqrt{p^2+(1-p)^2}\overline{\sigma}_+} \\ & 1 & 0 & \frac{p\sigma_+^{(1)}}{\sigma\sigma_+} & \frac{\sigma_+^{(1)}}{\sigma_+} & \frac{\sigma_+^{(1)}}{\overline{\sigma}_+} \\ & & 1 & \frac{(1-p)\sigma_-^{(1)}}{\sigma\sigma_-} & 0 & 0 \\ & & & 1 & \frac{p}{\sigma} & * \\ & & & & 1 & * \\ & & & & & 1 \end{bmatrix}.$$

### 4.4.2 Type I error $\alpha$ allocation and critical values

For our two-stage enrichment design, we split the overall alpha (e.g., $\alpha = 0.025$) between the two stages, following a similar method described by Lin et al. (2019). In Stage I, $\alpha_1$, a

fraction of the overall alpha, is allocated to test the global hypothesis $H_0$.

$$\begin{aligned}\alpha_1 &= P(\text{Reject } H_0|H_0)\\ &= P_0(Z^{(1)} < -c_1 \text{ or } Z_+^{(1)} < -c_2)\\ &= P_0(Z^{(1)} < -c_1) + P_0(Z^{(1)} \geq -c_1 \text{ or } Z_+^{(1)} < -c_2)\\ &= \alpha_{1a} + \alpha_{1b}.\end{aligned}$$

The critical value $c_1$ is obtained by allocating $\alpha_{1a}$, a portion of $\alpha_1$, for testing $H_{0a}$. Then $c_2$ can be solved for testing $H_{0+}$ in the above equation.

For Stage II, the overall alpha$\alpha - \alpha_1$, is left to spend between the mutually exclusive scenarios $II_A$ and $II_B$. We allocate $\alpha_2$, a fraction of $\alpha - \alpha_1$ for the tests in scenario $II_A$ and the rest $\alpha_{2*} = \alpha - \alpha_1 - \alpha_2$ for scenario $II_B$.

$$\begin{aligned}\alpha - \alpha_1 = &P(\text{Accept } H_0 \text{ at Stage } I, \text{ Reject } H_0 \text{ at Stage } II \text{ in case of } II_A)\\ &+ P(\text{Accept } H_0 \text{ at Stage } I, \text{ Reject } H_{0+} \text{ at Stage } II \text{ in case of } II_B)\\ = &\alpha_2 + \alpha_{2*}.\end{aligned}$$

A futility criterion for marker-negative subjects will be used to determine $\alpha_2$ and $\alpha_{2*}$. This is done through a pre-specified threshold value $c_0$ for the test statistic $Z_-^{(1)}$ through futility probability $\mathcal{F}_p = P_0(Z_-^{(1)} \geq c_0)$. For example, if we want the futility probability to be 75% (50%), then from $P_0(Z_-^{(1)} \geq c_0) = 0.75$ (0.50), $c_0 = -0.6745$ (0.0).

**Scenario** $II_A$ : If testing for treatment effect on marker negative group is not futile, i.e., $Z_-^{(1)} < c_0$, the study is continued with both marker-status cohorts and test $H_0$ at Stage II. The alpha is controlled by

$$\begin{aligned}\alpha_2 &= P(\text{Accept } H_0 \text{ at Stage I, Reject } H_0 \text{ at Stage } II \text{ in case of II}_A)\\ &= P_0(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z < -b_1)\\ &\quad + P_0(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z \geq -b_1, Z_+ < -b_2)\\ &= \alpha_{2a} + (\alpha_2 - \alpha_{2a}).\end{aligned}$$

That is, the critical value $b_1$ is obtained by allocating a portion $\alpha_{2a}$ of $\alpha_2$ for testing $H_{0a}$. Then $b_2$ can be solved for testing $H_{0+}$ in the above equation.

**Scenario** $II_B$ : when the test for treatment effect on marker negative group is futile, i.e., $Z_-^{(1)} \geq c_0$, the study is continued with enriching marker-postive cohort only and test for $H_{0+}$ at Stage $II_B$. The Type I error is controlled by

$$\begin{aligned}\alpha_{2*} &= P(\text{Accept } H_0 \text{ at Stage I, Reject } H_{0+} \text{ at Stage } II \text{ in case of } II_B\ ) \\ &= P_0(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z_-^{(1)} \geq c_0, \overline{Z}_+ < -b_3),\end{aligned}$$

with the critical value $b_3$ solved numerically using the correlation matrix.

The strategy to allocating $\alpha$ to either $II_A$ or $II_B$ is an important design consideration (Lin, Shih, and Lu, 2019). To utilize full $\alpha$ in both $II_A$ and $II_B$ scenarios, we split $\alpha - \alpha_1$ into $II_A$ and $II_B$ as the follows. Since the trial will only be in one scenario or the other, it would be ideal to maximize the $\alpha$ in both scenarios. Toward this end, we can first rewrite $\alpha_2$ and $\alpha_{2*}$, respectively, as

$$\begin{aligned}\alpha_2 = [&P_0(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z < -b_1)|Z_-^{(1)} < c_0) \\ &+ P_0(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z \geq -b_1, Z_+ < -b_2)|Z_-^{(1)} < c_0]P_0(Z_-^{(1)} < c_0)\end{aligned}$$

and

$$\alpha_{2*} = P_0(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, \overline{Z}_+ < -b_3|Z_-^{(1)} \geq c_0)P_0(Z_-^{(1)} \geq c_0).$$

Next, if we split $\alpha - \alpha_1$ into $II_A$ and $II_B$ with the same proportion as the odds of $P_0(Z_0^{(1)} < c_0)$ to $P_0(Z_-^{(1)} \geq c_0)$, i.e.,

$$\frac{\alpha_2}{\alpha_{2*}} = \frac{P_0(Z_-^{(1)} < c_0)}{P_0(Z_-^{(1)} \geq c_0)} = \frac{1 - \mathcal{F}_p}{\mathcal{F}_p},$$

then we have

$$\begin{aligned}\alpha - \alpha_1 = [&P_0(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z < -b_1)|Z_-^{(1)} < c_0) \\ &+ P_0(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z \geq -b_1, Z_+ < -b_2)|Z_-^{(1)} < c_0],\end{aligned}$$

and

$$\alpha - \alpha_1 = P_0(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, \overline{Z}_+ < -b_3 | Z_-^{(1)} \geq c_0).$$

This indicates that with the above alpha allocation strategy, the corresponding type I error in Stage II is $\alpha - \alpha_1$ for either $II_A$ or $II_B$. When the odds of nonfutility versus futility $\frac{P_0(Z_-^{(1)} < c_0)}{P_0(Z_-^{(1)} \geq c_0)}$ is predetermined, the critical values can be calculated.

Table 4.1 gives some examples for the calculated critical values based on some commonly used design parameters.

Table 4.1: Critcal Values When $\alpha = 0.025,\ \alpha_+ = 0.004, \alpha_2 = 0.021, r = 0.5$

| $\mathcal{F}_p$ | Info | p | $c_0$ | $c_1$ | $c_2$ | $b_1$ | $b_2$ | $b_3$ |
|---|---|---|---|---|---|---|---|---|
| 0.5 | 0.3 | 0.3 | 0 | -2.878 | -2.866 | -2.517 | -2.250 | -2.023 |
| | | 0.4 | 0 | -2.878 | -2.848 | -2.506 | -2.215 | -2.021 |
| | | 0.5 | 0 | -2.878 | -2.816 | -2.487 | -2.165 | -2.019 |
| | 0.5 | 0.3 | 0 | -2.878 | -2.866 | -2.509 | -2.234 | -2.156 |
| | | 0.4 | 0 | -2.878 | -2.848 | -2.503 | -2.196 | -2.012 |
| | | 0.5 | 0 | -2.878 | -2.816 | -2.488 | -2.142 | -2.009 |
| 0.75 | 0.3 | 0.3 | -0.674 | -2.878 | -2.866 | -2.701 | -2.245 | -2.023 |
| | | 0.4 | -0.674 | -2.878 | -2.848 | -2.677 | -2.206 | -2.021 |
| | | 0.5 | -0.674 | -2.878 | -2.816 | -2.636 | -2.153 | -2.019 |
| | 0.5 | 0.3 | -0.674 | -2.878 | -2.866 | -2.716 | -2.224 | -2.015 |
| | | 0.4 | -0.674 | -2.878 | -2.848 | -2.698 | -2.180 | -2.012 |
| | | 0.5 | -0.674 | -2.878 | -2.816 | -2.663 | -2.122 | -2.009 |

### 4.4.3 Global and marginal power

Under the alternative, we have

$$Z^{(1)} \sim AN(\frac{pd_+^{(1)} r(1-r)\log\theta_+ + (1-p)d_-^{(1)} r(1-r)\log\theta_-}{\sqrt{p^2 d_+^{(1)} r(1-r) + (1-p)^2 d_-^{(1)} r(1-r)}}, 1) \ ,$$

$$Z_+^{(1)} \sim AN(\sqrt{d_+^{(1)} r(1-r)}\log\theta_+, 1) \ ,$$

$$Z_-^{(1)} \sim AN(\sqrt{d_-^{(1)} r(1-r)}\log\theta_-, 1),$$

$$Z \sim AN(\frac{p(d_+^{(2)} + d_+^{(2A)}) r(1-r)\log\theta_+ + (1-p)(d_-^{(2)} + d_-^{(2A)}) r(1-r)\log\theta_-}{\sqrt{p^2(d_+^{(2)} + d_+^{(2A)}) r(1-r) + (1-p)^2(d_-^{(2)} + d_-^{(2A)}) r(1-r)}}, 1) \ ,$$

$$Z_1 \sim AN(\sqrt{(d_+^{(2)} + d_+^{(2A)})r(1-r)}\log\theta_+, 1)\ ,$$
$$\overline{Z}_1 \sim AN(\sqrt{(d_+^{(2)} + d_+^{(2B)})r(1-r)}\log\theta_+, 1). \tag{4.1}$$

The global power is

$$\begin{aligned}
1-\beta &= P(\text{Reject } H_0|H_1) \\
&= P(\text{Reject } H_{0a} \text{ or Reject } H_{0+}|H_1) \\
&= P_1(Z^{(1)} < -c_1 \textit{ or } Z_+^{(1)} < -c_2) \\
&\quad + P_1(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z < -b_1) \\
&\quad + P_1(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z \geq -b_1, Z_+ < -b_2) \\
&\quad + P_1(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z_-^{(1)} \geq c_0, \overline{Z}_+ < -b_3) \\
&= p_1 + p_{2a} + p_{2+} + \overline{p}_{2+},
\end{aligned}$$

where

$$\begin{aligned}
p_1 &= P(\text{Reject } H_{0a} \text{ or Reject } H_{0+} \text{ at Stage } I|H_1) \\
&= P_1(Z^{(1)} < -c_1 \textit{ or } Z_+^{(1)} < -c_2),
\end{aligned}$$

$$\begin{aligned}
p_{2a} &= P(\text{Accept } H_0 \text{ at Stage } I \text{ and Reject } H_{0a} \text{ at Stage } II_A\ |H_1) \\
&= P_1(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z < -b_1),
\end{aligned}$$

$$\begin{aligned}
p_{2+} &= P(\text{Accept } H_0 \text{ at Stage } I, \text{ Accept } H_{0a} \text{ and at Stage } II_A \text{ but Reject } H_{0+} \text{ at Stage } II_A\ |H_1) \\
&= P_1(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z \geq -b_1, Z_+ < -b_2),
\end{aligned}$$

$$\begin{aligned}
\overline{p}_{2+} &= P(\text{Accept } H_0 \text{ at Stage } I \text{ and Reject } H_{0+} \text{ at Stage } II_B\ |H_1) \\
&= P_1(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z_-^{(1)} \geq c_0, \overline{Z}_+ < -b_3).
\end{aligned}$$

Power for testing the treatment effect in the overall cohort is

$$\begin{aligned}
1-\beta_a &= P(\text{Reject } H_{0a}|H_1) \\
&= P(\text{Reject } H_{0a} \text{ at Stage } I) + \text{P}(\text{ Reject } H_{0a} \text{ at Stage } II_A) \\
&= P_1(Z^{(1)} < -c_1) + p_{2a} \\
&= p_{1a} + p_{2a}.
\end{aligned}$$

Power for testing the treatment effect in the marker-positive cohort is

$$\begin{aligned}
1-\beta_+ &= P(\text{Reject } H_{0+}|H_1) \\
&= P(\text{Reject } H_{0+} \text{ at Stage } I) + P(\text{Reject } H_{0+} \text{ at Stage } II_A) \\
&\quad + P(\text{Reject } H_{0+} \text{ at Stage } II_B) \\
&= P_1(Z_+^{(1)} < -c_2) + P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z_1 < -b_2) \\
&\quad + P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} \geq c_0, \overline{Z}_+ < -b_3).
\end{aligned}$$

### 4.4.4 Sample size calculations

Given the global type I error and power, assuming that we have the estimated prevalence rate $p$, the sample size need to detect the treatment effect can be found based on the formulas in Section 4.4.2 and Section 4.4.3 after we specify the design parameters in these sections.

To be more specific, given the design parameters shown in Section 4.4.2, we find the critical values first based on the formulas in Section 4.4.2. Notice that the critical values are based on the distributions under null hypothesis and do not depend on the sample size. Next using these critical values and the formulas shown in Section 4.4.3, we can determine the sample size needed to achieve the specified power of specific type (global, overall or marker positive), through optimization programming algorithms to find the solution and round up the the nearest integer. R-code is developed to calculate the needed sample size (or the number of events) and results are illustrated in Chapter 5.

# Chapter 5

# Numeric examples

## 5.1 Simulation setup

Consider a total of $N = 1000$ subjects for Stage I and Stage II, with a prevalence rate of $p = 0.3$, 0.4, or 0.5 for biomarker positive ($S+$). Randomization to active treatment $T$ or placebo treatment $C$ will be stratified by biomarker status with a randomization ratio of $r$ to $1 - r$, to $T$ or $C$, respectively. In Stage I, we plan to enroll $tN$ ($t = 0.7$) subjects (i.e., 7 months from the start of the study is needed for enrollment). A decision is made at interim analysis $T_1$ (information time: $info = 0.3$ *or* $info = 0.5$), to continue enrolling biomarker-unselected subjects (under scenario $II_A$) or to enroll only biomarker positive subjects under scenario $II_B$, with a plan to enroll additional $(1 - t)N$ subjects in Stage II. After the recruitment is complete, the study will follow-up to time $T_2$, when final analysis is performed. In addition, we assume that subject recruitment follows a uniform distribution, and survival times are exponentially distributed.

The primary composite hypothesis is about survival in marker positive subjects and/or in overall subjects. The null hypothesis is that the hazard rates for treatment and placebo group are equal within marker positive subjects and within the overall population. The hazard rate for active treated $S+$ group is $\lambda_{+T}$, and hazard rate for placebo treated $S+$ group is $\lambda_{+C}$; and the hazard rates for marker negative group $S-$ are $\lambda_{-T}$ for active treated, and $\lambda_{-C}$ for placebo, respectively.

We consider the simple case without biomarker misclassification ($\lambda_{sen} = \lambda_{spec} = 1.0$) and simulate the data for both Stage I and Stage II for 3000 times to obtain the observed global power through simulation. For each of 3000 simulation runs, data from patients enrolled in Stage I and from patients enrolled under Scenarios $II_A$ and $II_B$ are simulated independently. When the futility criteria is not met at interim analysis time in the final

data analysis time $T_1$, we use all data from patients enrolled in Stage I and data from patients enrolled in Stage II under Scenario $II_A$ in the final data analysis time $T_2$. If the futility criteria is met at interim analysis time $T_1$, we use data from marker positive patients enrolled in Stage I and data from patients enrolled in Stage II under Scenario $II_B$ in the final data analysis time $T_2$.

## 5.2 Nominal versus observed type I error rate under different scenarios

We check type I error rate under the following parameters: the hazard rate for treated $S+$ group is $\lambda_{+T} = 1/15$ and hazard rate for placebo treated $S+$ group is $\lambda_{+C} = 1/15$. There are no treatment effect for marker negative group $S-$, where the hazard rates are $\lambda_{-T} = \lambda_{-C} = 1/10$.

We simulate the trials data for 3000 times, to obtain the empirical type I error rate. For each of 3000 simulation runs, at interim analysis time $T_1$, we first calculated the log rank statistics for marker postie cohort, marker negative cohort, and the overall population. A decision is made at $T_1$ for Stage II enrollment: two mutually exclusive scenario $II_A$ or $II_B$.

The nominal and empirical type I error rates are shown in Table 5.1. As we can see from Table 5.1, the empirical type I error rate is close to the nominal type I error 0.025, across different prevalence rate (0.3, 0.4, 0.5), different information time of interim analysis (information time of 0.3 and 0.5), different futility probability (0.5 and 0.75), and different null hypothesis (all hazard rates are equal to 1/10 for biomaker negative cohort and equal to 1/15 for biomarker positive cohort, but with no treatment effect).

Table 5.1: The Nominal and Empirical Global Type I Error Rate for $H_0$ when $\lambda_{-T} = \lambda_{-C} = 1/10$, d=750, 3000 runs

| | | | | Empirical | |
|---|---|---|---|---|---|
| $\mathcal{F}_p$ | Info | $p$ | Nominal | $\lambda_{+T} = \lambda_{+C} = 1/10$ | $\lambda_{+T} = \lambda_{+C} = 1/15$ |
| 0.5 | 0.3 | 0.3 | 0.025 | 0.0233 | 0.0257 |
| | | 0.4 | 0.025 | 0.0247 | 0.0290 |
| | | 0.5 | 0.025 | 0.0237 | 0.0303 |
| | 0.5 | 0.3 | 0.025 | 0.0267 | 0.0267 |
| | | 0.4 | 0.025 | 0.0280 | 0.0280 |
| | | 0.5 | 0.025 | 0.0283 | 0.0283 |
| 0.75 | 0.3 | 0.3 | 0.025 | 0.0217 | 0.0233 |
| | | 0.4 | 0.025 | 0.0237 | 0.0277 |
| | | 0.5 | 0.025 | 0.0223 | 0.0293 |
| | 0.5 | 0.3 | 0.025 | 0.0260 | 0.0240 |
| | | 0.4 | 0.025 | 0.0240 | 0.0267 |
| | | 0.5 | 0.025 | 0.0247 | 0.0280 |

## 5.3 Theoretical versus empirical power under different scenarios

In this simulation, the critical values are based on $\alpha = 0.025, \alpha_1 = 0.004$, $r = 0.5$. The theoretical and empirical power for global, overall, and positive groups are shown in Table 5.2.

As we can see from Table 5.2, when there is treatment effect only in biomarker positive cohort and no treatment effect in biomarker negative cohort, the empirical powers (global power, overall power, and positive cohort power) are also close to the corresponding theoretical powers, across different prevalence rate (0.3, 0.4, 0.5), different information time of interim analysis (information time of 0.3 and 0.5), different futility probability (0.5 and 0.75). In the current set-up, since there is no treatment effect in biomarker negative cohort, the overall power (power for overall population) is low.

Table 5.2: The Theoretical and Empirical Power for $H_1$, $H_{1a}$ and $H_{1+}$ when $\lambda_{-T} = \lambda_{-C} = \lambda_{+C} = 1/10$ and $\lambda_{+T} = 1/15$ , N=1000, d=750, 3000 runs

| | | | Global Power | | Overall Power | | Positive Cohort Power | |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{F}_p$ | Info | $p$ | Theoretical | Empirical | Theoretical | Empirical | Theoretical | Empirical |
| 0.5 | 0.3 | 0.3 | 0.846 | 0.850 | 0.070 | 0.074 | 0.834 | 0.837 |
| | | 0.4 | 0.928 | 0.922 | 0.174 | 0.180 | 0.912 | 0.908 |
| | | 0.5 | 0.969 | 0.969 | 0.320 | 0.320 | 0.950 | 0.952 |
| | 0.5 | 0.3 | 0.849 | 0.858 | 0.066 | 0.078 | 0.836 | 0.843 |
| | | 0.4 | 0.946 | 0.923 | 0.113 | 0.175 | 0.931 | 0.906 |
| | | 0.5 | 0.970 | 0.970 | 0.321 | 0.326 | 0.950 | 0.951 |
| 0.75 | 0.3 | 0.3 | 0.884 | 0.848 | 0.042 | 0.056 | 0.874 | 0.838 |
| | | 0.4 | 0.944 | 0.923 | 0.103 | 0.150 | 0.930 | 0.909 |
| | | 0.5 | 0.975 | 0.969 | 0.195 | 0.296 | 0.956 | 0.952 |
| | 0.5 | 0.3 | 0.849 | 0.857 | 0.066 | 0.057 | 0.836 | 0.845 |
| | | 0.4 | 0.930 | 0.923 | 0.164 | 0.149 | 0.913 | 0.908 |
| | | 0.5 | 0.976 | 0.971 | 0.239 | 0.302 | 0.957 | 0.953 |

Figures 5.1, 5.2, and 5.3 show the contour plots of power surfaces for global (testing $H_1$), overall population (testing $H_{1a}$) and marker-positive population (testing $H_{1+}$) hypotheses, respectively, across $-0.10 \geq \delta \geq -0.40$ and $-0.10 \geq \delta_+ \geq -0.40$ by $n$ and $p$ assuming $\alpha = 0.025, \alpha_1 = 0.004, w_+ = p, w_- = 1-p, c_0 = 0, \delta = p\delta_+ + (1-p)\delta_-$, $r = 0.5$.

From Figure 5.1, the global power increases with increasing treatment effect for overall population and/or positive population (decreasing $\delta$ and/or decreasing $\delta_+$).

From Figure 5.2, the power for overall population increases with increasing treatment effect for overall population when treatment effect for positive population is fixed (decreasing $\delta$).

From Figure 5.3, the power for positive subgroup increases with increasing treatment effect for positive population (decreasing $\delta_+$) but decreases with increasing treatment effect for overall population (decreasing $\delta$).

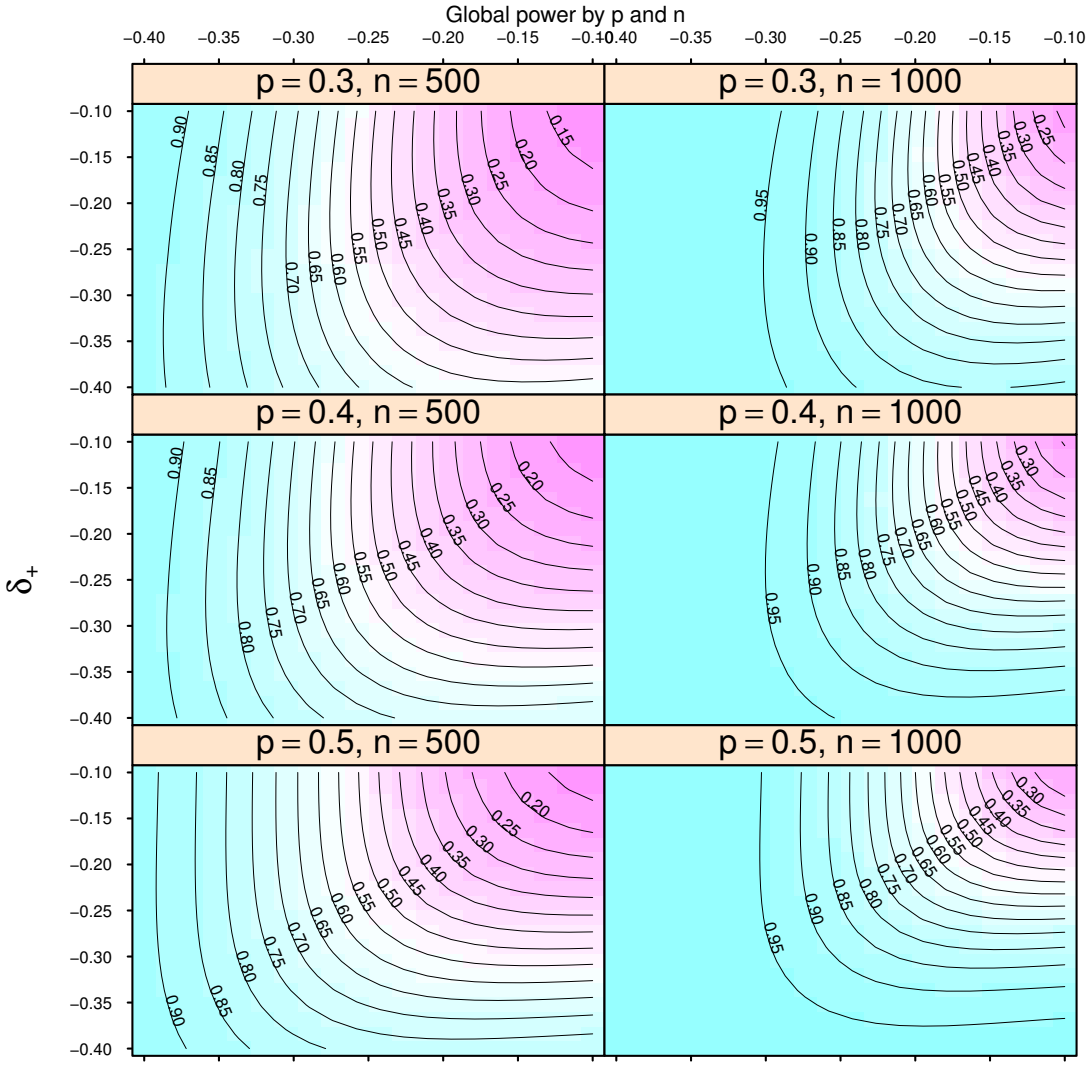



Figure 5.1: Contour plot of global power surface

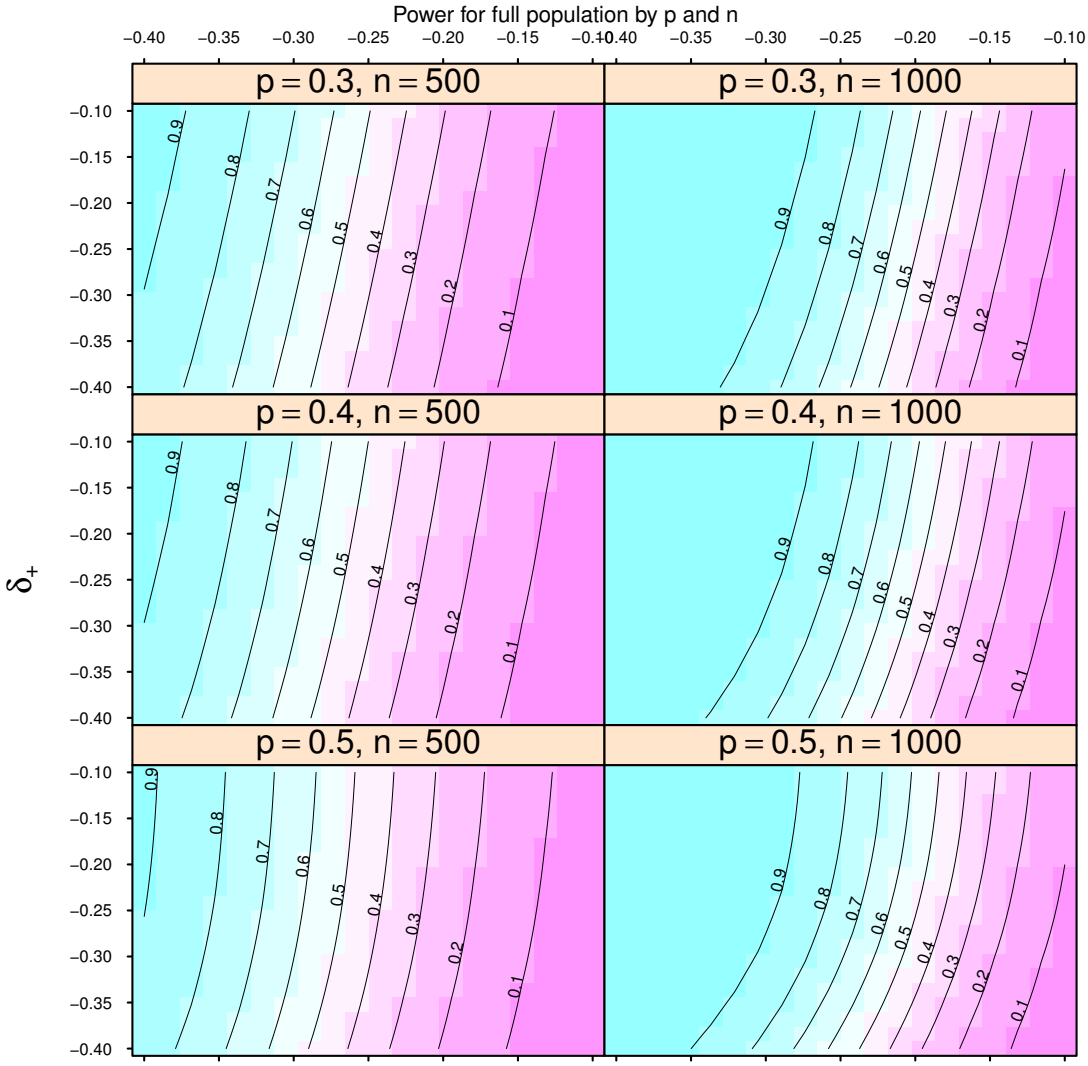



Figure 5.2: Contour plot of power surface for overall population

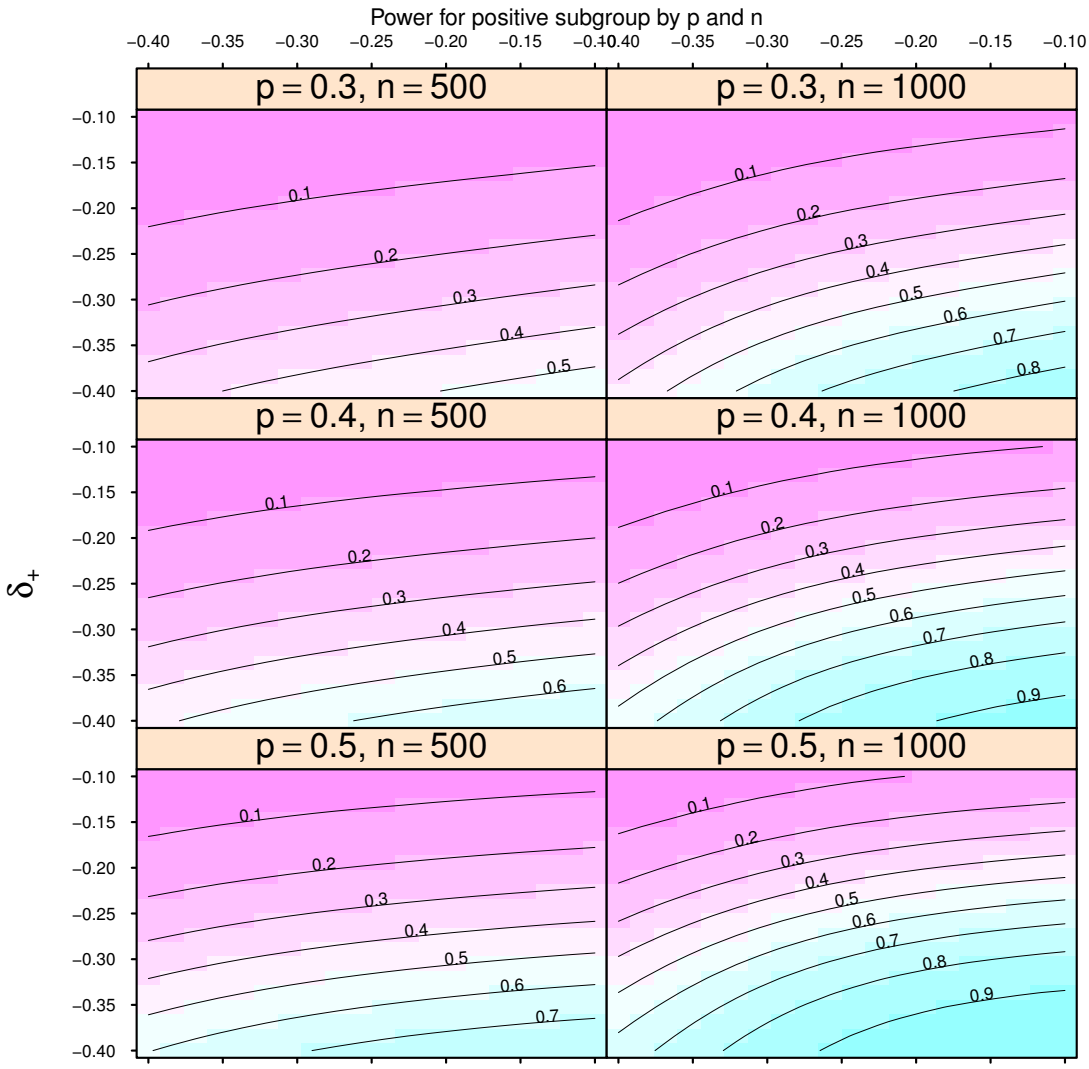



Figure 5.3: Contour plot of power surface for biomarker positive subgroup

Table 5.3 shows the sample size needed to achieve specified global and marginal power for biomarker positive subgroup in some commonly used situations. In the table, we assume the same $\alpha$ allocation, $\delta = -0.15$ (the log-hazard ratio for the overall cohort), and $\delta_+ = -0.3, -0.4$, or $-0.5$ (the log-hazard ratio for the marker positive cohort).

From the table, for example, if a target of 90% global power for testing $H_1$ is requested when the prevalence rate is 0.4 and interim analysis is performed at information time of 0.5, a total sample size of 871 is needed. If the target of 80% biomarker positive marginal power is requested for testing $H_{1+}$ when the prevalence rate is 0.4 and interim analysis is

performed at information time of 0.5, a total sample size of 678 is needed.

Table 5.3: Total Sample Size to Achieve Specified Global Power $H_1$ when $\alpha = 0.025, \alpha_1 = 0.004, \mathcal{F}_p = 0.5$, $r = 0.5$, and $\delta = p \log \theta_+ + (1-p) \log \theta_- = -0.15$

| | | Global Power under $H_1$ | | | | Marginal Power under $H_{1+}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 90% power | | 80% power | | 90% power | | 80% power | |
| | | Information | | Information | | Information | | Information | |
| $p$ | $\delta_+$ | 0.3 | 0.5 | 0.3 | 0.5 | 0.3 | 0.5 | 0.3 | 0.5 |
| 0.3 | -0.3 | 2005 | 2019 | 1492 | 1503 | 3056 | 3598 | 1907 | 2061 |
| | -0.4 | 1183 | 1184 | 879 | 878 | 1307 | 1328 | 946 | 957 |
| | -0.5 | 744 | 735 | 556 | 547 | 774 | 768 | 574 | 584 |
| 0.4 | -0.3 | 1592 | 1565 | 1174 | 1172 | 1860 | 1929 | 1316 | 1344 |
| | -0.4 | 879 | 871 | 661 | 652 | 920 | 913 | 685 | 678 |
| | -0.5 | 555 | 541 | 416 | 408 | 563 | 552 | 424 | 413 |
| 0.5 | -0.3 | 1271 | 1255 | 950 | 943 | 1367 | 1364 | 1004 | 1000 |
| | -0.4 | 701 | 688 | 529 | 519 | 717 | 703 | 539 | 528 |
| | -0.5 | 442 | 433 | 336 | 326 | 445 | 472 | 339 | 329 |

# Part II

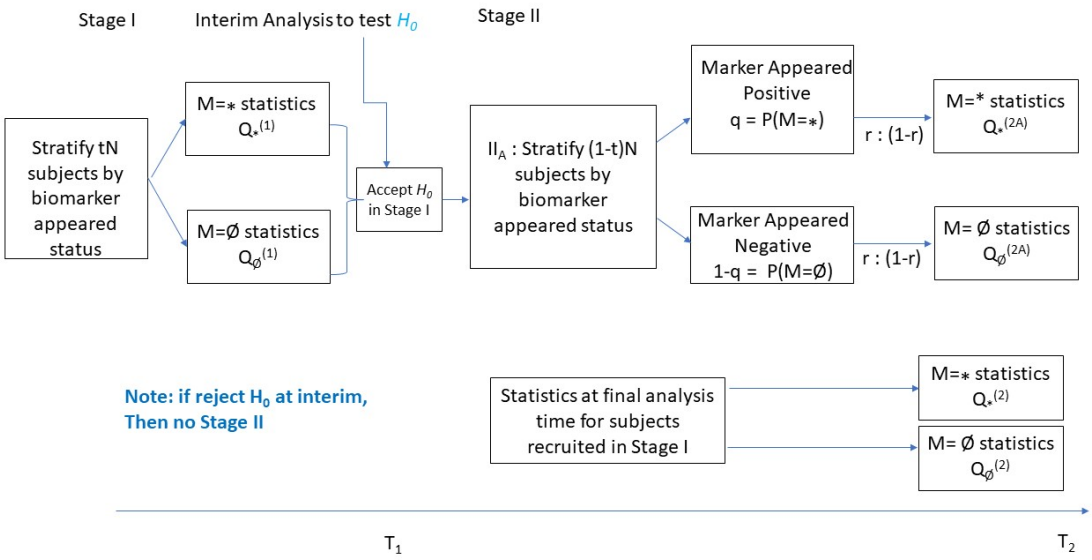
# Two-Stage Stratified Design with Misclassification Adjustments

# Chapter 6

# Methods for two-stage stratified design with biomarker misclassification adjustment

## 6.1 Design diagram

The following shows the diagram of the two-stage stratified design based on marker appearance status.




where $Q_i^j$ is log rank statistic, with $i = *, \phi$ (marker-appeared positive, appeared negative) and $j$ =1, 2, 2$A$ for analysis time/scenarios, $M$ denotes the marker-appeared status, $t$ represents the fraction of the total sample size $N$ that is allocated to Stage I, $q$ is the prevalence rate of biomarker-appeared positive, and $r$ is the randomization proportion to active treatment group. The definition of $Q_i^j$ can be found in 6.4.2 and 6.4.3.

## 6.2 Positive and negative predictive values

Denote true marker positive prevalence $P(D = +) = p$; true marker index $D = +$ or $-$, for positive or negative true marker status, respectively; marker-appeared status index $M = *$ or $\phi$, for appeared positive or appeared negative status, respectively. The stratified randomization designs are based on marker-appeared status with sensitivity $\lambda_{sen}$ and specificity $\lambda_{spec}$.

Probability of $M = *$, where marker-appeared status is positive, is

$$q = P(M = *) = p\lambda_{sen} + (1 - p)(1 - \lambda_{spec}).$$

The positive predictive value (PPV) $= P(D = +|M = *)$ is

$$\tau = \frac{p\lambda_{sen}}{p\lambda_{sen} + (1 - p)(1 - \lambda_{spec})} = \frac{p\lambda_{sen}}{q}.$$

The negative predictive value (NPV) $= P(D = -|M = \phi)$ is

$$\eta = \frac{(1 - p)\lambda_{spec}}{1 - P(M = *)} = \frac{(1 - p)\lambda_{spec}}{1 - q}.$$

Note: when $\lambda_{sen} = \lambda_{spec} = 1$, $q = p$ and $\tau = \eta = 1$.

Let the ratio of randomization assignment of active treatment to control be $r$ to $1 - r$, where $0 < r < 1$.

## 6.3 Difference between number of events and expected number of events when stratified by marker-appeared status under misclassification

Consider the difference between observed number of death $d_{iTk}$ and the expected number of death $E_{iTk}$ in active treatment group at $k^{th}$ event time in marker-appeared stratum $i$, where $i = *$ for marker-appeared positive and $i = \phi$ for marker-appeared negative, we have

$$d_{iTk} - E_{iTk} = d_{iTk} - \frac{d_{i+k}\overline{Y}_{iTk}}{\overline{Y}_{i+k}}.$$

Let $\overline{Y}_{*+k}$ be the total number of subjects, including both active-treated and control, at risk at $k^{th}$ event time.

The estimated variance of $d_{*Tk} - E_{*Tk}$ is

$$\widehat{Var}(d_{*Tk} - E_{*Tk}) = \widehat{Var}(d_{*Tk}) = \frac{\overline{Y}_{*Tk}\overline{Y}_{*Ck}d_{*+k}(\overline{Y}_{*+k} - d_{*+k})}{\overline{Y}^2_{*+k}(\overline{Y}_{*+k} - 1)}.$$

Similarly we have the formula for marker-appeared negative strata when $*$ is replaced by $\phi$.

Let $Q_i$ be the log-rank statistic in the marker-appeared status strata $i$

$$Q_i = \sum_{k=1}^{K}\left(d_{iTk} - \frac{d_{i+k}\overline{Y}_{iTk}}{\overline{Y}_{i+k}}\right)$$

where $i = *, \phi$. We have

$$\begin{aligned}\widehat{Var}(Q_i) &= \sum_{k=1}^{K}\widehat{Var}(d_{iTk}) \\ &= \sum_{k=1}^{K}\left\{\frac{\overline{Y}_{iTk}\overline{Y}_{iCk}d_{i+k}(\overline{Y}_{i+k} - d_{i+k})}{\overline{Y}^2_{i+k}(\overline{Y}_{i+k} - 1)}\right\}\end{aligned}$$

Under null,

$$\frac{Q_i}{\sqrt{\widehat{Var}(Q_i)}} \xrightarrow{d} N(0,1).$$

## 6.4 Asympototic distribution of adjusted log-rank statistics with marker misclassification

### 6.4.1 Expected number of events

For Stage I enrolled subjects in marker-appeared positive group, let $n^{(1)}_{*+} = \tau n^{(1)}_{*}$ be the expected total number of subjects in the marker-appeared positive group with true marker positive status, and $n^{(1)}_{*-} = (1-\tau)n^{(1)}_{*}$ be the expected number of subjects in marker-appeared positive group with true marker negative status, where $n^{(1)}_{*}$ is the expected number of subjects with marker-appeared positive status.

Similarly, let $n^{(1)}_{\phi+} = (1-\eta)n^{(1)}_{\phi}$ be the expected total number of subjects in marker-appeared negative group with true marker positive status and $n^{(1)}_{\phi-} = \eta n^{(1)}_{\phi}$ be the expected

total number of subjects in the marker-appeared negative group with true marker negative status, where $n_{\phi}^{(1)}$ is the expected number of subjects with marker-appeared negative status.

Let $n_i^{(1)}$ be the expected number of subjects enrolled at Stage I with true marker status $i$ , $i = +$ indicating true marker positive; $i = -$ indicating true marker negative.

At pre-determined interim analysis time $T_1$, let the probability to observe an event in true marker positive group (including both active and control treatment) be $\pi_+^{(1)}$, and let the probability to observe an event in true marker negative group (including both active and control treatment) be $\pi_-^{(1)}$. Then, the expected number of events at interim analysis time $T_1$ for marker-appeared positive group is

$$E(D_*^{(1)}) = \tau n_*^{(1)}\pi_+^{(1)} + (1-\tau)n_*^{(1)}\pi_-^{(1)},$$

the expected number of events at interim analysis time $T_1$ for marker-appeared negative group is

$$E(D_{\phi}^{(1)}) = (1-\eta)n_{\phi}^{(1)}\pi_+^{(1)} + \eta n_{\phi}^{(1)}\pi_-^{(1)}.$$

### 6.4.2 Adjusted log rank statistics at interim analysis

Let $Q_*^{(1)}$ and $Q_{\phi}^{(1)}$ be the log rank statistics at the interim analysis for marker-appeared positive strata and marker-appeared negative strata, respectively, then

$$E(Q_*^{(1)}) = r(1-r)\tau n_*^{(1)}\pi_+^{(1)}\log\frac{\lambda_{+T}}{\lambda_{+C}} + r(1-r)(1-\tau)n_*^{(1)}\pi_-^{(1)}\log\frac{\lambda_{-T}}{\lambda_{-C}}, \tag{6.1}$$

$$E(Q_{\phi}^{(1)}) = r(1-r)(1-\eta)n_{\phi}^{(1)}\pi_+^{(1)}\log\frac{\lambda_{+T}}{\lambda_{+C}} + r(1-r)\eta n_{\phi}^{(1)}\pi_-^{(1)}\log\frac{\lambda_{-T}}{\lambda_{-C}}. \tag{6.2}$$

From Equations 6.1 and 6.2, we get

$$\begin{aligned} & r(1-r)n_+^{(1)}\pi_+^{(1)}\log\frac{\lambda_{+T}}{\lambda_{+C}} \\ &= \frac{\tau n_*^{(1)} + (1-\eta)n_{\phi}^{(1)}}{n_*^{(1)}n_{\phi}^{(1)}}\frac{\eta n_{\phi}^{(1)}E(Q_*^{(1)}) - (1-\tau)n_*^{(1)}E(Q_{\phi}^{(1)})}{\tau+\eta-1} \\ &= \frac{\tau q + (1-\eta)(1-q)}{q(1-q)}\frac{\eta(1-q)E(Q_*^{(1)}) - (1-\tau)qE(Q_{\phi}^{(1)})}{\tau+\eta-1}, \end{aligned} \tag{6.3}$$

$$
\begin{aligned}
& r(1-r)n_-^{(1)}\pi_-^{(1)}\log\frac{\lambda_{-T}}{\lambda_{-C}} \\
&\qquad = \frac{(1-\tau)n_*^{(1)}+\eta n_\phi^{(1)}}{n_*^{(1)}n_\phi^{(1)}}\frac{-(1-\eta)n_\phi^{(1)}E(Q_*^{(1)})+\tau n_*^{(1)}E(Q_\phi^{(1)})}{\tau+\eta-1} \\
&\qquad = \frac{(1-\tau)q+\eta(1-q)}{q(1-q)}\frac{-(1-\eta)(1-q)E(Q_*^{(1)})+\tau qE(Q_\phi^{(1)})}{\tau+\eta-1}. \qquad (6.4)
\end{aligned}
$$

Let

$$
Q_+^{(1)} = \frac{\tau q+(1-\eta)(1-q)}{q(1-q)}\frac{\eta(1-q)Q_*^{(1)}-(1-\tau)qQ_\phi^{(1)}}{\tau+\eta-1}, \qquad (6.5)
$$

$$
Q_-^{(1)} = \frac{(1-\tau)q+\eta(1-q)}{q(1-q)}\frac{-(1-\eta)(1-q)Q_*^{(1)}+\tau qQ_\phi^{(1)}}{\tau+\eta-1}. \qquad (6.6)
$$

Then we have

$$
\begin{aligned}
E(Q_+^{(1)}) &= r(1-r)n_+^{(1)}\pi_+^{(1)}\log\frac{\lambda_{+T}}{\lambda_{+C}}, \\
E(Q_-^{(1)}) &= r(1-r)n_-^{(1)}\pi_-^{(1)}\log\frac{\lambda_{-T}}{\lambda_{-C}},
\end{aligned}
$$

and their variances

$$
Var(Q_+^{(1)}) = \left(\frac{\tau q+(1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2\left((1-q)^2\eta^2Var(Q_*^{(1)})+q^2(1-\tau)^2Var(Q_\phi^{(1)})\right),
$$

$$
Var(Q_-^{(1)}) = \left(\frac{(1-\tau)q+\eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2\left((1-q)^2(1-\eta)^2Var(Q_*^{(1)})+q^2\tau^2Var(Q_\phi^{(1)})\right).
$$

Therefore we use $Q_+^{(1)}$ and $Q_-^{(1)}$, unbiased estimators for the true effects, as test statistics to test treatment effects on true marker positive and negative groups, respectively. This also implies that the adjusted log rank statistic for the overall population (based on true marker positive group and true marker negative group) at the interim analysis time $T_1$ is

$$
Q^{(1)} = pQ_+^{(1)}/\sigma_+^{(1)}+(1-p)Q_-^{(1)}/\sigma_-^{(1)},
$$

and

$$Var(Q^{(1)}) = p^2+(1-p)^2-2p(1-p)(\eta(1-\eta)(1-q)^2Var(Q_*^{(1)})+\tau(1-\tau)q^2Var(Q_\phi^{(1)}))/(\sigma_+^{(1)}\sigma_-^{(1)}).$$

The adjusted standardized log rank statistic for true marker positive group at interim analysis $T_1$ is

$$Z_+^{(1)} = \frac{Q_+^{(1)}}{\sqrt{Var(Q_+^{(1)})}} = \frac{Q_+^{(1)}}{\sigma_+^{(1)}},$$

where

$$\sigma_+^{(1)} = \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)\sqrt{(1-q)^2\eta^2Var(Q_*^{(1)}) + q^2(1-\tau)^2Var(Q_\phi^{(1)})}.$$

Similarly the adjusted standardized log rank statistic for true marker negative group at interim analysis $T_1$ is

$$Z_-^{(1)} = \frac{Q_-^{(1)}}{\sqrt{Var(Q_-^{(1)})}} = \frac{Q_-^{(1)}}{\sigma_-^{(1)}},$$

where

$$\sigma_-^{(1)} = \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)\sqrt{(1-q)^2(1-\eta)^2Var(Q_*^{(1)}) + q^2\tau^2Var(Q_\phi^{(1)})}.$$

The adjusted standardized log rank statistic for overall population at interim analysis $T_1$ is

$$Z^{(1)} = \frac{Q^{(1)}}{\sqrt{Var(Q^{(1)})}} = \frac{Q^{(1)}}{\sigma^{(1)}} = \frac{pZ_+^{(1)} + (1-p)Z_-^{(1)}}{\sigma^{(1)}},$$

where

$$(\sigma^{(1)})^2 = p^2+(1-p)^2-2p(1-p)(\eta(1-\eta)(1-q)^2Var(Q_*^{(1)})+\tau(1-\tau)q^2Var(Q_\phi^{(1)}))/(\sigma_+^{(1)}\sigma_0^{(1)}).$$

### 6.4.3 Adjusted log rank statistics at final analysis

**Log rank statistics for subjects recruited at Stage I at final analysis time $T_2$**

Let $Q_*^{(2)}$ and $Q_\phi^{(2)}$ be the log rank statistics at final analysis time $T_2$ for the subjects recruited

at Stage I (recruited between study start time $T_0$ and the interim analysis time $T_1$), then

$$E(Q_*^{(2)}) = r(1-r)\tau n_*^{(1)}\pi_+^{(2)}\log\frac{\lambda_{+T}}{\lambda_{+C}} + r(1-r)(1-\tau)n_*^{(1)}\pi_-^{(2)}\log\frac{\lambda_{-T}}{\lambda_{-C}}, \tag{6.7}$$

$$E(Q_\phi^{(2)}) = r(1-r)(1-\eta)n_\phi^{(1)}\pi_+^{(2)}\log\frac{\lambda_{+T}}{\lambda_{+C}} + r(1-r)\eta n_\phi^{(1)}\pi_-^{(2)}\log\frac{\lambda_{-T}}{\lambda_{-C}}. \tag{6.8}$$

From Equations 6.7 and 6.8, we get

$$r(1-r)n_+^{(1)}\pi_+^{(2)}\log\frac{\lambda_{+T}}{\lambda_{+C}} = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)}\frac{\eta(1-q)E(Q_*^{(2)}) - (1-\tau)qE(Q_\phi^{(2)})}{\tau+\eta-1},$$

$$r(1-r)n_-^{(1)}\pi_-^{(2)}\log\frac{\lambda_{-T}}{\lambda_{-C}} = \frac{(1-\tau q + \eta(1-q)}{q(1-q)}\frac{-(1-\eta)(1-q)E(Q_*^{(2)}) + \tau qE(Q_\phi^{(2)})}{\tau+\eta-1}.$$

Let

$$Q_+^{(2)} = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)}\frac{\eta(1-q)Q_*^{(2)} - (1-\tau)qQ_\phi^{(2)}}{\tau+\eta-1},$$

$$Q_-^{(2)} = \frac{(1-\tau q + \eta(1-q)}{q(1-q)}\frac{-(1-\eta)(1-q)Q_*^{(2)} + \tau qQ_\phi^{(2)}}{\tau+\eta-1},$$

Then we have

$$E(Q_+^{(2)}) = r(1-r)n_+^{(1)}\pi_+^{(2)}\log\frac{\lambda_{+T}}{\lambda_{+C}},$$

$$E(Q_-^{(2)}) = r(1-r)n_-^{(1)}\pi_-^{(2)}\log\frac{\lambda_{-T}}{\lambda_{-C}}.$$

and their variances

$$Var(Q_+^{(2)}) = \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2\left((1-q)^2\eta^2Var(Q_*^{(2)}) + q^2(1-\tau)^2Var(Q_\phi^{(2)})\right),$$

$$Var(Q_-^{(2)}) = \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2\left((1-q)^2(1-\eta)^2Var(Q_*^{(2)}) + q^2\tau^2Var(Q_\phi^{(2)})\right).$$

Therefore the test statistics $Q_+^{(2)}$ and $Q_-^{(2)}$, unbiased estimators for the true effects, are constructed as test statistics to test treatment effects at final analysis time $T_2$ on true marker positive and negative groups, respectively, for subjects enrolled at Stage I.

**Log rank statistics for subjects recruited at Stage II**

Let $Q_*^{(2A)}$ and $Q_\phi^{(2A)}$ be the log rank statistics at final analysis time $T_2$ for the subjects

recruited at Stage II, in addition, let the probability to observe an event at final analysis time $T_2$ in true marker positive group be $\widetilde{\pi}_+$ (including both active and control treatment), and let $\widetilde{\pi}_-$ be the probability to observe an event in true marker negative group (including both active and control treatment), then

$$E(Q_*^{(2A)}) = r(1-r)\tau n_*^{(2A)}\widetilde{\pi}_+ \log\frac{\lambda_{+T}}{\lambda_{+C}} + r(1-r)(1-\tau)n_*^{(2A)}\widetilde{\pi}_- \log\frac{\lambda_{-T}}{\lambda_{-C}}, \tag{6.9}$$

$$E(Q_\phi^{(2A)}) = r(1-r)(1-\eta)n_\phi^{(2A)}\widetilde{\pi}_+ \log\frac{\lambda_{+T}}{\lambda_{+C}} + r(1-r)\eta n_\phi^{(2A)}\widetilde{\pi}_- \log\frac{\lambda_{-T}}{\lambda_{-C}}. \tag{6.10}$$

From Equations 6.9 and 6.10, we get

$$r(1-r)n_+^{(2A)}\widetilde{\pi}_+ \log\frac{\lambda_{+T}}{\lambda_{+C}} = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)}\frac{\eta(1-q)E(Q_*^{(2A)}) - (1-\tau)qE(Q_\phi^{(2A)})}{\tau+\eta-1},$$

$$r(1-r)n_-^{(2A)}\widetilde{\pi}_- \log\frac{\lambda_{-T}}{\lambda_{-C}} = \frac{(1-\tau)q + \eta(1-q)}{q(1-q)}\frac{-(1-\eta)(1-q)E(Q_*^{(2A)}) + \tau qE(Q_\phi^{(2A)})}{\tau+\eta-1}.$$

Let

$$Q_+^{(2A)} = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)}\frac{\eta(1-q)Q_*^{(2A)} - (1-\tau)qQ_\phi^{(2A)}}{\tau+\eta-1},$$

$$Q_-^{(2A)} = \frac{(1-\tau)q + \eta(1-q)}{q(1-q)}\frac{-(1-\eta)(1-q)Q_*^{(2A)} + \tau qQ_\phi^{(2A)}}{\tau+\eta-1},$$

then we have

$$E(Q_+^{(2A)}) = r(1-r)n_+^{(2A)}\widetilde{\pi}_+ \log\frac{\lambda_{+T}}{\lambda_{+C}},$$
$$E(Q_-^{(2A)}) = r(1-r)n_-^{(2A)}\widetilde{\pi}_- \log\frac{\lambda_{-T}}{\lambda_{-C}},$$

and their variances are

$$Var(Q_+^{(2A)}) = \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left(((1-q)\eta)^2 Var(Q_*^{(2A)}) + (q(1-\tau))^2 Var(Q_\phi^{(2A)})\right),$$

$$Var(Q_-^{(2A)}) = \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left(((1-q)(1-\eta))^2 Var(Q_*^{(2A)}) + (q\tau)^2 Var(Q_\phi^{(2A)})\right).$$

Therefore the test statistics $Q_+^{(2A)}$ and $Q_-^{(2A)}$, unbiased estimators for the true effects, are constructed as test statistics to test treatment effects on true marker positive and negative groups, respectively, for subjects enrolled at Stage II. This also implies that at the final analysis time $T_2$, the following statistics $Q_+$ and $Q_-$ based on the rank statistics for testing treatment effects on marker-appeared status can be used for testing treatment effects on the true marker positive and negative groups, respectively, and $Q$ can be used for testing treatment effect on the overall population.

$$\begin{aligned} Q_+ &= Q_+^{(2)} + Q_+^{(2A)}, \\ Var(Q_+) &= \sigma_+^2 = Var(Q_+^{(2)}) + Var(Q_+^{(2A)}), \\ Q_- &= Q_-^{(2)} + Q_-^{(2A)}, \\ Var(Q_-) &= \sigma_-^2 = Var(Q_-^{(2)}) + Var(Q_-^{(2A)}), \end{aligned}$$

and

$$Q = pQ_+/\sigma_+ + (1-p)Q_-/\sigma_-,$$

$$Var(Q) = p^2 + (1-p)^2 - 2p(1-p)(\eta(1-\eta)(1-q)^2 Var(Q_*) + \tau(1-\tau)q^2 Var(Q_\phi))/(\sigma_+\sigma_-).$$

The standardized log rank statistic for the true marker-positive group at the final analysis time $T_2$ is

$$Z_+ = \frac{Q_+}{\sqrt{Var(Q_+)}} = \frac{Q_+^{(2)} + Q_+^{(2A)}}{\sigma_+}$$

where

$$\begin{aligned} \sigma_+^2 = &\left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left((1-q)^2\eta^2 Var(Q_*^{(2)}) + q^2(1-\tau)^2 Var(Q_\phi^{(2)})\right) \\ &+ \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left(((1-q)\eta)^2 Var(Q_*^{(2A)}) + (q(1-\tau))^2 Var(Q_\phi^{(2A)})\right). \end{aligned}$$

The standardized log rank statistic for the true marker-negative group is

$$Z_- = \frac{Q_-}{\sqrt{Var(Q_-)}} = \frac{Q_-}{\sigma_-^{(1)}},$$

where

$$\sigma_-^2 = \left(\frac{(1-\tau)q+\eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left((1-q)^2(1-\eta)^2 Var(Q_*^{(2)}) + q^2\tau^2 Var(Q_\phi^{(2)})\right) + \left(\frac{(1-\tau)q+\eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left(((1-q)(1-\eta))^2 Var(Q_*^{(2A)}) + (q\tau)^2 Var(Q_\phi^{(2A)})\right).$$

The standardized log rank statistic for overall population after the misclassification adjustment is

$$Z = \frac{Q}{\sqrt{Var(Q)}} = \frac{Q}{\sigma} = \frac{pZ_1 + (1-p)Z_0}{\sigma}$$

where

$$\sigma^2 = p^2 + (1-p)^2 - 2p(1-p)(\eta(1-\eta)(1-q)^2 Var(Q_*) + \tau(1-\tau)q^2 Var(Q_\phi))/(\sigma_+\sigma_-).$$

As we can see, to obtain log rank statistic for true marker positive group and true marker negative group shown by equation (6.3) and (6.4), we use both equation (6.1) and (6.2), since $\pi_+$ and $\pi_-$ are different for different analysis time point ($T_1$ or $T_2$). In this setting, we can adjust the misclassification without additional distributional assumptions. This is in contrast to the adaptive enrichment designs in Part III, for which if we restrict our enrolling criteria to only marker-appeared positive strata (after interim analysis), we only have one equation observed and can not solve two unknown parameters without additional assumptions, for true marker positive group and true marker negative group for the subjects who are enrolled in Stage II.

## 6.5 Correlations between test statistics

Let

$$\begin{aligned}
A &= \frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)},\\
B &= \frac{(1-\tau)q+\eta(1-q)}{q(1-q)(\tau+\eta-1)},\\
F_1 &= (1-q)^2\eta^2 Var(Q_*^{(1)}) + q^2(1-\tau)^2 Var(Q_\phi^{(1)}),\\
F_2 &= (1-q)^2\eta^2 Var(Q_*) + q^2(1-\tau)^2 Var(Q_\phi),
\end{aligned}$$

$$
\begin{aligned}
G_1 &= \eta(1-\eta)(1-q)^2 Var(Q_*^{(1)}) + \tau(1-\tau)q^2 Var(Q_\phi^{(1)}),\\
G_2 &= \eta(1-\eta)(1-q)^2 Var(Q_*) + \tau(1-\tau)q^2 Var(Q_\phi),\\
H_1 &= (1-q)^2(1-\eta)^2 Var(Q_*^{(1)}) + q^2\tau^2 Var(Q_\phi^{(1)}),\\
H_2 &= (1-q)^2(1-\eta)^2 Var(Q_*) + q^2\tau^2 Var(Q_\phi),\\
C_1 &= \frac{p}{\sqrt{p^2+(1-p)^2-2p(1-p)G_1}},\\
D_1 &= \frac{1-p}{\sqrt{p^2+(1-p)^2-2p(1-p)G_1}},\\
C_2 &= \frac{p}{\sqrt{p^2+(1-p)^2-2p(1-p)G_2}},\\
D_2 &= \frac{1-p}{\sqrt{p^2+(1-p)^2-2p(1-p)G_2}},\\
E &= C_1C_2\frac{\sigma_+^{(1)}}{\sigma_+} + D_1D_2\frac{\sigma_-^{(1)}}{\sigma_-} - ABG_1\left(\frac{C_1D_2}{\sigma_-\sigma_1^{(1)}} + \frac{C_2D_1}{\sigma_+\sigma_-^{(1)}}\right).
\end{aligned}
$$

The covariance between standardized log rank statistics for true marker positive and true marker negative test at interim analysis is

$$
\begin{aligned}
Cov(Z_+^{(1)}, Z_-^{(1)}) &= -\left\{AB\left(\eta(1-\eta)(1-q)^2 Var(Q_*^{(1)}) + \tau(1-\tau)q^2 Var(Q_\phi^{(1)})\right)\right\}/\sigma_+^{(1)}\sigma_-^{(1)}\\
&= -\frac{ABG_1}{\sigma_+^{(1)}\sigma_-^{(1)}}
\end{aligned}
$$

The covariance between standardized log rank statistics for true marker positive subgroup and overall population at interim analysis is

$$
Cov(Z_+^{(1)}, Z^{(1)}) = C_1 + D_1 Cov(Z_+^{(1)}, Z_-^{(1)}) = C_1 - \frac{ABD_1G_1}{\sigma_+^{(1)}\sigma_-^{(1)}}.
$$

The covariance between standardized log rank statistics for true marker negative subgroup and overall population at interim analysis is

$$
Cov(Z_-^{(1)}, Z^{(1)}) = D_1 + C_1 Cov(Z_+^{(1)}, Z_-^{(1)}) = D_1 - \frac{ABC_1G_1}{\sigma_+^{(1)}\sigma_-^{(1)}}.
$$

The covariance between standardized log rank statistics for true marker positive subgroup and true marker negative subgroup at final analysis time is

$$\begin{aligned} Cov(Z_+, Z_-) &= -AB\left\{\left(\eta(1-\eta)(1-q)^2 Var(Q_*) + \tau(1-\tau)q^2 Var(Q_\phi)\right)\right\}/\sigma_+\sigma_- \\ &= -\frac{ABG_2}{\sigma_+\sigma_-}. \end{aligned}$$

The covariance between standardized log rank statistics for true marker positive subgroup and overall population at final analysis time is

$$\begin{aligned} Cov(Z_+, Z) &= C_2 + D_2 Cov(Z_+, Z_-) \\ &= C_2 - \frac{ABD_2G_2}{\sigma_+\sigma_-}. \end{aligned}$$

The covariance between standardized log rank statistics for true marker negative subgroup and overall population at final analysis time is

$$\begin{aligned} Cov(Z_-, Z) &= D_2 + C_2 Cov(Z_+, Z_-) \\ &= D_2 - \frac{ABC_2G_2}{\sigma_+\sigma_-}. \end{aligned}$$

The covariance between standardized log rank statistics for true marker negative subgroup (interim) and true marker negative subgroup (final) is

$$Cov(Z_-, Z_-^{(1)}) = \sigma_-^{(1)}/\sigma_-.$$

The covariance between standardized log rank statistics for true marker positive subgroup (interim) and true marker positive subgroup (final) is

$$Cov(Z_+, Z_+^{(1)}) = \sigma_+^{(1)}/\sigma_+.$$

The covariance between standardized log rank statistics for true marker positive subgroup (interim) and overall population (final) is

$$\begin{aligned} Cov(Z, Z_+^{(1)}) &= C_2 Cov(Z_+, Z_+^{(1)}) + D_2 Cov(Z_-, Z_+^{(1)}) \\ &= C_2 \frac{\sigma_+^{(1)}}{\sigma_+} - \frac{ABD_2G_1}{\sigma_0 \sigma_+^{(1)}}. \end{aligned}$$

The covariance between standardized log rank statistics for true marker negative subgroup (interim) and overall population (final) is

$$\begin{aligned} Cov(Z, Z_-^{(1)}) &= C_2 Cov(Z_+, Z_-^{(1)}) + D_2 Cov(Z_-, Z_-^{(1)}) \\ &= -\frac{ABC_2G_1}{\sigma_+ \sigma_-^{(1)}} + D_2 \sigma_-^{(1)} / \sigma_-. \end{aligned}$$

The covariance between standardized log rank statistics for true marker positive subgroup (interim) and marker negative subgroup (final) is

$$\begin{aligned} Cov(Z_-, Z_+^{(1)}) &= -AB \left\{ \eta(1-\eta)(1-q)^2 Var(Q_*^{(1)}) + q^2(1-\tau)\tau Var(Q_\phi^{(1)}) \right\} / \sigma_- \sigma_1^{(1)} \\ &= \frac{-ABG_1}{\sigma_- \sigma_+^{(1)}}. \end{aligned}$$

The covariance between standardized log rank statistics for true marker negative subgroup (interim) and marker positive subgroup (final) is

$$\begin{aligned} Cov(Z_+, Z_-^{(1)}) &= -AB \left\{ \eta(1-\eta)(1-q)^2 Var(Q_*^{(1)}) + q^2(1-\tau)\tau Var(Q_\phi^{(1)}) \right\} / \sigma_+ \sigma_0^{(1)} \\ &= \frac{-ABG_1}{\sigma_+ \sigma_-^{(1)}}. \end{aligned}$$

The covariance between standardized log rank statistics for overall population (interim) and marker positive subgroup (final) is

$$\begin{aligned} Cov(Z_+, Z^{(1)}) &= C_1 Cov(Z_+, Z_+^{(1)}) + D_1 Cov(Z_+, Z_-^{(1)}) \\ &= C_1 \frac{\sigma_+^{(1)}}{\sigma_+} - \frac{ABD_1G_1}{\sigma_1 \sigma_-^{(1)}}. \end{aligned}$$

The covariance between standardized log rank statistics for overall population (interim) and marker negative subgroup (final) is

$$\begin{aligned} Cov(Z_-, Z^{(1)}) &= C_1 Cov(Z_-, Z_+^{(1)}) + D_1 Cov(Z_-, Z_-^{(1)}) \\ &= \frac{-ABC_1G_1}{\sigma_-\sigma_+^{(1)}} + D_1\frac{\sigma_0^{(1)}}{\sigma_-}. \end{aligned}$$

The covariance between standardized log rank statistics for overall population (interim) and overall population (final) is

$$\begin{aligned} Cov(Z, Z^{(1)}) &= Cov(\frac{pZ_+ + (1-p)Z_-}{\sigma}, \frac{pZ_+^{(1)} + (1-p)Z_-^{(1)}}{\sigma^{(1)}}) \\ &= C_1C_2\frac{\sigma_+^{(1)}}{\sigma_+} + D_1D_2\frac{\sigma_0^{(1)}}{\sigma_-} - ABG_1(\frac{C_1D_2}{\sigma_-\sigma_+^{(1)}} + \frac{C_2D_1}{\sigma_+\sigma_-^{(1)}}). \end{aligned}$$

In summary, the correlation matrix between the standardized log rank statistics is:

$$Cov\left(\left(Z^{(1)}, Z_+^{(1)}, Z_-^{(1)}, Z, Z_+, Z_-\right)^T\right) =$$

$$\begin{bmatrix} 1 & C_1 - \frac{ABD_1G_1}{\sigma_+^{(1)}\sigma_-^{(1)}} & D_1 - \frac{ABC_1G_1}{\sigma_+^{(1)}\sigma_-^{(1)}} & E & C_1\frac{\sigma_+^{(1)}}{\sigma_+} - \frac{ABD_1G_1}{\sigma_+\sigma_-^{(1)}} & \frac{-ABC_1G_1}{\sigma_-\sigma_+^{(1)}} + D_1\frac{\sigma_-^{(1)}}{\sigma_-} \\ & 1 & \frac{-ABG_1}{\sigma_+^{(1)}\sigma_-^{(1)}} & C_2\frac{\sigma_+^{(1)}}{\sigma_+} - \frac{ABD_2G_1}{\sigma_-\sigma_+^{(1)}} & \sigma_+^{(1)}/\sigma_+ & \frac{-ABG_1}{\sigma_-\sigma_+^{(1)}} \\ & & 1 & -\frac{ABC_2G_1}{\sigma_+\sigma_-^{(1)}} + D_2\sigma_0^{(1)}/\sigma_- & \frac{-ABG_1}{\sigma_+\sigma_-^{(1)}} & \sigma_-^{(1)}/\sigma_- \\ & & & 1 & C_2 - \frac{ABD_2G_2}{\sigma_+\sigma_-} & D_2 - \frac{ABC_2G_2}{\sigma_+\sigma_-} \\ & & & & 1 & -\frac{ABG_2}{\sigma_+\sigma_-} \\ & & & & & 1 \end{bmatrix}$$

## 6.6 Asymptotic distribution of test statistics

Given $t, N, p, \lambda_{sen}, \lambda_{spec}$, then $q, \tau, \eta$ are fixed. Under the alternative, given $\pi_+^{(1)}$, $\pi_+^{(2)}$, $\widetilde{\pi}_+$, $\pi_-^{(1)}$, $\pi_-^{(2)}, \widetilde{\pi}_-$, and assumption of proportional hazard $\log\theta_+ = \log\frac{\lambda_{+T}}{\lambda_{+C}}$, and $\log\theta_- = \log\frac{\lambda_{-T}}{\lambda_{-C}}$. Asymptotically, we have

$$\begin{aligned} Z_+^{(1)} &\sim AN(\sqrt{tNr(1-r)}\frac{\pi_+^{(1)}\log\theta_+}{\sqrt{m_1}}, 1)\ , \\ Z_-^{(1)} &\sim AN(\sqrt{tNr(1-r)}\frac{\pi_-^{(1)}\log\theta_-}{\sqrt{m_0}}, 1)\ , \end{aligned} \tag{6.11}$$

$$Z^{(1)} \sim AN(\frac{\sqrt{tNr(1-r)}\{p\pi_+^{(1)}\log\theta_+/\sqrt{m_1}+(1-p)\pi_-^{(1)}\log\theta_-/\sqrt{m_0}\}}{\sqrt{p^2+(1-p)^2-2p(1-p)m_2/\sqrt{m_1m_0}}},1)\ ,$$

$$Z_+ \sim AN(\sqrt{Nr(1-r)}\frac{t\pi_+^{(2)}+(1-t)\widetilde{\pi}_+}{\sqrt{m_3}}\log\theta_+,1)\ ,$$

$$Z_- \sim AN(\sqrt{Nr(1-r)}\frac{t\pi_-^{(2)}+(1-t)\widetilde{\pi}_-}{\sqrt{m_4}}\log\theta_-,1)\ ,$$

$$Z \sim AN(\frac{\sqrt{Nr(1-r)}\{p[t\pi_+^{(2)}+(1-t)\widetilde{\pi}_+]\frac{\log\theta_+}{\sqrt{m_3}}+(1-p)[t\pi_-^{(2)}+(1-t)\widetilde{\pi}_-]\frac{\log\theta_-}{\sqrt{m_4}}\}}{\sqrt{p^2+(1-p)^2-2p(1-p)m_5/\sqrt{m_3m_4}}},1),$$

where

$$\begin{aligned}
m_1 &= \eta^2(1-q)^2\{\tau q\pi_+^{(1)}+(1-\tau)q\pi_-^{(1)}\}A^2+(1-\tau)^2q^2\{(1-\eta)(1-q)\pi_+^{(1)}\\
&\quad +\eta(1-q)\pi_-^{(1)}\}A^2,\\
m_0 &= (1-\eta)^2(1-q)^2\{\tau q\pi_+^{(1)}+(1-\tau)q\pi_-^{(1)}\}B^2+\tau^2q^2\{(1-\eta)(1-q)\pi_+^{(1)}\\
&\quad +\eta(1-q)\pi_-^{(1)}\}B^2,\\
m_2 &= AB\eta(1-\eta)(1-q)^2\{r(1-r)tN[\tau q\pi_+^{(1)}+AB(1-\tau)q\pi_-^{(1)}]\}\\
&\quad +\tau(1-\tau)q^2\{r(1-r)tN[(1-\eta)(1-q)\pi_+^{(1)}+\eta(1-q)\pi_-^{(1)}]\},\\
m_3 &= A^2\eta^2(1-q)^2\{t[\tau q\pi_+^{(2)}+(1-\tau)q\pi_-^{(2)}]+(1-t)[\tau q\widetilde{\pi}_++(1-\tau)q\widetilde{\pi}_-]\}\\
&\quad +A^2(1-\tau)^2q^2\{t[(1-\eta)(1-q)\pi_+^{(2)}+\eta(1-q)\pi_-^{(2)}]\\
&\quad +(1-t)[(1-\eta)(1-q)\widetilde{\pi}_++\eta(1-q)\widetilde{\pi}_-]\},\\
m_4 &= B^2(1-\eta)^2(1-q)^2\{t[\tau q\pi_+^{(2)}+(1-\tau)q\pi_-^{(2)}]+(1-t)[\tau q\widetilde{\pi}_++(1-\tau)q\widetilde{\pi}_-]\}\\
&\quad +B^2\tau^2q^2\{t[(1-\eta)(1-q)\pi_+^{(2)}+\eta(1-q)\pi_-^{(2)}]\\
&\quad +(1-t)[(1-\eta)(1-q)\widetilde{\pi}_++\eta(1-q)\widetilde{\pi}_-]\},\\
m_5 &= AB\eta(1-\eta)(1-q)^2r(1-r)\{tN[\tau q\pi_+^{(2)}+(1-\tau)q\pi_-^{(2)}]\\
&\quad +(1-t)N[\tau q\widetilde{\pi}_++(1-\tau)q\widetilde{\pi}_-]\}\\
&\quad +AB\tau(1-\tau)q^2r(1-r)\{tN[(1-\eta)(1-q)\pi_+^{(2)}+\eta(1-q)\pi_-^{(2)}]\\
&\quad +(1-t)N[(1-\eta)(1-q)\widetilde{\pi}_++\eta(1-q)\widetilde{\pi}_-]\}.
\end{aligned}$$

## 6.7 Type I error $\alpha$ allocation and critical values

For the two-stage stratified designs, we can split the overall alpha (e.g, $\alpha = 0.025$) between the two stages, following a traditional sequential design. In Stage I, a fraction of the overall alpha, $\alpha_1$, is allocated to test the global hypothesis $H_0$

$$\begin{aligned}
\alpha_1 &= P(\text{Reject } H_0 | H_0) \\
&= P_0(Z^{(1)} < -c_1 \text{ or } Z_+^{(1)} < -c_2) \\
&= P_0(Z^{(1)} < -c_1) + P_0(Z^{(1)} \geq -c_1 \text{ or } Z_+^{(1)} < -c_2) \\
&= \alpha_{1a} + \alpha_{1b}.
\end{aligned}$$

The critical value $c_1$ is obtained by allocating a portion $\alpha_{1a}$ of $\alpha_1$, for testing $H_{0a}$, Then $c_2$ can be solved for testing $H_{0+}$ in the above equation.

For Stage II, the overall alpha is left with $\alpha - \alpha_1$, where

$$\begin{aligned}
\alpha - \alpha_1 &= P(\text{Accept } H_0 \text{ at Stage } I, \text{ Reject } H_0 \text{ at Stage } II) \\
&= \alpha_2.
\end{aligned}$$

We have

$$\begin{aligned}
\alpha_2 &= P(\text{Accept } H_0 \text{ at Stage I, Reject } H_0 \text{ at Stage } II) \\
&= P_0(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z < -b_1) \\
&\qquad + P_0(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z \geq -b_1, Z_+ < -b_2) \\
&= \alpha_{2a} + (\alpha_2 - \alpha_{2a}).
\end{aligned}$$

That is, the critical value $b_1$ is obtained by allocating $\alpha_{2a}$, a portion of $\alpha_2$, for testing $H_{0a}$. Then $b_2$ can be solved for testing $H_{0+}$ in the above equation.

Tables 6.1 and 6.2 show the critical values for some commonly used situations. Information time is 0.3 and 0.5, respectively. From the tables, we can see when the prevalence rate, $\lambda_{sen}$ , and $\lambda_{spec}$ are fixed, changing information time form 0.3 to 0.5 does not change the

critical value for overall population ($c_1$) and for marker-positive population ($c_2$) at interim analysis, but change the critical values for overall population ($b_1$) and for marker-positive population ($b_2$) at final analysis time.

Table 6.1: Critical Values When $\alpha = 0.025,\ \alpha_+ = 0.004, \alpha_2 = 0.021, r = 0.5$ and info=0.3

| $\lambda_{sen}$ | $\lambda_{spec}$ | $p$ | $c_1$ | $c_2$ | $b_1$ | $b_2$ |
|---|---|---|---|---|---|---|
| 1.0 | 1.0 | 0.3 | -2.878 | -2.866 | -2.287 | -2.255 |
| | | 0.4 | -2.878 | -2.848 | -2.286 | -2.224 |
| | | 0.5 | -2.878 | -2.816 | -2.284 | -2.178 |
| 0.95 | 0.95 | 0.3 | -2.878 | -2.871 | -2.288 | -2.267 |
| | | 0.4 | -2.878 | -2.857 | -2.286 | -2.238 |
| | | 0.5 | -2.878 | -2.826 | -2.285 | -2.192 |
| 0.9 | 0.9 | 0.3 | -2.878 | -2.875 | -2.288 | -2.276 |
| | | 0.4 | -2.878 | -2.864 | -2.287 | -2.252 |
| | | 0.5 | -2.878 | -2.836 | -2.285 | -2.205 |
| 0.85 | 0.85 | 0.3 | -2.878 | -2.877 | -2.289 | -2.283 |
| | | 0.4 | -2.878 | -2.870 | -2.288 | -2.264 |
| | | 0.5 | -2.878 | -2.845 | -2.286 | -2.219 |
| 0.8 | 0.8 | 0.3 | -2.878 | -2.878 | -2.289 | -2.287 |
| | | 0.4 | -2.878 | -2.874 | -2.288 | -2.274 |
| | | 0.5 | -2.878 | -2.854 | -2.286 | -2.233 |
| 0.75 | 0.75 | 0.3 | -2.878 | -2.878 | -2.289 | -2.289 |
| | | 0.4 | -2.878 | -2.877 | -2.288 | -2.282 |
| | | 0.5 | -2.878 | -2.861 | -2.270 | -2.231 |
| 0.7 | 0.7 | 0.3 | -2.878 | -2.878 | -2.290 | -2.290 |
| | | 0.4 | -2.878 | -2.878 | -2.274 | -2.272 |
| | | 0.5 | -2.878 | -2.867 | -2.287 | -2.258 |
| 1 | 0.8 | 0.3 | -2.878 | -2.876 | -2.288 | -2.279 |
| | | 0.4 | -2.878 | -2.865 | -2.287 | -2.253 |
| | | 0.5 | -2.878 | -2.835 | -2.285 | -2.204 |
| 0.8 | 1 | 0.3 | -2.878 | -2.872 | -2.288 | -2.268 |
| | | 0.4 | -2.878 | -2.860 | -2.287 | -2.245 |
| | | 0.5 | -2.878 | -2.834 | -2.285 | -2.202 |

Table 6.2: Critical Values When $\alpha = 0.025,\ \alpha_+ = 0.004, \alpha_2 = 0.021, r = 0.5$ and info=0.5

| $\lambda_{sen}$ | $\lambda_{spec}$ | $p$ | $c_1$ | $c_2$ | $b_1$ | $b_2$ |
|---|---|---|---|---|---|---|
| 1.0 | 1.0 | 0.3 | -2.878 | -2.866 | -2.271 | -2.240 |
| | | 0.4 | -2.878 | -2.848 | -2.269 | -2.210 |
| | | 0.5 | -2.878 | -2.816 | -2.266 | -2.164 |
| 0.95 | 0.95 | 0.3 | -2.878 | -2.871 | -2.272 | -2.252 |
| | | 0.4 | -2.878 | -2.857 | -2.270 | -2.224 |
| | | 0.5 | -2.878 | -2.826 | -2.266 | -2.177 |
| 0.9 | 0.9 | 0.3 | -2.878 | -2.875 | -2.273 | -2.261 |
| | | 0.4 | -2.878 | -2.864 | -2.271 | -2.237 |
| | | 0.5 | -2.878 | -2.836 | -2.267 | -2.191 |
| 0.85 | 0.85 | 0.3 | -2.878 | -2.877 | -2.274 | -2.268 |
| | | 0.4 | -2.878 | -2.870 | -2.272 | -2.249 |
| | | 0.5 | -2.878 | -2.845 | -2.268 | -2.205 |
| 0.8 | 0.8 | 0.3 | -2.878 | -2.878 | -2.274 | -2.272 |
| | | 0.4 | -2.878 | -2.874 | -2.273 | -2.259 |
| | | 0.5 | -2.878 | -2.854 | -2.269 | -2.219 |
| 0.75 | 0.75 | 0.3 | -2.878 | -2.878 | -2.274 | -2.274 |
| | | 0.4 | -2.878 | -2.877 | -2.274 | -2.267 |
| | | 0.5 | -2.878 | -2.861 | -2.270 | -2.231 |
| 0.7 | 0.7 | 0.3 | -2.878 | -2.878 | -2.275 | -2.275 |
| | | 0.4 | -2.878 | -2.878 | -2.274 | -2.272 |
| | | 0.5 | -2.878 | -2.867 | -2.271 | -2.243 |
| 1 | 0.8 | 0.3 | -2.878 | -2.876 | -2.273 | -2.264 |
| | | 0.4 | -2.878 | -2.865 | -2.271 | -2.239 |
| | | 0.5 | -2.878 | -2.835 | -2.267 | -2.189 |
| 0.8 | 1 | 0.3 | -2.878 | -2.872 | -2.272 | -2.254 |
| | | 0.4 | -2.878 | -2.860 | -2.270 | -2.231 |
| | | 0.5 | -2.878 | -2.834 | -2.267 | -2.188 |

## 6.8 Global and marginal power

The global power is

$$\begin{aligned}
1-\beta &= P(\text{Reject } H_0|H_1) \\
&= P(\text{Reject } H_{0a} \text{ or Reject } H_{0+}|H_1) \\
&= P_1(Z^{(1)} < -c_1 \textit{ or } Z_+^{(1)} < -c_2) \\
&\quad + P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z < -b_1) \\
&\quad + P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z \geq -b_1, Z_+ < -b_2) \\
&= p_1 + p_{2a} + p_{2+}
\end{aligned}$$

where

$$\begin{aligned}
p_1 &= P(\text{Reject } H_{0a} \text{ or Reject } H_{0+} \text{ at Stage } I|H_1) \\
&= P_1(Z^{(1)} < -c_1 \textit{ or } Z_+^{(1)} < -c_2),
\end{aligned}$$

and

$$\begin{aligned}
p_{2a} &= P(\text{Accept } H_0 \text{ at Stage } I \text{ and Reject } H_{0a} \text{ at Stage } II_A \ |H_1) \\
&= P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z < -b_1).
\end{aligned}$$

Power for testing the treatment effect in the overall cohort is

$$\begin{aligned}
1-\beta_a &= P(\text{Reject } H_{0a}|H_1) \\
&= P(\text{Reject } H_{0a} \text{ at Stage } I) + \text{P}(\text{ Reject } H_{0a} \text{ at Stage } II_A) \\
&= P_1(Z^{(1)} < -c_1) + p_{2a} \\
&= p_{1a} + p_{2a}.
\end{aligned}$$

Power for testing the treatment effect in the marker-positive cohort is

$$1-\beta_+ = P(\text{Reject } H_{0+}|H_1)$$

$$= P(\text{Reject } H_{0+} \text{ at Stage } I) + P(\text{Reject } H_{0+} \text{ at Stage } II)$$
$$= P_1(Z_+^{(1)} < -c_2) + P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_+ < -b_2).$$

## 6.9 Sample size calculations

Given the global type I error and power, assuming that we have the estimated prevalence rate $p$, sensitivity $\lambda_{sen}$ and specificity $\lambda_{spec}$ from previous studies, the sample size needed to detect the treatment effect can be found based on the formulas in Section 6.7 and Section 6.8 after we specify the design parameters in these sections.

To be more specific, given the design parameters shown in Section 6.7, we find the critical values first based on the formulas in Section 6.7. Notice that the critical values are based on the distributions under null hypothesis and do not depend on the sample size. Next using these critical values and the formulas shown in Section 6.8, we can determine the sample size needed to achieve the specified power of specific type (global, overall or marker positive), through optimization programming algorithms to find the solution and round up the the nearest integer. R-code is developed to calculate the needed sample size (the number of events) and results are illustrated in Chapter 7.

# Chapter 7

# Numeric examples

.

## 7.1 Simulation set-up

Consider a total number of subjects $N$=1000 for both Stage I and Stage II, with a prevalence rate p (0.3, 0.4, or 0.5) for biomarker positive ($S+$). Subjects enrollment is expected to complete in 10 months. Randomization to active treatment T or control treatment C will be stratified by marker-appeared status with a randomization ratio of $r$ to $1-r$, to $T$ or $C$, respectively. After the recruitment is complete, the study will follow-up to interim analysis time $T_1$ and final analysis time $T_2$. Assume subject recruitment follow a uniform distribution. Survival times are exponentially distributed.

The null hypothesis is the hazard rates for treatment and control group are equal. Exponential distributions are assumed to simulate the trials. The hazard rate for treated in true marker positive $S+$ group is $\lambda_{+T}$, and hazard rate for control treated in true marker positive $S+$ group is $\lambda_{+C}$. The hazard rates for true marker negative group $S-$ are $\lambda_{-T}$ and $\lambda_{-C}$.

We consider 9 different combinations of sensitivity and specificity ($\lambda_{sen} = \lambda_{spec} = 0.7$ to 1.0, $\lambda_{sen} = 1$ and $\lambda_{spec} = 0.8$, or $\lambda_{sen} = 0.8$ and $\lambda_{spec} = 1.0$), and simulate the trials for 3000 times for each combination of sensitivity and specificity.

The nominal and empirical global type I error rate are shown in Table 7.1. As we can see from Table 7.1, the empirical type I error rates are close to the nominal type I error rate 0.025, across different prevalence rate (0.3, 0.4, 0.5), different sensitivity $\lambda_{sen}$ and specificity $\lambda_{spec}$ combinations from 1.0 to 0.7.

Table 7.1: The Nominal and Empirical Global Type I Error Rate for $H_0$ when $\lambda_{-T} = \lambda_{-C} = 1/10$ and $\lambda_{+T} = \lambda_{+C} = 1/15$, Info=0.5, N=1000, d=750, 3000 runs

| $\lambda_{sen}$ | $\lambda_{spec}$ | $p$ | Nominal global type I | Empirical global type I |
|---|---|---|---|---|
| 1.0 | 1.0 | 0.3 | 0.025 | 0.0250 |
| | | 0.4 | 0.025 | 0.0273 |
| | | 0.5 | 0.025 | 0.0250 |
| 0.95 | 0.95 | 0.3 | 0.025 | 0.0240 |
| | | 0.4 | 0.025 | 0.0273 |
| | | 0.5 | 0.025 | 0.0257 |
| 0.9 | 0.9 | 0.3 | 0.025 | 0.0253 |
| | | 0.4 | 0.025 | 0.0260 |
| | | 0.5 | 0.025 | 0.0270 |
| 0.85 | 0.85 | 0.3 | 0.025 | 0.0270 |
| | | 0.4 | 0.025 | 0.0277 |
| | | 0.5 | 0.025 | 0.0220 |
| 0.8 | 0.8 | 0.3 | 0.025 | 0.0290 |
| | | 0.4 | 0.025 | 0.0287 |
| | | 0.5 | 0.025 | 0.0250 |
| 0.75 | 0.75 | 0.3 | 0.025 | 0.0280 |
| | | 0.4 | 0.025 | 0.0257 |
| | | 0.5 | 0.025 | 0.0250 |
| 0.7 | 0.7 | 0.3 | 0.025 | 0.0253 |
| | | 0.4 | 0.025 | 0.0263 |
| | | 0.5 | 0.025 | 0.0263 |
| 1 | 0.8 | 0.3 | 0.025 | 0.0263 |
| | | 0.4 | 0.025 | 0.0277 |
| | | 0.5 | 0.025 | 0.0283 |
| 0.8 | 1 | 0.3 | 0.025 | 0.0267 |
| | | 0.4 | 0.025 | 0.0250 |
| | | 0.5 | 0.025 | 0.0247 |

## 7.2 Theoretical versus empirical power under different scenarios

The theoretical and empirical global, overall, and positive subgroup power are shown in Table 7.2.

In this simulation, we use $\alpha = 0.025, \alpha_+ = 0.004$, $r = 0.5$ to calculate the critical values. As we can see from Table 7.2, when there is treatment effect only in biomarker positive cohort and no treatment effect in biomarker negative cohort, the empirical powers (global power, overall power, and positive cohort power) are close to the corresponding theoretical powers, across different prevalence rate (0.3, 0.4, 0.5) under different combination of biomarker sensitivity $\lambda_{sen}$ and specificity $\lambda_{spec}$ (from 1.0 to 0.7). In the current set-up, since there is no treatment effect in biomarker negative cohort, the overall power is low.

Table 7.2: The Theoretical and Empirical Power when $\lambda_{-T} = \lambda_{-C} = \lambda_{+C} = 1/10$ and $\lambda_{+T} = 1/15$, Info = 0.5, N=1000, d=750, 3000 runs

| | | | Global Power | | Overall Power | | Positive Group Power | |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{sen}$ | $\lambda_{spec}$ | $p$ | Theoretical | Empirical | Theoretical | Empirical | Theoretical | Empirical |
| 1.0 | 1.0 | 0.3 | 0.772 | 0.764 | 0.099 | 0.094 | 0.754 | 0.747 |
| | | 0.4 | 0.895 | 0.888 | 0.233 | 0.226 | 0.874 | 0.869 |
| | | 0.5 | 0.958 | 0.954 | 0.445 | 0.448 | 0.936 | 0.937 |
| 0.95 | 0.95 | 0.3 | 0.688 | 0.685 | 0.099 | 0.096 | 0.660 | 0.664 |
| | | 0.4 | 0.849 | 0.848 | 0.247 | 0.255 | 0.814 | 0.812 |
| | | 0.5 | 0.938 | 0.931 | 0.481 | 0.480 | 0.897 | 0.895 |
| 0.9 | 0.9 | 0.3 | 0.595 | 0.600 | 0.097 | 0.099 | 0.553 | 0.559 |
| | | 0.4 | 0.790 | 0.789 | 0.257 | 0.265 | 0.732 | 0.727 |
| | | 0.5 | 0.910 | 0.906 | 0.517 | 0.530 | 0.840 | 0.835 |
| 0.85 | 0.85 | 0.3 | 0.495 | 0.504 | 0.092 | 0.098 | 0.440 | 0.445 |
| | | 0.4 | 0.719 | 0.718 | 0.262 | 0.281 | 0.628 | 0.623 |
| | | 0.5 | 0.874 | 0.886 | 0.552 | 0.576 | 0.755 | 0.765 |
| 0.8 | 0.8 | 0.3 | 0.395 | 0.409 | 0.084 | 0.094 | 0.331 | 0.338 |
| | | 0.4 | 0.634 | 0.643 | 0.262 | 0.281 | 0.506 | 0.500 |
| | | 0.5 | 0.832 | 0.842 | 0.583 | 0.607 | 0.639 | 0.632 |
| 0.75 | 0.75 | 0.3 | 0.300 | 0.313 | 0.073 | 0.075 | 0.233 | 0.246 |
| | | 0.4 | 0.544 | 0.563 | 0.253 | 0.281 | 0.376 | 0.376 |
| | | 0.5 | 0.786 | 0.802 | 0.608 | 0.641 | 0.495 | 0.499 |
| 0.7 | 0.7 | 0.3 | 0.213 | 0.226 | 0.061 | 0.064 | 0.153 | 0.163 |
| | | 0.4 | 0.447 | 0.483 | 0.235 | 0.261 | 0.253 | 0.267 |
| | | 0.5 | 0.742 | 0.761 | 0.626 | 0.664 | 0.340 | 0.338 |
| 1 | 0.8 | 0.3 | 0.558 | 0.567 | 0.095 | 0.096 | 0.512 | 0.522 |
| | | 0.4 | 0.781 | 0.783 | 0.259 | 0.264 | 0.718 | 0.723 |
| | | 0.5 | 0.913 | 0.917 | 0.514 | 0.523 | 0.846 | 0.848 |
| 0.8 | 1 | 0.3 | 0.674 | 0.666 | 0.099 | 0.098 | 0.643 | 0.634 |
| | | 0.4 | 0.823 | 0.807 | 0.252 | 0.263 | 0.777 | 0.762 |
| | | 0.5 | 0.917 | 0.913 | 0.508 | 0.518 | 0.855 | 0.851 |

Figures 7.1, 7.2, and 7.3 show the contour plots of power surfaces for global (testing $H_1$), overall population (testing $H_{1a}$) and marker-positive population (testing $H_{1+}$) hypotheses, respectively, across $-0.10 \geq \delta \geq -0.40$ and $-0.10 \geq \delta_+ \geq -0.40$ by $n, p$ assuming $\alpha = 0.025, \alpha_1 = 0.004, w_+ = p, w_- = 1 - p, \delta = p\delta_+ + (1-p)\delta_-$, $r = 0.5$.

The power increases as $n$, $p$, $\lambda_{sen}$, or $\lambda_{spec}$ increases as well. For example, with $p = 0.5$, and $n = 500$, $\delta = -0.15$, and $\delta_+ = -0.4$, Figure 7.1.A with $\lambda_{sen} = \lambda_{spec} = 1$ shows power 70%. However, Figure 7.1.D shows power only about 45%. We also see the $\lambda_{sen}$ has less impact than $\lambda_{spec}$ in terms of power.

From Figure 7.1, the global power increases with increasing treatment effect for overall population and positive population (decreasing $\delta$ and/or decreasing $\delta_+$).

From Figure 7.2, the power for overall population increases with increasing treatment effect for overall population when treatment effect for positive population is fixed (decreasing $\delta$).

From Figure 7.3, the power for positive subgroup increases with increasing treatment effect for positive population (decreasing $\delta_+$) but decreases with increasing treatment effect for overall population (decreasing $\delta$).

Tables 7.3 and 7.4 show the sample size needed to achieve specified global and marginal power for marker positive subgroup, respectively. To illustrate sample size calculation, with the same $\alpha$ allocation, $\lambda_{sen} = \lambda_{spec} = 0.8$, and treatment effects $\delta = -0.15$ (the log-hazard ratio for the overall cohort), and $\delta_+ = -0.4$ (the log-hazard ratio for the marker positive cohort). For instance, assuming a target of 90% global power for testing $H_1$ when prevalence is 0.4 and interim analysis is performed at information time of 0.5, a total sample size 2062 is needed (see Table 7.3). If the target of 80% biomarker positive marginal power is needed for testing $H_{1+}$ when prevalence is 0.4 and interim analysis is performed at information time of 0.5, a total sample size of 2043 is needed (see Table 7.4).

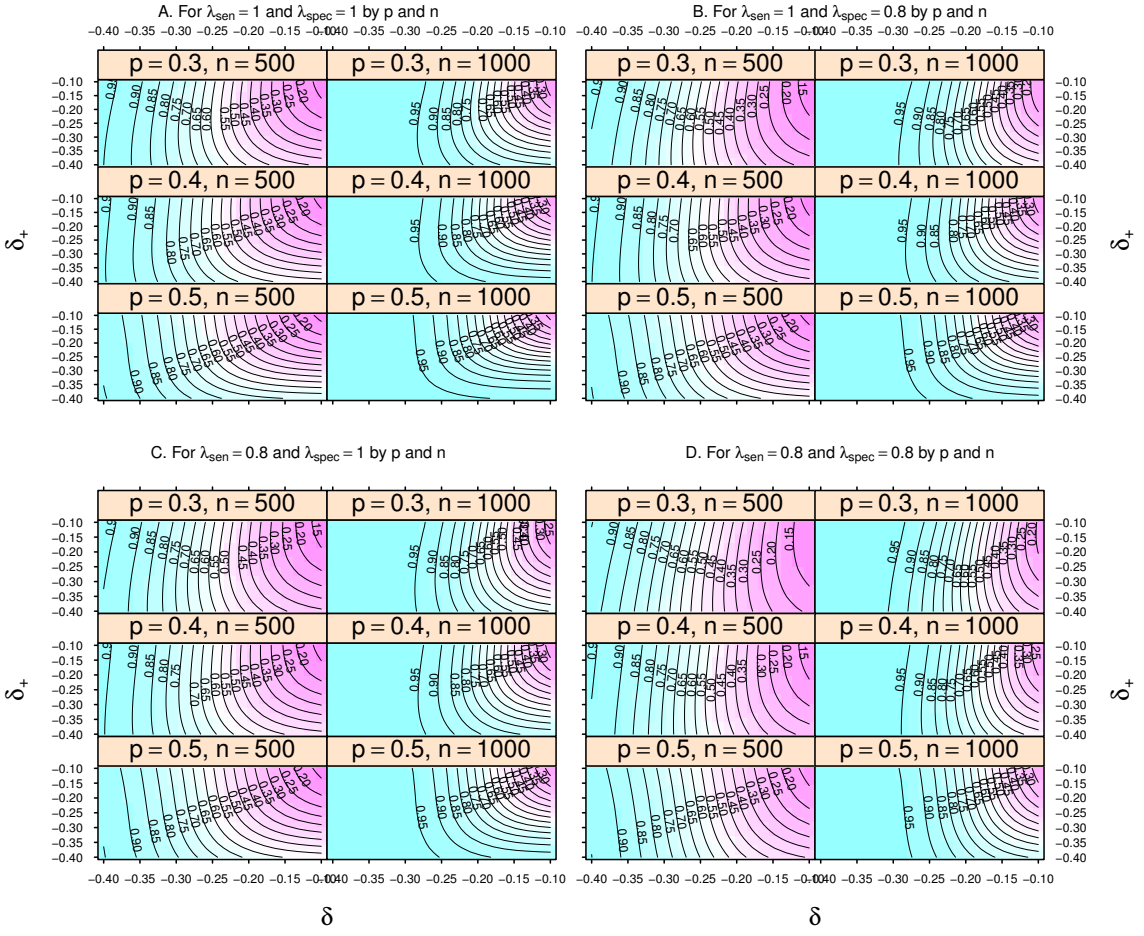



Figure 7.1: Contour plot of global power surface

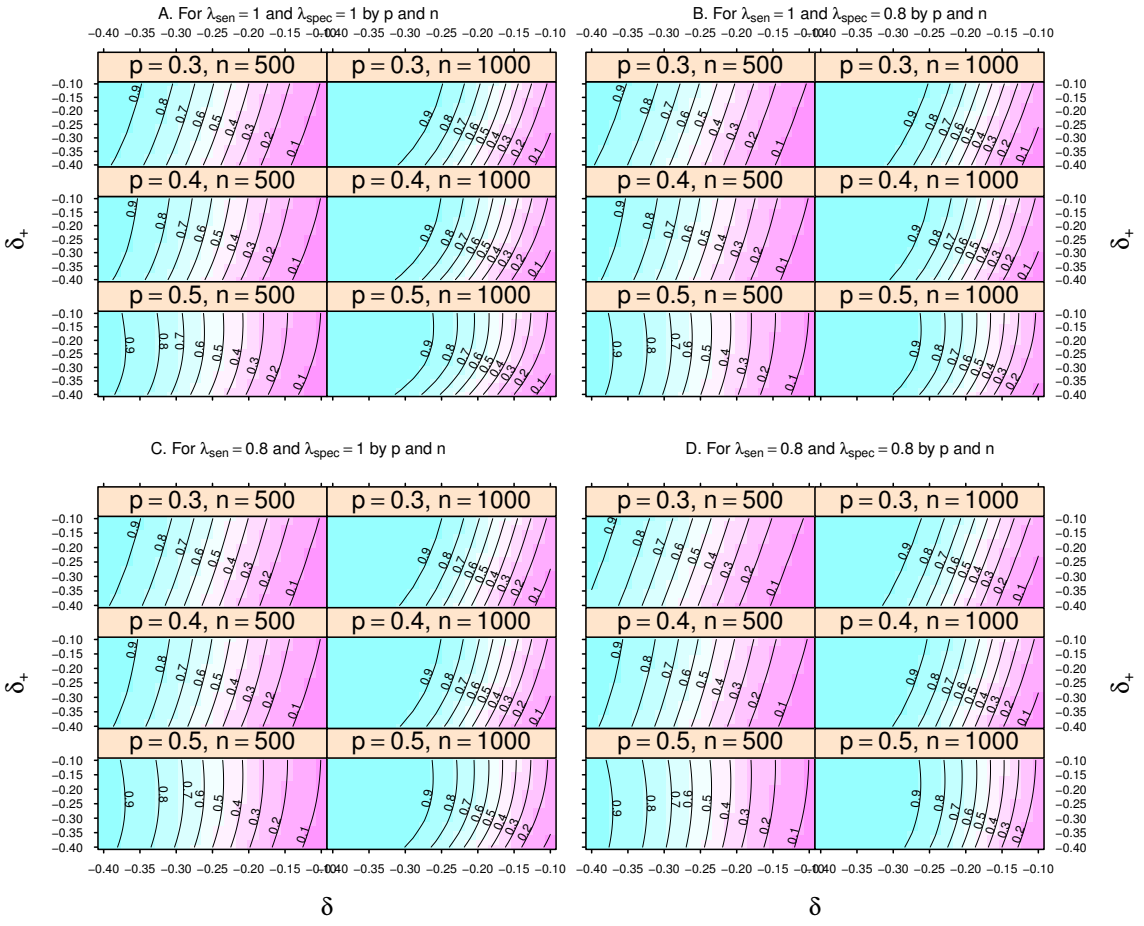



Figure 7.2: Contour plot of power surface for overall population

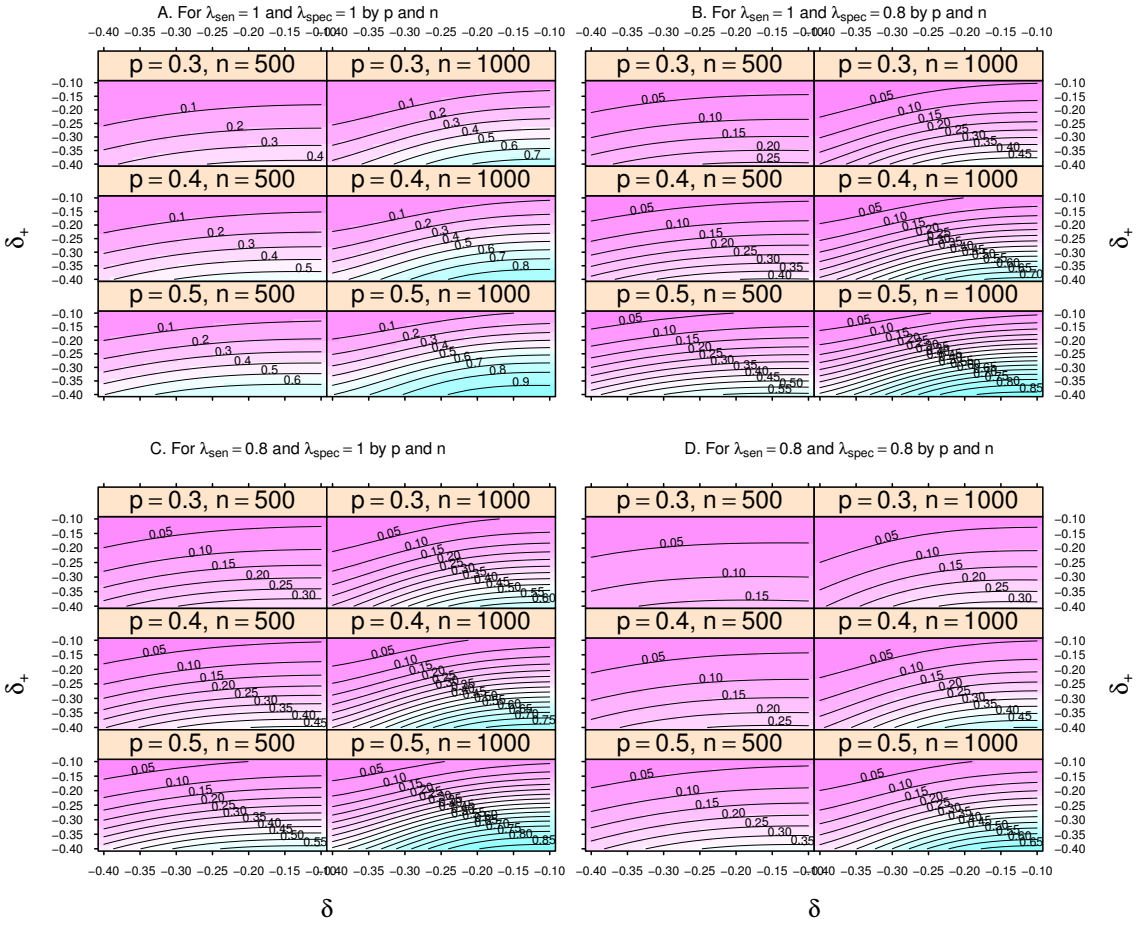



Figure 7.3: Contour plot of power surface for marker positive subgroup

Table 7.3: Total Sample Size to Achieve Specified Global Power $H_1$ when $\alpha = 0.025, \alpha_+ = 0.004$, $r = 0.5$, and $\delta = p \log \theta_+ + (1-p) \log \theta_- = -0.15, \delta_+ = \log \theta_+ = -0.4$.

| | | | 90% power | | 80% power | |
|---|---|---|---|---|---|---|
| $p$ | $\lambda_{sen}$ | $\lambda_{spec}$ | Info=0.3 | Info=0.5 | Info=0.3 | Info=0.5 |
| 0.3 | 1 | 1 | 1399 | 1385 | 1061 | 1049 |
| | 0.95 | 0.95 | 1647 | 1631 | 1249 | 1235 |
| | 0.9 | 0.9 | 1959 | 1938 | 1487 | 1470 |
| | 0.85 | 0.85 | 2356 | 2331 | 1794 | 1772 |
| | 0.8 | 0.8 | 2878 | 2843 | 2203 | 2174 |
| | 0.75 | 0.75 | 3606 | 3557 | 2783 | 2744 |
| | 0.7 | 0.7 | 4730 | 4667 | 3702 | 3649 |
| | 1 | 0.8 | 2094 | 2071 | 1591 | 1572 |
| | 0.8 | 1 | 1695 | 1679 | 1286 | 1272 |
| 0.4 | 1 | 1 | 1065 | 1056 | 808 | 800 |
| | 0.95 | 0.95 | 1225 | 1214 | 929 | 919 |
| | 0.9 | 0.9 | 1434 | 1422 | 1087 | 1076 |
| | 0.85 | 0.85 | 1712 | 1696 | 1298 | 1283 |
| | 0.8 | 0.8 | 2085 | 2062 | 1582 | 1563 |
| | 0.75 | 0.75 | 2589 | 2560 | 1971 | 1947 |
| | 0.7 | 0.7 | 3292 | 3252 | 2525 | 2490 |
| | 1 | 0.8 | 1471 | 1458 | 1115 | 1103 |
| | 0.8 | 1 | 1318 | 1307 | 999 | 989 |
| 0.5 | 1 | 1 | 839 | 833 | 636 | 630 |
| | 0.95 | 0.95 | 946 | 938 | 717 | 710 |
| | 0.9 | 0.9 | 1090 | 1080 | 826 | 818 |
| | 0.85 | 0.85 | 1287 | 1276 | 974 | 964 |
| | 0.8 | 0.8 | 1559 | 1544 | 1177 | 1165 |
| | 0.75 | 0.75 | 1926 | 1906 | 1453 | 1436 |
| | 0.7 | 0.7 | 2388 | 2363 | 1803 | 1782 |
| | 1 | 0.8 | 1075 | 1065 | 815 | 807 |
| | 0.8 | 1 | 1051 | 1043 | 796 | 788 |

Table 7.4: Total Sample Size to Achieve Specified Marginal Power $H_{1+}$ when $\alpha = 0.025, \alpha_1 = 0.004$, $r = 0.5$, and $\delta = p \log \theta_+ + (1-p) \log \theta_- = -0.15, \delta_+ = \log \theta_+ = -0.4$.

| | | | 90% power | | 80% power | |
|---|---|---|---|---|---|---|
| $p$ | $\lambda_{sen}$ | $\lambda_{spec}$ | Info=0.3 | Info=0.5 | Info=0.3 | Info=0.5 |
| 0.3 | 1 | 1 | 1561 | 1567 | 1162 | 1162 |
| | 0.95 | 0.95 | 1947 | 1983 | 1428 | 1438 |
| | 0.9 | 0.9 | 2524 | 2664 | 1803 | 1840 |
| | 0.85 | 0.85 | 3507 | 4136 | 2364 | 2463 |
| | 0.8 | 0.8 | 14594 | 8538 | 3258 | 3561 |
| | 0.75 | 0.75 | 24612 | 14406 | 4837 | 6852 |
| | 0.7 | 0.7 | 40364 | 23643 | 30359 | 17734 |
| | 1 | 0.8 | 2820 | 3055 | 1983 | 2037 |
| | 0.8 | 1 | 2027 | 2072 | 1482 | 1495 |
| 0.4 | 1 | 1 | 1116 | 1107 | 842 | 836 |
| | 0.95 | 0.95 | 1318 | 1311 | 989 | 984 |
| | 0.9 | 0.9 | 1606 | 1616 | 1194 | 1196 |
| | 0.85 | 0.85 | 2053 | 2109 | 1499 | 1516 |
| | 0.8 | 0.8 | 2828 | 3082 | 1982 | 2043 |
| | 0.75 | 0.75 | 11111 | 6687 | 2825 | 3056 |
| | 0.7 | 0.7 | 22606 | 13251 | 4600 | 7854 |
| | 1 | 0.8 | 1662 | 1676 | 1233 | 1236 |
| | 0.8 | 1 | 1440 | 1440 | 1077 | 1075 |
| 0.5 | 1 | 1 | 855 | 845 | 647 | 641 |
| | 0.95 | 0.95 | 977 | 968 | 738 | 762 |
| | 0.9 | 0.9 | 1151 | 1143 | 865 | 860 |
| | 0.85 | 0.85 | 1416 | 1420 | 1054 | 1055 |
| | 0.8 | 0.8 | 1865 | 1914 | 1356 | 1374 |
| | 0.75 | 0.75 | 2820 | 3232 | 1888 | 1989 |
| | 0.7 | 0.7 | 13913 | 8240 | 3094 | 4084 |
| | 1 | 0.8 | 1133 | 1125 | 852 | 847 |
| | 0.8 | 1 | 1101 | 1093 | 830 | 824 |

# Chapter 8

# Keytruda trial examples

## 8.1 Misclassification of predictive biomarkers

Immunotherapy is a new paradigm for the treatment of non-small-cell cancer (NSCLC), and targeting the PD-1/PD-L1 pathway is a promising therapeutic option. Pembrolizumab is a new immunotherapy that blocks the PD-1 pathway and restores the body's immune response against cancer cells and allows the immune system to recognize and kill cancer cells. We use the KEYNOTE-10 trial as an example to illustrate our method. This was a (phase 2/3) randomized trial to study Pembrolizumab versus Docetaxel for previously treated, PD-L1-positive, advance NSCLC patients (Herbst et al., 2016). This trial stratified qualified subjects by biomarker's TPS (tumor proportion score $\geq$50% vs 1-49% ), which measured the extent of PD-L1 expression, then randomized subjects with 1:1:1 ratio to three treatment groups within each high and low TPS stratum. The companion diagnostic assay for PD-L1 expression was the Dako EnVision FLEX+HRP-Polymer kit using the 22C3 antibody clone, which was validated in the phase 1 KEYNOTE-001 trial (Garon et al., 2015). Here, we use the phase 1 KEYNOTE-001 data as the basis to "redesign" the KEYNOTE-10 as an "imaginary" two-stage stratified trial to illustrate our method. For illustration, also, since there was no significant difference between the two test doses of Pembrolizumab, we only look at the Pembrolizumab 2 mg (Pem) versus Docetaxel (Dox), which is the control/standard-of-care.

From the phase 1 KEYNOTE-001 trial, the prevalence was about 0.39 for TPS <1%, 0.38 for TPS = 1-49%, and 0.23 for TPS $\geq$50%; $\lambda_{sen} = \lambda_{spec} = 0.8$. Thus, we estimate the prevalence rate of the PD-L1 true "strongly positive" (TPS $\geq$ 50%) among the PD-L1 positive (TPS > 1%) NSCLC patients being p $\approx$ 0.40, the appeared PD-L1 "strongly positive" prevalence is $q = P(M = *) = p\lambda_{sen} + (1-p)(1-\lambda_{spec}) = 0.40 \times 0.80 + (1 -$

$0.40)(1-0.80) = 0.44$. Hence $PPV = \tau = \frac{p\lambda_{sen}}{p\lambda_{sen}+(1-p)(1-\lambda_{spec})} = \frac{p\lambda_{sen}}{q} = \frac{0.4\times 0.8}{0.44} = 0.73$, and $NPV = \eta = \frac{(1-p)\lambda_{spec}}{1-P(M=*)} = \frac{(1-p)\lambda_{spec}}{1-q} = \frac{(1-0.4)\times 0.8}{1-0.44} = 0.86$. The real phase 2/3 KEYNOTE-10 trial had both overall survival and progression-free survival (PFS) as primary end-points, We only use PFS for the "imaginary" trial for illustration purpose. Suppose that the overall (one-sided) $\alpha = 0.025$ and an interim analysis planned at $info = 0.5$, with $\alpha_1 = 0.004$ allocated for Stage I. As stated in the work of Herbst et al. (2016), the study aimed to show a benefit of Pem over Dox in PFS in patients with TPS $\geq$ 50% as well as in the whole TPS $\geq$ 1% cohort, so we allocate $\alpha_{1a} = \alpha_{1b} = \frac{\alpha_1}{2} = 0.002$. Then $\alpha_2 = 0.021$ for Stage II, an equal amount 0.0105 ($= \alpha_2/2$) of $\alpha_2$ is used for testing treatment effect on the overall cohort and for testing treatment effect on the PD-L1 strongly positive subset.

From the reports by Herbst et al. (2016), the total enrollment time is 19 months for 688 subjects. We expect a total of 688 $\times$ 0.92= 632 PFS events at the end of the trial. In Stage I, we plan to enroll 70% subjects (13.3 months from the start of the study). A decision is made at interim analysis $T_1$ (information: $info = 0.5$, when 316= 632 $\times$ 0.5 PFS is expected). If the null hypothesis is accepted at interim analysis, enroll additional 30% subjects in Stage II. After the recruitment is complete, the study will follow-up to calendar time $T_2$. When the final analysis is performed, the total number of 632 PFS are observed, with 461 PFS events from Stage I enrolled subjects and 171 PFS events from Stage II enrolled subjects. The number of PFS events from Stage I and Stage II enrolled subjects were determined, to make sure the number of events arrives approximately at the same calendar time.

To illustrate power calculation, we take the treatment effect information from Herbst et al. (2016). Assume PFS times are exponentially distributed. The hazard rate for treated S+ group is $\lambda_{+T} = 9.90$, and hazard rate for control treated S+ group is $\lambda_{+C} = 5.85$. The hazard rates for marker negative group $S-$ are $\lambda_{-T} = 5.14$ and $\lambda_{-C} = 5.85$. These design parameters lead to the critical values $(c_1, c_2, b_1, b_2) = (-2.878, -2.848, -2.269, -2.211)$ when $\lambda_{sen} = \lambda_{spec} = 1$, and critical values $(c_1, c_2, b_1, b_2) = (-2.878, -2.874, -2.273, -2.260)$ when $\lambda_{sen} = \lambda_{spec} = 0.8$. With a total of $N = 688$ patients and prevalence rate $p = 0.4$, we expected the power to test global hypothesis $H_1$ to be 97%, the power to test $H_{1+}$ to be 97%, and the power to test $H_{1a}$ to be 5.7%, when $\lambda_{sen} = \lambda_{spec} = 1$. However, when

$\lambda_{sen} = \lambda_{spec} = 0.8$, with a total of $N = 688$ patients and prevalence rate $p = 0.4$, we expected the power to test global hypothesis $H_1$ to be 72%, the power to test $H_{1+}$ to be 70%, and the power to test $H_{1a}$ to be 8.0%.

To illustrate sample size calculation, with the same $\alpha$ allocation, $\lambda_{sen} = \lambda_{spec} = 0.8$, and treatment effects $\delta$ = the log-hazard ratio (Dox vs Pem) = $-0.128$ for the overall cohort, and $\delta_+$ = the log-hazard ratio (Dox vs Pem) = $-0.528$ for the marker positive cohort. Trials usually aim a power of either 80% or 90% for the marker positive subset. Assuming a target of 90% power for testing $H_{0+}$, a total sample size of 1125 is needed. The global power for the composite hypothesis is 91 %, and the power for the overall cohort is 10%. If the target of 80% power is for testing $H_{1+}$, a total sample size of 855 is needed. The global power for the composite hypothesis $H_1$ is 82 %, and the power for the overall cohort $H_{1a}$ is 9.2%.

# Part III

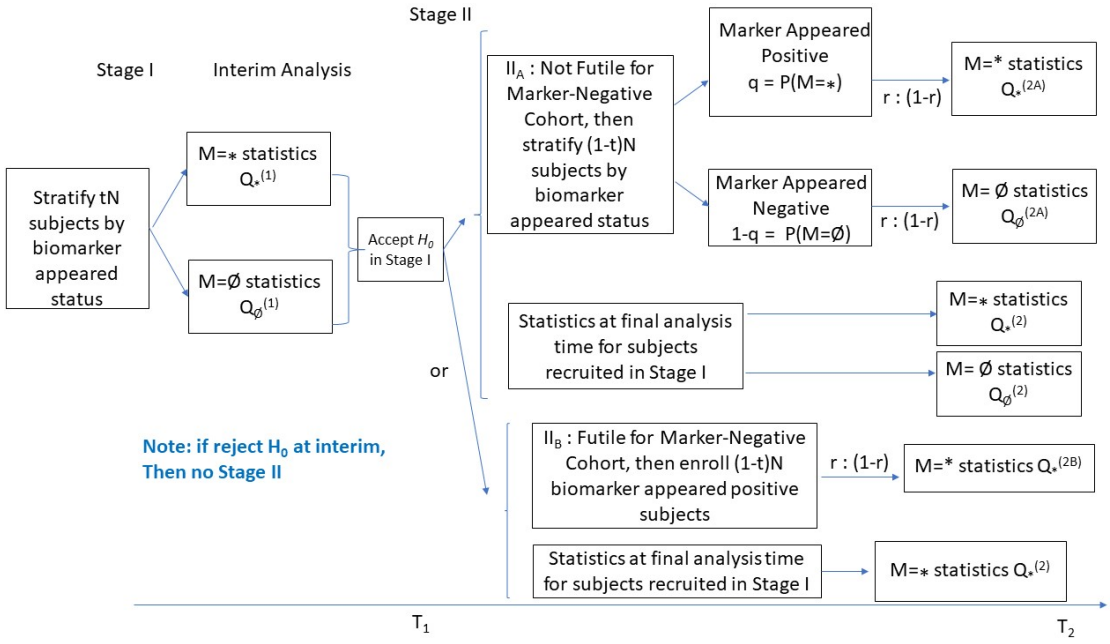
# Two-Stage Enrichment Design with Misclassification Adjustments

# Chapter 9

# Methods for biomarker misclassification adjustment

## 9.1 Design diagram

The following shows the diagram of the two-stage enrichment design based on marker appearance status.




where $Q_i^j$ is log rank statistics, where $i = *, \phi$ (marker-appeared positive, appeared negative) and $j$ =$1, 2, 2A, 2B$ for analysis time/scenarios, $M$ denotes the marker-apppeared status, $t$ represents the fraction of the total sample size $N$ that is allocated to Stage I, $q$ is the prevalence rate for biomarker-appeared positive status, and $r$ is the randomization probability to active treatment group. The definition of $Q_i^j$ can be found in 9.3.2, 9.3.3 and 9.3.4.

## 9.2 Difference between observed number of events and expected number of events in stratas by marker-appeared status

Consider the difference between observed number of death $d_{iTk}$ and the expected number of death $E_{iTk}$ in active treatment group at $k^{th}$ event time in $i^{th}$ marker-appeared strata (where $i = * :$ marker-appeared positive; $i = \phi :$ marker-appeared negative), we have

$$d_{iTk} - E_{iTk} = d_{iTk} - \frac{d_{i+k}\overline{Y}_{iTk}}{\overline{Y}_{i+k}}.$$

Let the total number of subjects (including both active treated and control) at risk be $\overline{Y}_{*+k}$ at $k^{th}$ event time. The estimated variance of $d_{*Tk} - E_{*Tk}$ is

$$\widehat{Var}(d_{*Tk} - E_{*Tk}) = \widehat{Var}(d_{*Tk}) = \frac{\overline{Y}_{*Tk}\overline{Y}_{*Ck}d_{*+k}(\overline{Y}_{*+k} - d_{*+k})}{\overline{Y}^2_{*+k}(\overline{Y}_{*+k} - 1)}.$$

This can be applied to marker-appeared negative cohort as well.

Let $Q_i$ be log-rank statistic in $i^{th}$ marker-appeared status strata:

$$Q_i = \sum_{k=1}^{K}(d_{iTk} - \frac{d_{i+k}\overline{Y}_{iTk}}{\overline{Y}_{i+k}})$$

where $i = *, \phi$. We have

$$\begin{aligned}\widehat{Var}(Q_i) &= \sum_{k=1}^{K}\widehat{Var}(d_{iTk}) \\ &= \sum_{k=1}^{K}\left\{\frac{\overline{Y}_{iTk}\overline{Y}_{iCk}d_{i+k}(\overline{Y}_{i+k} - d_{i+k})}{\overline{Y}^2_{i+k}(\overline{Y}_{i+k} - 1)}\right\}\end{aligned}$$

and under null,

$$\frac{Q_i^2}{\sqrt{\widehat{Var}(Q_i)}} \xrightarrow{d} N(0,1).$$

## 9.3 Asympototic distribution of adjusted log rank statistics with marker misclassification

### 9.3.1 Expected number of events

Consider Stage I enrolled subjects who were in marker-appeared positive group. Let $n_{*+}^{(1)} = \tau n_{*}^{(1)}$ be expected total number of subjects in marker-appeared positive group with true marker positive status and $n_{*-}^{(1)} = (1-\tau)n_{*}^{(1)}$ be expected total number of subjects in marker-appeared positive group with true marker negative status, where $n_{*}^{(1)}$ is the expected number of subjects with marker-appeared positive status.

Similarly, let $n_{\phi+}^{(1)} = (1-\eta)n_{\phi}^{(1)}$ be expected total number of subjects in marker-appeared negative group with true marker positive status and $n_{\phi-}^{(1)} = \eta n_{\phi}^{(1)}$ be expected total number of subjects in marker-appeared negative group with true marker negative status, where $n_{\phi}^{(1)}$ is the expected number of subjects with marker-appeared negative status.

Let $n_{i}^{(1)}$ be the expected number of subjects with true $i^{th}$ marker status ($i = +$ for true marker positive; $i = -$ for true marker negative).

At pre-determined analysis time (index $k$), let the probability to observe an event in true marker $i^{th}$ group with $j^{th}$ treatment $\pi_{ij}^{(k)}$, where $i = +$ for true marker positive; $i = 0$ for true marker negative; $j = T$ for active treatment; $j = C$ for control treatment; $k = 1$ for subjects recruited at Stage I at interim analysis time ; $k = 2$ for subjects recruited at Stage I at final analysis time ; $k = 2A$ for subjects recruited at Stage II under IIA scenario at final analysis time; $k = 2B$ for subjects recruited at Stage II under IIB scenario at final analysis time ; and

$$\pi_{i}^{(k)} = r\pi_{iT}^{(k)} + (1-r)\pi_{iC}^{(k)},$$

where $i = +, -$.

Then, the expected number of events at time $T_1$ for marker-appeared positive group is

$$E(D_{*}^{(1)}) = \tau n_{*}^{(1)}(r\pi_{+T}^{(1)} + (1-r)\pi_{+C}^{(1)}) + (1-\tau)n_{*}^{(1)}(r\pi_{-T}^{(1)} + (1-r)\pi_{-C}^{(1)}).$$

The expected total number of events at time $T_1$ for marker-appeared negative group is

$$E(D_\phi^{(1)}) = (1-\eta)n_\phi^{(1)}(r\pi_{+T}^{(1)} + (1-r)\pi_{+C}^{(1)}) + \eta n_\phi^{(1)}(r\pi_{-T}^{(1)} + (1-r)\pi_{-C}^{(1)}).$$

### 9.3.2 Adjusted log rank statistics at interim analysis

Let $Q_*^{(1)}$ and $Q_\phi^{(1)}$ be log rank statistics at time of interim analysis for marker-appeared positive and marker-appeared negative strata, respectively, we have

$$\begin{aligned} E(Q_*^{(1)}) &= r(1-r)\tau n_*^{(1)}(r\pi_{+T}^{(1)} + (1-r)\pi_{+C}^{(1)})^{(1)} \log\frac{\lambda_{+T}}{\lambda_{+C}} \\ &\quad + r(1-r)(1-\tau)n_*^{(1)}(r\pi_{-T}^{(1)} + (1-r)\pi_{-C}^{(1)}) \log\frac{\lambda_{-T}}{\lambda_{-C}}, \qquad (9.1) \\ E(Q_\phi^{(1)}) &= r(1-r)(1-\eta)n_\phi^{(1)}(r\pi_{+T}^{(1)} + (1-r)\pi_{+C}^{(1)}) \log\frac{\lambda_{+T}}{\lambda_{+C}} \\ &\quad + r(1-r)\eta n_\phi^{(1)}(r\pi_{-T}^{(1)} + (1-r)\pi_{-C}^{(1)}) \log\frac{\lambda_{-T}}{\lambda_{-C}}. \qquad (9.2) \end{aligned}$$

From Equations 9.1 and 9.2, we get

$$\begin{aligned} &r(1-r)n_+^{(1)}(r\pi_{+T}^{(1)} + (1-r)\pi_{+C}^{(1)}) \log\frac{\lambda_{+T}}{\lambda_{+C}} \\ &\qquad = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)} \frac{\eta(1-q)E(Q_*^{(1)}) - (1-\tau)qE(Q_\phi^{(1)})}{\tau+\eta-1}, \\ &r(1-r)n_-^{(1)}(r\pi_{-T}^{(1)} + (1-r)\pi_{-C}^{(1)}) \log\frac{\lambda_{-T}}{\lambda_{-C}} \\ &\qquad = \frac{(1-\tau)q + \eta(1-q)}{q(1-q)} \frac{-(1-\eta)(1-q)E(Q_*^{(1)}) + \tau qE(Q_\phi^{(1)})}{\tau+\eta-1}. \end{aligned}$$

Let

$$\begin{aligned} Q_+^{(1)} &= \frac{\tau q + (1-\eta)(1-q)}{q(1-q)} \frac{\eta(1-q)Q_*^{(1)} - (1-\tau)qQ_\phi^{(1)}}{\tau+\eta-1}, \\ Q_-^{(1)} &= \frac{(1-\tau)q + \eta(1-q)}{q(1-q)} \frac{-(1-\eta)(1-q)Q_*^{(1)} + \tau qQ_\phi^{(1)}}{\tau+\eta-1}, \end{aligned}$$

we have

$$E(Q_+^{(1)}) = r(1-r)n_+^{(1)}(r\pi_{+T}^{(1)} + (1-r)\pi_{+C}^{(1)}) \log\frac{\lambda_{+T}}{\lambda_{+C}},$$

$$E(Q_-^{(1)}) = r(1-r)n_-^{(1)}(r\pi_{-T}^{(1)} + (1-r)\pi_{-C}^{(1)}) \log \frac{\lambda_{-T}}{\lambda_{-C}},$$

and their variances are

$$\begin{aligned} Var(Q_+^{(1)}) &= \left(\sigma_+^{(1)}\right)^2 \\ &= \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left((1-q)^2\eta^2 Var(Q_*^{(1)}) + q^2(1-\tau)^2 Var(Q_\phi^{(1)})\right), \end{aligned}$$

$$\begin{aligned} Var(Q_-^{(1)})) &= \left(\sigma_-^{(1)}\right)^2 \\ &= \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left((1-q)^2(1-\eta)^2 Var(Q_*^{(1)}) + q^2\tau^2 Var(Q_\phi^{(1)})\right). \end{aligned}$$

Therefore the test statistics $Q_+^{(1)}$ and $Q_-^{(1)}$, unbiased estimators for the true effects, are constructed as test statistics to test treatment effects on true marker positive and negative groups, respectively. This also implies that the adjusted log rank statistic for the overall population (based on true marker positive group and true marker negative group) at the interim analysis time $T_1$ is

$$Q^{(1)} = pQ_+^{(1)}/\sigma_+^{(1)} + (1-p)Q_-^{(1)}/\sigma_-^{(1)}$$

with its variance

$$Var(Q^{(1)}) = p^2 + (1-p)^2 - 2p(1-p)(\eta(1-\eta)(1-q)^2 Var(Q_*^{(1)}) + \tau(1-\tau)q^2 Var(Q_\phi^{(1)}))/(\sigma_+^{(1)}\sigma_-^{(1)})$$

The standardized adjusted log rank statistic for the true marker positive group at the interim analysis $T_1$ is

$$Z_+^{(1)} = \frac{Q_1^{(1)}}{\sqrt{Var(Q_+^{(1)})}} = \frac{Q_+^{(1)}}{\sigma_+^{(1)}},$$

where

$$\sigma_+^{(1)} = \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right) \sqrt{(1-q)^2\eta^2 Var(Q_*^{(1)}) + q^2(1-\tau)^2 Var(Q_\phi^{(1)})}.$$

The standardized adjusted log rank statistic for true marker negative group at interim analysis $T_1$ is

$$Z_-^{(1)} = \frac{Q_-^{(1)}}{\sqrt{Var(Q_-^{(1)})}} = \frac{Q_-^{(1)}}{\sigma_-^{(1)}},$$

where

$$\sigma_-^{(1)} = \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)\sqrt{(1-q)^2(1-\eta)^2 Var(Q_*^{(1)}) + q^2\tau^2 Var(Q_\phi^{(1)})}.$$

The standardized adjusted log rank statistic for overall population at interim analysis $T_1$ is

$$Z^{(1)} = \frac{Q^{(1)}}{\sqrt{Var(Q^{(1)})}} = \frac{Q^{(1)}}{\sigma^{(1)}} = \frac{pZ_+^{(1)} + (1-p)Z_-^{(1)}}{\sqrt{p^2+(1-p)^2}}$$

where

$$(\sigma^{(1)})^2 = p^2+(1-p)^2-2p(1-p)(\eta(1-\eta)(1-q)^2 Var(Q_*^{(1)})+\tau(1-\tau)q^2 Var(Q_\phi^{(1)}))/(\sigma_+^{(1)}\sigma_0^{(1)}).$$

At interim analysis time $T_1$, a futility criterion ($Q_-^{(1)} \geq c_0$) for marker-negative subjects will be used, to decide the enrollment pattern for Stage II, as two mutually exclusive scenarios $II_A$ and $II_B$:

**Scenario** $II_A$ : Enroll full population (both marker-appeared positive and marker-appeared negative subjects) for the remaining $(1-t)N$, when $Q_-^{(1)} < c_0$;

**Scenario** $II_B$ : Stop enrolling marker-negative subjects and only enroll $(1-t)N$ marker-appeared positive subjects for Stage II, when $Q_-^{(1)} \geq c_0$.

### 9.3.3 Adjusted log rank statistics at final analysis under Scenario $II_A$

**Scenario** $II_A$ : We enroll $(1-t)N$ marker-unselected subjects, the corresponding test statistics at the final analysis time $t_2$ are derived as follows.

Let $Q_*^{(2)}$ and $Q_\phi^{(2)}$ be log rank statistics at final analysis $T_2$ for the subjects recruited at

Stage I:

$$E(Q_*^{(2)}) = r(1-r)\tau n_*^{(1)}(r\pi_{+T}^{(2)} + (1-r)\pi_{+C}^{(2)}) \log \frac{\lambda_{+T}}{\lambda_{+C}} + r(1-r)(1-\tau)n_*^{(1)}(r\pi_{-T}^{(2)} + (1-r)\pi_{-C}^{(2)}) \log \frac{\lambda_{-T}}{\lambda_{-C}}, \tag{9.3}$$

$$E(Q_\phi^{(2)}) = r(1-r)(1-\eta)n_\phi^{(1)}(r\pi_{+T}^{(2)} + (1-r)\pi_{+C}^{(2)}) \log \frac{\lambda_{+T}}{\lambda_{+C}} + r(1-r)\eta n_\phi^{(1)}(r\pi_{-T}^{(2)} + (1-r)\pi_{-C}^{(2)}) \log \frac{\lambda_{-T}}{\lambda_{-C}}. \tag{9.4}$$

From Equations 9.3 and 9.4, we get

$$r(1-r)n_+^{(1)}(r\pi_{+T}^{(2)} + (1-r)\pi_{+C}^{(2)}) \log \frac{\lambda_{+T}}{\lambda_{+C}} = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)} \frac{\eta(1-q)E(Q_*^{(2)}) - (1-\tau)qE(Q_\phi^{(2)})}{\tau+\eta-1},$$

$$r(1-r)n_-^{(1)}(r\pi_{-T}^{(2)} + (1-r)\pi_{-C}^{(2)}) \log \frac{\lambda_{-T}}{\lambda_{-C}} = \frac{(1-\tau)q + \eta(1-q)}{q(1-q)} \frac{-(1-\eta)(1-q)E(Q_*^{(2)}) + \tau qE(Q_\phi^{(2)})}{\tau+\eta-1}.$$

Let

$$Q_+^{(2)} = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)} \frac{\eta(1-q)Q_*^{(2)} - (1-\tau)qQ_\phi^{(2)}}{\tau+\eta-1},$$

$$Q_-^{(2)} = \frac{(1-\tau)q + \eta(1-q)}{q(1-q)} \frac{-(1-\eta)(1-q)Q_*^{(2)} + \tau qQ_\phi^{(2)}}{\tau+\eta-1},$$

we have

$$E(Q_+^{(2)}) = r(1-r)n_+^{(1)}(r\pi_{+T}^{(2)} + (1-r)\pi_{+C}^{(2)}) \log \frac{\lambda_{+T}}{\lambda_{+C}},$$

$$E(Q_-^{(2)}) = r(1-r)n_-^{(1)}(r\pi_{-T}^{(2)} + (1-r)\pi_{-C}^{(2)}) \log \frac{\lambda_{-T}}{\lambda_{-C}}.$$

and their variances are

$$Var(Q_+^{(2)}) = \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left((1-q)^2\eta^2 Var(Q_*^{(2)}) + q^2(1-\tau)^2 Var(Q_\phi^{(2)})\right),$$

$$Var(Q_-^{(2)}) = \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left((1-q)^2(1-\eta)^2 Var(Q_*^{(2)}) + q^2\tau^2 Var(Q_\phi^{(2)})\right).$$

Therefore the test statistics $Q_+^{(2)}$ and $Q_-^{(2)}$, unbiased estimators for the true effects, are constructed as test statistics to test treatment effects on true marker positive and negative groups, respectively, for the subjects enrolled at Stage I at final analysis time $T_2$.

Let $Q_*^{(2A)}$ and $Q_\phi^{(2A)}$ be the log rank statistics at the final analysis time $T_2$ for the subjects recruited at Stage II, then their expected values are

$$\begin{aligned} E(Q_*^{(2A)}) &= r(1-r)\tau n_*^{(2A)}(r\pi_{+T}^{(2A)} + (1-r)\pi_{+C}^{(2A)}) \log\frac{\lambda_{+T}}{\lambda_{+C}} \\ &\quad + r(1-r)(1-\tau)n_*^{(2A)}(r\pi_{-T}^{(2A)} + (1-r)\pi_{-C}^{(2A)}) \log\frac{\lambda_{-T}}{\lambda_{-C}}, \qquad (9.5) \\ E(Q_\phi^{(2A)}) &= r(1-r)(1-\eta)n_\phi^{(2A)}(r\pi_{+T}^{(2A)} + (1-r)\pi_{+C}^{(2A)}) \log\frac{\lambda_{+T}}{\lambda_{+C}} \\ &\quad + r(1-r)\eta n_\phi^{(2A)}(r\pi_{-T}^{(2A)} + (1-r)\pi_{-C}^{(2A)}) \log\frac{\lambda_{-T}}{\lambda_{-C}}. \qquad (9.6) \end{aligned}$$

From Equations 9.5 and 9.6, we get

$$\begin{aligned} &r(1-r)n_+^{(2A)}(r\pi_{+T}^{(2A)} + (1-r)\pi_{+C}^{(2A)}) \log\frac{\lambda_{+T}}{\lambda_{+C}} \\ &\qquad = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)} \frac{\eta(1-q)E(Q_*^{(2A)}) - (1-\tau)qE(Q_\phi^{(2A)})}{\tau+\eta-1}, \end{aligned}$$

$$\begin{aligned} &r(1-r)n_-^{(2A)}(r\pi_{-T}^{(2A)} + (1-r)\pi_{-C}^{(2A)}) \log\frac{\lambda_{-T}}{\lambda_{-C}} \\ &\qquad = \frac{(1-\tau)q + \eta(1-q)}{q(1-q)} \frac{-(1-\eta)(1-q)E(Q_*^{(2A)}) + \tau qE(Q_\phi^{(2A)})}{\tau+\eta-1}. \end{aligned}$$

Let

$$Q_+^{(2A)} = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)} \frac{\eta(1-q)Q_*^{(2A)} - (1-\tau)qQ_\phi^{(2A)}}{\tau+\eta-1},$$

$$Q_{-}^{(2A)} = \frac{(1-\tau)q + \eta(1-q)}{q(1-q)} \frac{-(1-\eta)(1-q)Q_{*}^{(2A)} + \tau q Q_{\phi}^{(2A)}}{\tau + \eta - 1},$$

we have

$$E(Q_{+}^{(2A)}) = r(1-r)n_{+}^{(2A)}(r\pi_{+T}^{(2A)} + (1-r)\pi_{+C}^{(2A)}) \log \frac{\lambda_{+T}}{\lambda_{+C}},$$
$$E(Q_{-}^{(2A)}) = r(1-r)n_{-}^{(2A)}(r\pi_{-T}^{(2A)} + (1-r)\pi_{-C}^{(2A)}) \log \frac{\lambda_{-T}}{\lambda_{-C}},$$

and their variances are

$$Var(Q_{+}^{(2A)}) = \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left(((1-q)\eta)^2 Var(Q_{*}^{(2A)}) + (q(1-\tau))^2 Var(Q_{\phi}^{(2A)})\right),$$

$$Var(Q_{-}^{(2A)}) = \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left(((1-q)(1-\eta))^2 Var(Q_{*}^{(2A))} + (q\tau)^2 Var(Q_{\phi}^{(2A)})\right).$$

Therefore the test statistics $Q_{+}^{(2A)}$ and $Q_{-}^{(2A)}$, unbiased estimators for the true effects, are constructed as test statistics to test treatment effects on true marker positive and negative groups, respectively, for subjects enrolled at Stage II under Scenario $II_A$. This also implies that, at final analysis time $T_2$, the adjusted log rank statistics and their variances for the true marker positive group $Q_{+}$ and true marker negative group $Q_{-}$ can be found

$$Q_{+} = Q_{+}^{(2)} + Q_{+}^{(2A)},$$
$$Var(Q_{+}) = Var(Q_{+}^{(2)}) + Var(Q_{+}^{(2A)}),$$

and

$$Q_{-} = Q_{-}^{(2)} + Q_{-}^{(2A)},$$
$$Var(Q_{-}) = Var(Q_{-}^{(2)}) + Var(Q_{-}^{(2A)}).$$

The log rank statistic for overall population after misclassification adjustment is

$$Q = pQ_{+}/\sigma_{+} + (1-p)Q_{-}/\sigma_{-},$$

and its variance

$$Var(Q) = p^2 + (1-p)^2 - 2p(1-p)(\eta(1-\eta)(1-q)^2 Var(Q_*) + \tau(1-\tau)q^2 Var(Q_\phi))/(\sigma_+\sigma_-).$$

The standardized adjusted log rank statistic for true marker positive group at final analysis $T_2$ is

$$Z_+ = \frac{Q_+}{\sqrt{Var(Q_+)}} = \frac{Q_+^{(2)} + Q_+^{(2A)}}{\sigma_+}$$

where

$$\begin{aligned}\sigma_+^2 = &\left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left((1-q)^2\eta^2 Var(Q_*^{(2)}) + q^2(1-\tau)^2 Var(Q_\phi^{(2)})\right) \\ &+ \left(\frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left(((1-q)\eta)^2 Var(Q_*^{(2A)}) + (q(1-\tau))^2 Var(Q_\phi^{(2A)})\right).\end{aligned}$$

The standardized adjusted log rank statistic for true marker negative group at final analysis $T_2$ is

$$Z_- = \frac{Q_-}{\sqrt{Var(Q_-)}} = \frac{Q_-}{\sigma_-^{(1)}},$$

where

$$\begin{aligned}\sigma_-^2 = &\left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left((1-q)^2(1-\eta)^2 Var(Q_*^{(2)}) + q^2\tau^2 Var(Q_\phi^{(2)})\right) \\ &+ \left(\frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}\right)^2 \left(((1-q)(1-\eta))^2 Var(Q_*^{(2A)}) + (q\tau)^2 Var(Q_\phi^{(2A)})\right).\end{aligned}$$

The standardized adjusted log rank statistic for overall population at final analysis $T_2$:

$$Z = \frac{Q}{\sqrt{Var(Q)}} = \frac{Q}{\sigma} = \frac{pZ_+ + (1-p)Z_-}{\sigma},$$

where

$$\sigma^2 = \; p^2 + (1-p)^2 - 2p(1-p)(\eta(1-\eta)(1-q)^2 Var(Q_*) + \tau(1-\tau)q^2 Var(Q_\phi))/(\sigma_+\sigma_-).$$

### 9.3.4 Adjusted log rank statistics at final analysis under Scenario $II_B$

1) **Construct Log rank Statistics $Q_+^{(2B)}$ at Final Analysis**

We enroll $(1-t)N$ marker-appeared positive subjects, the corresponding test statistics at final analysis time $T_2$ are as follows. Let $Q_*^{(2B)}$ be log rank statistic at final analysis time $T_2$ for the subjects recruited at Stage II under IIB scenario, while $Q_\phi^{(2B)}$ is no longer available because the futility rule is applied and the marker-appeared negative stratum is not enrolled. Previously we showed that we can obtain adjusted log rank statistics at time $T_2$ for true marker positive group for subjects recruited at Stage II

$$\begin{aligned} E(Q_+^{(2B)}) &= n_+^{(2B)}(r\pi_{+T}^{(2B)} + (1-r)\pi_{+C}^{(2B)}) \log \frac{\lambda_{+T}}{\lambda_{+C}} \\ &= \frac{\tau q + (1-\eta)(1-q)}{q(1-q)} \frac{\eta(1-q)E(Q_*^{(2B)}) - (1-\tau)qE(Q_\phi^{(2B)})}{\tau+\eta-1}, \end{aligned}$$

where $Q_*^{(2B)}$ is observed and $Q_\phi^{(2B)}$ is not observed. Next, a method is shown to estimate $Q_\phi^{(2B)}$, then use it to obtain $Q_+^{(2B)}$ at the final analysis.

2) **Methods to solve $Q_\phi^{(2B)}$**

$$\begin{aligned} E(Q_\phi^{(2B)}) = r(1-r)(1-\eta)n_\phi^{(2B)}(r\pi_{+T}^{(2B)} + (1-r)\pi_{+C}^{(2B)}) \log \frac{\lambda_{+T}}{\lambda_{+C}} \\ + r(1-r)\eta n_\phi^{(2B)}(r\pi_{-T}^{(2B)} + (1-r)\pi_{-C}^{(2B)}) \log \frac{\lambda_{-T}}{\lambda_{-C}} \end{aligned}$$

- If assuming proportional hazard and exponential distribution, we need $\pi_{+T}^{(2B)}$, $\pi_{+C}^{(2B)}$, $\pi_{-T}^{(2B)}$, $\pi_{-C}^{(2B)}$ to estimate $Q_\phi^{(2B)}$. Similarly, we need $\pi_{+T}^{(2)}, \pi_{+C}^{(2)}, \pi_{-T}^{(2)}, \pi_{-C}^{(2)}$ to estimate $U_\phi^{(2)}$ (will be presented later).

$$\begin{array}{rl} & \text{Needed} \\ Q_\phi^{(2B)}: & \pi_{+T}^{(2B)}, \pi_{+C}^{(2B)}, \pi_{-T}^{(2B)}, \pi_{-C}^{(2B)} \\ U_\phi^{(2)}: & \pi_{+T}^{(2)}, \pi_{+C}^{(2)}, \pi_{-T}^{(2)}, \pi_{-C}^{(2)} \end{array}$$

- We use method described by Schoenfeld (1983). Denote an accrual period $a$ as the period during which patients enter the study, and a follow-up period $f$ as

the period from the end of accrual until the analysis of the data. The proportion of patients that will survive is the average of the survival curve from Time $f$ to Time $a+f$, provided that patients enter the trial at a constant rate. If the survival is exponentially distributed within each treatment group with the true marker status, one can estimate the proportion expected to die as $\pi_{ij}^{(k)}$ from

$$\pi_{ij} = 1 - \exp\left(-\frac{f}{\lambda_{ij}}\right)\left[1 - \exp\left(-\frac{a}{\lambda_{ij}}\right)\right] \Big/ \left(\frac{a}{\lambda_{ij}}\right)$$

where $a$ is accrual time and $f$ is follow-up time.

- Use information from interim analysis time $T_1$, we can obtain a total of 4 equations

$$\begin{aligned}
*T &: (1-\tau)\pi_{-T}^{(1)} + \tau\pi_{+T}^{(1)} = E(\tfrac{d_{*T}}{n_{*T}}),\\
*C &: (1-\tau)\pi_{-C}^{(1)} + \tau\pi_{+C}^{(1)} = E(\tfrac{d_{*C}}{n_{*C}}),\\
\phi T &: \eta\pi_{-T}^{(1)} + (1-\eta)\pi_{+T}^{(1)} = E(\tfrac{d_{\phi T}}{n_{\phi T}}),\\
\phi C &: \eta\pi_{-C}^{(1)} + (1-\eta)\pi_{+C}^{(1)} = E(\tfrac{d_{\phi C}}{n_{\phi C}}),
\end{aligned}$$

where each $\pi_{ij}$ is a function of $\lambda_{ij}$ by Schoenfeld method, where $i = +, -$; $j = T, C$. Now, we have 4 linear equations to solve 4 unknowns, we can obtain close form solution for $\lambda_{+T}, \lambda_{+C}, \lambda_{-T}, \lambda_{-C}$.

- With $\lambda_{ij}$, we can estimate $\pi_{ij}^{(k)}$, where $k = 2, 2B$, since we know the accrual time and follow-up time for subject enrolled after interim analysis.

- Assume the **constant log hazard ratios** $\log\frac{\lambda_{+T}}{\lambda_{+C}}$ and $\log\frac{\lambda_{-T}}{\lambda_{-C}}$, and estimate them using the information from interim analysis.

$$\log\frac{\lambda_{+T}}{\lambda_{+C}} = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)} \frac{\eta(1-q)E(Q_*^{(1)}) - (1-\tau)qE(Q_\phi^{(1)})}{\tau+\eta-1} \times \frac{1}{n_+^{(1)}(r\pi_{+T}^{(1)} + (1-r)\pi_{+C}^{(1)})},$$

$$\log\frac{\lambda_{-T}}{\lambda_{-C}} = \frac{(1-\tau)q + \eta(1-q)}{q(1-q)} \frac{-(1-\eta)(1-q)E(Q_*^{(1)}) + \tau qE(Q_\phi^{(1)})}{\tau+\eta-1}$$

$$\times \frac{1}{n_-^{(1)}(r\pi_{-T}^{(1)} + (1-r)\pi_{-C}^{(1)})}.$$

3) **Estimate $Q_\phi^{(2B)}$**

With $\lambda_{ij}$ in hand, we can estimate $\pi_{ij}^{(2B)}$, and then $Q_\phi^{(2B)}$ as follows.

$$\begin{aligned} Q_\phi^{(2B)} = n_\phi^{(2B)}\{&(1-\eta)\frac{r\pi_{+T}^{(2B)} + (1-r)\pi_{+C}^{(2B)}}{n_+^{(1)}[r\pi_{+T}^{(1)} + (1-r)\pi_{+C}^{(1)}]} \\ &\times A[\eta(1-q)Q_*^{(1)} - (1-\tau)qQ_\phi^{(1)}] \\ &+ \eta\frac{r\pi_{-T}^{(2B)} + (1-r)\pi_{-C}^{(2B)}}{n_-^{(1)}[r\pi_{-T}^{(1)} + (1-r)\pi_{-C}^{(1)}]}B[-(1-\eta)(1-q)Q_*^{(1)} + \tau qQ_\phi^{(1)}]\}, \end{aligned}$$

where

$$\begin{aligned} A &= \frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)}, \\ B &= \frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)}. \end{aligned}$$

4) **Estimate $U_\phi^{(2)}$**

Apply the same method to estimate $\pi_{ij}^{(2)}$ and the unobserved log rank statistic $U_\phi^{(2)}$ for marker-appeared negative group at final analysis time under scenario $II_B$, to distinguish it from $Q_\phi^{(2)}$, which is observed under $II_A$ scenario.

$$\begin{aligned} U_\phi^{(2)} = n_\phi^{(1)}\{&(1-\eta)\frac{r\pi_{+T}^{(2)} + (1-r)\pi_{+C}^{(2)}}{n_+^{(1)}[r\pi_{+T}^{(1)} + (1-r)\pi_{+C}^{(1)}]} \\ &\times A[\eta(1-q)Q_*^{(1)} - (1-\tau)qQ_\phi^{(1)}] \\ &+ \eta\frac{r\pi_{-T}^{(2)} + (1-r)\pi_{-C}^{(2)}}{n_-^{(1)}[r\pi_{-T}^{(1)} + (1-r)\pi_{-C}^{(1)}]}B[-(1-\eta)(1-q)Q_*^{(1)} + \tau qQ_\phi^{(1)}]\} \end{aligned}$$

5) **Estimate log rank statistic at final analysis time uner $II_B$**

Finally, under scenario IIB, let $\overline{Q}_+$ be log rank statistic for marker positive group at

final analysis time $T_2$. *Test statistic* $\overline{Q}_+$ *under* $II_B$ *Scenario* is

$$\begin{aligned}
\overline{Q}_+ &= U_+^{(2)} + Q_+^{(2B)} \\
&= A\eta(1-q)(Q_*^{(2)} + Q_*^{(2B)}) - A(1-\tau)q(U_\phi^{(2)} + Q_\phi^{(2B)}) \\
&= A\eta(1-q)(Q_*^{(2)} + Q_*^{(2B)}) \\
&\quad - A(1-\tau)q[E(n_\phi^{(2B)})\eta(1-\eta)(1-q)(\frac{A\widehat{\pi}_+^{(2B)}}{n_+^{(1)}\widehat{\pi}_+^{(1)}} - \frac{B\widehat{\pi}_-^{(2B)}}{n_-^{(1)}\widehat{\pi}_-^{(1)}}) \\
&\quad + n_\phi^{(1)}\eta(1-\eta)(1-q)(\frac{A\widehat{\pi}_+^{(2)}}{n_+^{(1)}\widehat{\pi}_+^{(1)}} - \frac{B\widehat{\pi}_-^{(2)}}{n_-^{(1)}\widehat{\pi}_-^{(1)}})]Q_*^{(1)} \\
&\quad - A(1-\tau)q[-(1-\eta)(1-\tau)qA(\frac{E(n_\phi^{(2B)})\widehat{\pi}_+^{(2B)}}{n_+^{(1)}\widehat{\pi}_+^{(1)}} + \frac{n_\phi^{(1)}\widehat{\pi}_+^{(2)}}{n_-^{(1)}\widehat{\pi}_+^{(1)}}) \\
&\quad + \eta(1-\tau)qB(\frac{E(n_\phi^{(2B)})\widehat{\pi}_+^{(2B)}}{n_+^{(1)}\widehat{\pi}_+^{(1)}} + \frac{n_\phi^{(1)}\widehat{\pi}_+^{(2)}}{n_-^{(1)}\widehat{\pi}_+^{(1)}})]Q_\phi^{(1)}.
\end{aligned}$$

In addition, we have

$$\begin{aligned}
E(\overline{Q}_+) &= E(U_+^{(2)}) + E(Q_+^{(2B)}) \\
&= A\eta(1-q)(E(Q_*^{(2)}) + E(Q_*^{(2B)})) - A(1-\tau)q(E(U_\phi^{(2)}) + E(Q_\phi^{(2B)})) \\
&= A\eta(1-q)(E(Q_*^{(2)}) + E(Q_*^{(2B)})) \\
&\quad - A(1-\tau)q[n_\phi^{(2B)}\eta(1-\eta)(1-q)(\frac{A\pi_+^{(2B)}}{n_+^{(1)}\pi_+^{(1)}} - \frac{B\pi_-^{(2B)}}{n_-^{(1)}\pi_-^{(1)}}) \\
&\quad + n_\phi^{(1)}\eta(1-\eta)(1-q)(\frac{A\pi_+^{(2)}}{n_+^{(1)}\pi_+^{(1)}} - \frac{B\pi_-^{(2)}}{n_-^{(1)}\pi_-^{(1)}})]E(Q_*^{(1)}) \\
&\quad - A(1-\tau)q[-(1-\eta)(1-\tau)qA(\frac{n_\phi^{(2B)}\pi_+^{(2B)}}{n_+^{(1)}\pi_+^{(1)}} + \frac{n_\phi^{(1)}\pi_+^{(2)}}{n_-^{(1)}\pi_+^{(1)}}) \\
&\quad + \eta(1-\tau)qB(\frac{n_\phi^{(2B)}\pi_+^{(2B)}}{n_+^{(1)}\pi_+^{(1)}} + \frac{n_\phi^{(1)}\pi_+^{(2)}}{n_-^{(1)}\pi_+^{(1)}})]E(Q_\phi^{(1)}), \\
Var(\overline{Q}_+) &= Var(U_+^{(2)}) + Var(Q_+^{(2B)}) \\
&= \{A\eta(1-q)\}^2(Var(Q_*^{(2)}) + Var(Q_*^{(2B)})) \\
&\quad + \{A(1-\tau)q[n_\phi^{(2B)}\eta(1-\eta)(1-q)(\frac{A\pi_+^{(2B)}}{n_+^{(1)}\pi_+^{(1)}} - \frac{B\pi_-^{(2B)}}{n_-^{(1)}\pi_-^{(1)}})
\end{aligned}$$

$$+ n_{\phi}^{(1)}\eta(1-\eta)(1-q)(\frac{A\pi_{+}^{(2)}}{n_{+}^{(1)}\pi_{+}^{(1)}} - \frac{B\pi_{-}^{(2)}}{n_{-}^{(1)}\pi_{-}^{(1)}})]\}^2 Var(Q_{*}^{(1)})$$
$$+ \{A(1-\tau)q[-(1-\eta)(1-\tau)qA(\frac{n_{\phi}^{(2B)}\pi_{+}^{(2B)}}{n_{+}^{(1)}\pi_{+}^{(1)}} + \frac{n_{\phi}^{(1)}\pi_{+}^{(2)}}{n_{-}^{(1)}\pi_{+}^{(1)}})$$
$$+ \eta(1-\tau)qB(\frac{n_{\phi}^{(2B)}\pi_{+}^{(2B)}}{n_{+}^{(1)}\pi_{+}^{(1)}} + \frac{n_{\phi}^{(1)}\pi_{+}^{(2)}}{n_{-}^{(1)}\pi_{+}^{(1)}})]\}^2 Var(Q_{\phi}^{(1)})$$
$$- \{A\eta(1-q)\}A(1-\tau)q[n_{\phi}^{(2B)}\eta(1-\eta)(1-q)(\frac{A\pi_{1t}^{(2B)}}{n_{+}^{(1)}\pi_{+}^{(1)}} - \frac{B\pi_{-}^{(2B)}}{n_{-}^{(1)}\pi_{-}^{(1)}})$$
$$+ n_{\phi}^{(1)}\eta(1-\eta)(1-q)(\frac{A\pi_{+}^{(2)}}{n_{+}^{(1)}\pi_{1t}^{(1)}} - \frac{B\pi_{-}^{(2)}}{n_{-}^{(1)}\pi_{-}^{(1)}})]Var(Q_{*}^{(1)}).$$

Let $\overline{\sigma}_{+} = \sqrt{Var(\overline{Q}_{+})}$, then we have

$$\overline{Z}_{+} = \frac{\overline{Q}_{+}}{\overline{\sigma}_{+}}.$$

## 9.4 Correlations between standardized adjusted log rank statistics

Let

$$A = \frac{\tau q + (1-\eta)(1-q)}{q(1-q)(\tau+\eta-1)},$$
$$B = \frac{(1-\tau)q + \eta(1-q)}{q(1-q)(\tau+\eta-1)},$$
$$F_1 = (1-q)^2\eta^2 Var(Q_{*}^{(1)}) + q^2(1-\tau)^2 Var(Q_{\phi}^{(1)}),$$
$$F_2 = (1-q)^2\eta^2 Var(Q_{*}) + q^2(1-\tau)^2 Var(Q_{\phi}),$$
$$G_1 = \eta(1-\eta)(1-q)^2 Var(Q_{*}^{(1)}) + \tau(1-\tau)q^2 Var(Q_{\phi}^{(1)}),$$
$$G_2 = \eta(1-\eta)(1-q)^2 Var(Q_{*}) + \tau(1-\tau)q^2 Var(Q_{\phi}),$$
$$H_1 = (1-q)^2(1-\eta)^2 Var(Q_{*}^{(1)}) + q^2\tau^2 Var(Q_{\phi}^{(1)}),$$
$$H_2 = (1-q)^2(1-\eta)^2 Var(Q_{*}) + q^2\tau^2 Var(Q_{\phi}),$$
$$C_1 = \frac{p}{\sqrt{p^2 + (1-p)^2 - 2p(1-p)G_1}},$$
$$D_1 = \frac{1-p}{\sqrt{p^2 + (1-p)^2 - 2p(1-p)G_1}},$$

$$C_2 = \frac{p}{\sqrt{p^2+(1-p)^2-2p(1-p)G_2}},$$
$$D_2 = \frac{1-p}{\sqrt{p^2+(1-p)^2-2p(1-p)G_2}},$$
$$E = C_1C_2\frac{\sigma_+^{(1)}}{\sigma_+} + D_1D_2\frac{\sigma_-^{(1)}}{\sigma_-} - ABG_1\left(\frac{C_1D_2}{\sigma_-\sigma_1^{(1)}} + \frac{C_2D_1}{\sigma_+\sigma_-^{(1)}}\right),$$

and

$$\begin{aligned}
E_1 = Var(Q_*^{(1)})\{&\eta^2(1-q)^2 \\
&- \eta^2(1-q)^2(1-\tau)q(1-\eta)[A\frac{n_\phi^{(2B)}(r\pi_{+T}^{(2B)}+(1-r)\pi_{+C}^{(2B)})}{n_+^{(1)}(r\pi_{+T}^{(1)}+(1-r)\pi_{+C}^{(1)})} \\
&+ A\frac{n_\phi^{(1)}(r\pi_{+T}^{(2)}+(1-r)\pi_{+C}^{(2)})}{n_+^{(1)}(r\pi_{+T}^{(1)}+(1-r)\pi_{+C}^{(1)})} - B\frac{n_\phi^{(2B)}(r\pi_{-T}^{(2B)}+(1-r)\pi_{-C}^{(2B)})}{n_-^{(1)}(r\pi_{-T}^{(1)}+(1-r)\pi_{-C}^{(1)})} \\
&- B\frac{n_\phi^{(1)}(r\pi_{-T}^{(2)}+(1-r)\pi_{-C}^{(2)})}{n_-^{(1)}(r\pi_{-T}^{(1)}+(1-r)\pi_{-C}^{(1)})}]\} \\
&+ Var(Q_\phi^{(1)})(1-\tau)^2q^2\{-(1-\tau)q(1-\eta)[A\frac{n_\phi^{(2B)}(r\pi_{+T}^{(2B)}+(1-r)\pi_{+C}^{(2B)})}{n_+^{(1)}(r\pi_{+T}^{(1)}+(1-r)\pi_{+C}^{(1)})} \\
&+ A\frac{n_\phi^{(1)}(r\pi_{+T}^{(2)}+(1-r)\pi_{+C}^{(2)})}{n_+^{(1)}(r\pi_{+T}^{(1)}+(1-r)\pi_{+C}^{(1)})}] \\
&+ \tau q\eta[B\frac{n_\phi^{(2B)}(r\pi_{-T}^{(2B)}+(1-r)\pi_{-C}^{(2B)})}{n_-^{(1)}(r\pi_{-T}^{(1)}+(1-r)\pi_{-C}^{(1)})} - B\frac{n_\phi^{(1)}(r\pi_{-T}^{(2)}+(1-r)\pi_{-C}^{(2)})}{n_-^{(1)}(r\pi_{-T}^{(1)}+(1-r)\pi_{-C}^{(1)})}]\},
\end{aligned}$$

$$\begin{aligned}
E_2 = Var(Q_*^{(1)})\{&-\eta(1-\eta)(1-q)^2 \\
&+ \eta(1-\eta)^2q(1-q)^2(1-\tau)[A\frac{n_\phi^{(2B)}(r\pi_{+T}^{(2B)}+(1-r)\pi_{+C}^{(2B)})}{n_+^{(1)}(r\pi_{+T}^{(1)}+(1-r)\pi_{+C}^{(1)})} \\
&+ A\frac{n_\phi^{(1)}(r\pi_{+T}^{(2)}+(1-r)\pi_{+C}^{(2)})}{n_+^{(1)}(r\pi_{+T}^{(1)}+(1-r)\pi_{+C}^{(1)})} - B\frac{n_\phi^{(2B)}(r\pi_{-T}^{(2B)}+(1-r)\pi_{-C}^{(2B)})}{n_-^{(1)}(r\pi_{-T}^{(1)}+(1-r)\pi_{-C}^{(1)})} \\
&- B\frac{n_\phi^{(1)}(r\pi_{-T}^{(2)}+(1-r)\pi_{-C}^{(2)})}{n_-^{(1)}(r\pi_{-T}^{(1)}+(1-r)\pi_{-C}^{(1)})}]\} \\
&+ Var(Q_\phi^{(1)})(1-\tau)\tau q^2\{q(1-\tau)(1-\eta)[A\frac{n_\phi^{(2B)}(r\pi_{+T}^{(2B)}+(1-r)\pi_{+C}^{(2B)})}{n_+^{(1)}(r\pi_{+T}^{(1)}+(1-r)\pi_{+C}^{(1)})}
\end{aligned}$$

$$+ A\frac{n_{\phi}^{(1)}(r\pi_{+T}^{(2)} + (1-r)\pi_{+C}^{(2)})}{n_{+}^{(1)}(r\pi_{+T}^{(1)} + (1-r)\pi_{+C}^{(1)})}] - q\tau\eta[B\frac{n_{\phi}^{(2B)}(r\pi_{-T}^{(2B)} + (1-r)\pi_{-C}^{(2B)})}{n_{-}^{(1)}(r\pi_{-T}^{(1)} + (1-r)\pi_{-C}^{(1)})}$$
$$+ B\frac{n_{\phi}^{(1)}(r\pi_{-T}^{(2)} + (1-r)\pi_{-C}^{(2)})}{n_{-}^{(1)}(r\pi_{-T}^{(1)} + (1-r)\pi_{-C}^{(1)})}]\}.$$

The covariance between standardized adjusted log rank statistics for true marker positive and true marker negative test at interim analysis is

$$\begin{aligned} Cov(Z_{+}^{(1)}, Z_{-}^{(1)}) &= -\left\{AB\left(\eta(1-\eta)(1-q)^2 Var(Q_{*}^{(1)}) + \tau(1-\tau)q^2 Var(Q_{\phi}^{(1)})\right)\right\}/\sigma_{+}^{(1)}\sigma_{-}^{(1)} \\ &= -\frac{ABG_1}{\sigma_{+}^{(1)}\sigma_{-}^{(1)}}. \end{aligned}$$

The covariance between standardized adjusted log rank statistics for true marker positive subgroup and overall population at interim analysis is

$$Cov(Z_{+}^{(1)}, Z^{(1)}) = C_1 + D_1 Cov(Z_{+}^{(1)}, Z_{-}^{(1)}) = C_1 - \frac{ABD_1G_1}{\sigma_{+}^{(1)}\sigma_{-}^{(1)}}.$$

The covariance between standardized adjusted log rank statistics for true marker negative subgroup and overall population at interim analysis is

$$Cov(Z_{-}^{(1)}, Z^{(1)}) = D_1 + C_1 Cov(Z_{+}^{(1)}, Z_{-}^{(1)}) = D_1 - \frac{ABC_1G_1}{\sigma_{+}^{(1)}\sigma_{-}^{(1)}}.$$

The covariance between standardized adjusted log rank statistics for true marker positive subgroup and true marker negative subgroup at final analysis time is

$$\begin{aligned} Cov(Z_{+}, Z_{-}) &= -AB\left\{\left(\eta(1-\eta)(1-q)^2 Var(Q_{*}) + \tau(1-\tau)q^2 Var(Q_{\phi})\right)\right\}/\sigma_{+}\sigma_{-} \\ &= -\frac{ABG_2}{\sigma_{+}\sigma_{-}}. \end{aligned}$$

The covariance between standardized adjusted log rank statistics for true marker positive subgroup and overall population at final analysis time is

$$Cov(Z_{+}, Z) = C_2 + D_2 Cov(Z_1, Z_0)$$

$$= C_2 - \frac{ABD_2G_2}{\sigma_+\sigma_-}.$$

The covariance between standardized adjusted log rank statistics for true marker negative subgroup and overall population at final analysis time is

$$\begin{aligned} Cov(Z_-, Z) &= D_2 + C_2 Cov(Z_+, Z_-) \\ &= D_2 - \frac{ABC_2G_2}{\sigma_+\sigma_-}. \end{aligned}$$

The covariance between standardized adjusted log rank statistics for true marker negative subgroup (interim) and true marker negative subgroup (final) is

$$Cov(Z_-, Z_-^{(1)}) = \sigma_-^{(1)}/\sigma_-.$$

The covariance between standardized adjusted log rank statistics for true marker positive subgroup (interim) and true marker positive subgroup (final) is

$$Cov(Z_+, Z_+^{(1)}) = \sigma_+^{(1)}/\sigma_+.$$

The covariance between standardized adjusted log rank statistics for true marker positive subgroup (interim) and overall population (final) is

$$\begin{aligned} Cov(Z, Z_+^{(1)}) &= C_2 Cov(Z_+, Z_+^{(1)}) + D_2 Cov(Z_-, Z_+^{(1)}) \\ &= C_2 \frac{\sigma_+^{(1)}}{\sigma_+} - \frac{ABD_2G_1}{\sigma_0\sigma_+^{(1)}}. \end{aligned}$$

The covariance between standardized adjusted log rank statistics for true marker negative subgroup (interim) and overall population (final) is

$$\begin{aligned} Cov(Z, Z_-^{(1)}) &= C_2 Cov(Z_+, Z_-^{(1)}) + D_2 Cov(Z_-, Z_-^{(1)}) \\ &= -\frac{ABC_2G_1}{\sigma_+\sigma_-^{(1)}} + D_2\sigma_-^{(1)}/\sigma_-. \end{aligned}$$

The covariance between standardized adjusted log rank statistics for true marker positive

subgroup (interim) and marker negative subgroup (final) is

$$\begin{aligned} Cov(Z_-, Z_+^{(1)}) &= -AB\left\{\eta(1-\eta)(1-q)^2 Var(Q_*^{(1)}) + q^2(1-\tau)\tau Var(Q_\phi^{(1)})\right\}/\sigma_-\sigma_1^{(1)} \\ &= \frac{-ABG_1}{\sigma_-\sigma_+^{(1)}}. \end{aligned}$$

The covariance between standardized adjusted log rank statistics for true marker negative subgroup (interim) and marker positive subgroup (final) is

$$\begin{aligned} Cov(Z_+, Z_-^{(1)}) &= -AB\left\{\eta(1-\eta)(1-q)^2 Var(Q_*^{(1)}) + q^2(1-\tau)\tau Var(Q_\phi^{(1)})\right\}/\sigma_+\sigma_0^{(1)} \\ &= \frac{-ABG_1}{\sigma_+\sigma_-^{(1)}}. \end{aligned}$$

The covariance between standardized adjusted log rank statistics for overall population (interim) and marker positive subgroup (final) is

$$\begin{aligned} Cov(Z_+, Z^{(1)}) &= C_1 Cov(Z_+, Z_+^{(1)}) + D_1 Cov(Z_+, Z_-^{(1)}) \\ &= C_1\frac{\sigma_+^{(1)}}{\sigma_+} - \frac{ABD_1G_1}{\sigma_1\sigma_-^{(1)}}. \end{aligned}$$

The covariance between standardized adjusted log rank statistics for overall population (interim) and marker negative subgroup (final) is

$$\begin{aligned} Cov(Z_-, Z^{(1)}) &= C_1 Cov(Z_-, Z_+^{(1)}) + D_1 Cov(Z_-, Z_-^{(1)}) \\ &= \frac{-ABC_1G_1}{\sigma_-\sigma_+^{(1)}} + D_1\frac{\sigma_0^{(1)}}{\sigma_-}. \end{aligned}$$

The covariance between standardized adjusted log rank statistics for overall population (interim) and overall population (final) is

$$\begin{aligned} Cov(Z, Z^{(1)}) &= Cov(\frac{pZ_+ + (1-p)Z_-}{\sigma}, \frac{pZ_+^{(1)} + (1-p)Z_-^{(1)}}{\sigma^{(1)}}) \\ &= C_1C_2\frac{\sigma_+^{(1)}}{\sigma_+} + D_1D_2\frac{\sigma_0^{(1)}}{\sigma_-} - ABG_1(\frac{C_1D_2}{\sigma_-\sigma_+^{(1)}} + \frac{C_2D_1}{\sigma_+\sigma_-^{(1)}}). \end{aligned}$$

The covariance between standardized adjusted log rank statistics for true marker positive

under IIB and true marker positive test statistics at interim analysis is

$$Cov(\overline{Z}_+, Z_+^{(1)}) = \frac{E_1}{\sigma_+^{(1)}\overline{\sigma}_+}.$$

The covariance between standardized adjusted log rank statistics for true marker positive under IIB and true marker negative test statistics at interim analysis is

$$Cov(\overline{Z}_+, Z_-^{(1)}) = \frac{E_2}{\sigma_-^{(1)}\overline{\sigma}_+}.$$

The covariance between standardized adjusted log rank statistics for true marker positive subgroup under IIB scenario and overall population at interim analysis is

$$\begin{aligned} Cov(\overline{Z}_+, Z^{(1)}) &= C_1 Cov(\overline{Z}_+, Z_+^{(1)}) + D_1 Cov(\overline{Z}_+, Z_-^{(1)}) \\ &= \frac{C_1 E_1}{\sigma_+^{(1)}\overline{\sigma}_1} + \frac{D_1 E_2}{\sigma_-^{(1)}\overline{\sigma}_1}. \end{aligned}$$

In summary, the correlation matrix for standardized adjusted log rank statistics is

$$Cov\left(\left(Z^{(1)}, Z_+^{(1)}, Z_-^{(1)}, Z, Z_+, \overline{Z}_+\right)^T\right) =$$

$$\begin{bmatrix} 1 & C_1 - \frac{ABD_1G_1}{\sigma_+^{(1)}\sigma_-^{(1)}} & D_1 - \frac{ABC_1G_1}{\sigma_+^{(1)}\sigma_-^{(1)}} & E & C_1\frac{\sigma_+^{(1)}}{\sigma_+} - \frac{ABD_1G_1}{\sigma_+\sigma_-^{(1)}} & \frac{C_1E_1}{\sigma_+^{(1)}\overline{\sigma}_+} + \frac{D_1E_2}{\sigma_-^{(1)}\overline{\sigma}_+} \\ & 1 & \frac{-ABG_1}{\sigma_+^{(1)}\sigma_-^{(1)}} & C_2\frac{\sigma_+^{(1)}}{\sigma_+} - \frac{ABD_2G_1}{\sigma_-\sigma_+^{(1)}} & \sigma_+^{(1)}/\sigma_+ & \frac{E_1}{\sigma_+^{(1)}\overline{\sigma}_+} \\ & & 1 & -\frac{ABC_2G_1}{\sigma_+\sigma_-^{(1)}} + D_2\sigma_-^{(1)}/\sigma_- & \frac{-ABG_1}{\sigma_+\sigma_-^{(1)}} & \frac{E_2}{\sigma_-^{(1)}\overline{\sigma}_+} \\ & & & 1 & C_2 - \frac{ABD_2G_2}{\sigma_+\sigma_-} & * \\ & & & & 1 & * \\ & & & & & 1 \end{bmatrix}$$

## 9.5 Asympototic distribution of test statistics under alternative

Given $t, N, p, \lambda_{sen}, \lambda_{spec}$, then $q, \tau, \eta$ are fixed. Under the alternative, given $\pi_+^{(1)}, \pi_+^{(2)}, \widetilde{\pi}_+$, $\pi_-^{(1)}, \pi_-^{(2)}, \widetilde{\pi}_-$, and $\pi_+^{(2b)}$, and the assumptions of proportional hazard $\log\theta_+ = \log\frac{\lambda_{+T}}{\lambda_{+C}}$, and $\log\theta_- = \log\frac{\lambda_{-T}}{\lambda_{-C}}$. Asymptotically, we have

$$Z_+^{(1)} \sim AN(\sqrt{tNr(1-r)}\frac{\pi_+^{(1)}\log\theta_+}{\sqrt{m_1}}, 1)\ , \tag{9.7}$$

$$Z_-^{(1)} \sim AN(\sqrt{tNr(1-r)}\frac{\pi_-^{(1)}\log\theta_-}{\sqrt{m_0}},1),$$

$$Z^{(1)} \sim AN(\frac{\sqrt{tNr(1-r)}\{p\pi_+^{(1)}\log\theta_+/\sqrt{m_1}+(1-p)\pi_-^{(1)}\log\theta_-/\sqrt{m_0}\}}{\sqrt{p^2+(1-p)^2-2p(1-p)m_2/\sqrt{m_1m_0}}},1)$$

$$Z_+ \sim AN(\sqrt{Nr(1-r)}\frac{t\pi_+^{(2)}+(1-t)\widetilde{\pi}_+}{\sqrt{m_3}}\log\theta_+,1)\ ,$$

$$Z_- \sim AN(\sqrt{Nr(1-r)}\frac{t\pi_-^{(2)}+(1-t)\widetilde{\pi}_-}{\sqrt{m_4}}\log\theta_-,1),$$

$$Z \sim AN(\frac{\sqrt{Nr(1-r)}\{p[t\pi_+^{(2)}+(1-t)\widetilde{\pi}_+]\frac{\log\theta_+}{\sqrt{m_3}}+(1-p)[t\pi_-^{(2)}+(1-t)\widetilde{\pi}_-]\frac{\log\theta_-}{\sqrt{m_4}}\}}{\sqrt{p^2+(1-p)^2-2p(1-p)m_5/\sqrt{m_3m_4}}},1),$$

$$\overline{Z}_+ \sim N(\sqrt{r(1-r)}\frac{tN\pi_+^{(2)}\log\theta_1+(1-t)N\pi_+^{(2B)}\log\theta_+/q}{A\sqrt{m_6}},1),$$

where

$$\begin{aligned}
m_1 &= \eta^2(1-q)^2\{\tau q\pi_+^{(1)}+(1-\tau)q\pi_-^{(1)}\}A^2+(1-\tau)^2q^2\{(1-\eta)(1-q)\pi_+^{(1)}\\
&\quad +\eta(1-q)\pi_-^{(1)}\}A^2,\\
m_0 &= (1-\eta)^2(1-q)^2\{\tau q\pi_+^{(1)}+(1-\tau)q\pi_-^{(1)}\}B^2+\tau^2q^2\{(1-\eta)(1-q)\pi_+^{(1)}\\
&\quad +\eta(1-q)\pi_-^{(1)}\}B^2,\\
m_2 &= AB\eta(1-\eta)(1-q)^2\{r(1-r)tN[\tau q\pi_+^{(1)}+AB(1-\tau)q\pi_-^{(1)}]\}\\
&\quad +\tau(1-\tau)q^2\{r(1-r)tN[(1-\eta)(1-q)\pi_+^{(1)}+\eta(1-q)\pi_-^{(1)}]\},\\
m_3 &= A^2\eta^2(1-q)^2\{t[\tau q\pi_+^{(2)}+(1-\tau)q\pi_-^{(2)}]+(1-t)[\tau q\widetilde{\pi}_{1t}+(1-\tau)q\widetilde{\pi}_-]\}\\
&\quad +A^2(1-\tau)^2q^2\{t[(1-\eta)(1-q)\pi_+^{(2)}+\eta(1-q)\pi_-^{(2)}]\\
&\quad +(1-t)[(1-\eta)(1-q)\widetilde{\pi}_{1t}+\eta(1-q)\widetilde{\pi}_-]\},\\
m_4 &= B^2(1-\eta)^2(1-q)^2\{t[\tau q\pi_+^{(2)}+(1-\tau)q\pi_-^{(2)}]+(1-t)[\tau q\widetilde{\pi}_++(1-\tau)q\widetilde{\pi}_-]\}\\
&\quad +B^2\tau^2q^2\{t[(1-\eta)(1-q)\pi_+^{(2)}+\eta(1-q)\pi_-^{(2)}]\\
&\quad +(1-t)[(1-\eta)(1-q)\widetilde{\pi}_++\eta(1-q)\widetilde{\pi}_-]\},\\
m_5 &= AB\eta(1-\eta)(1-q)^2r(1-r)\{tN[\tau q\pi_+^{(2)}+(1-\tau)q\pi_-^{(2)}]+(1-t)N[\tau q\widetilde{\pi}_+\\
&\quad +(1-\tau)q\widetilde{\pi}_-]\}\\
&\quad +AB\tau(1-\tau)q^2r(1-r)\{tN[(1-\eta)(1-q)\pi_+^{(2)}+\eta(1-q)\pi_-^{(2)}]\\
&\quad +(1-t)N[(1-\eta)(1-q)\widetilde{\pi}_++\eta(1-q)\widetilde{\pi}_-]\},
\end{aligned}$$

$$
\begin{aligned}
m_6 = \eta^2(1-q)^2\ & ([\tau q\pi_+^{(2B)} + (1-\tau)q\pi_-^{(2B)}](1-t)N/q + [\tau q\pi_+^{(2)} + (1-\tau)q\pi_-^{(2)}]tN) \\
& + (1-\tau)^2q^2\{A(\frac{N(1-t)(1-q)\pi_+^{(2B)}}{qn_+^{(1)}\pi_+^{(1)}} + \frac{n_\phi^{(1)}\pi_+^{(2)}}{n_+^{(1)}\pi_+^{(1)}})(1-\eta)\eta(1-q) \\
& - B(\frac{n_\phi^{(2B)}\pi_-^{(2B)}}{n_-^{(1)}\pi_-^{(1)}} + \frac{n_\phi^{(1)}\pi_-^{(2)}}{n_-^{(1)}\pi_-^{(1)}})(1-q)(1-\eta)\eta\}^2tN[\tau q\pi_+^{(1)} + (1-\tau)q\pi_+^{(0)}] \\
& + (1-\tau)^2q^2\{-(1-\eta)(1-\tau)qA(\frac{N(1-t)(1-q)\pi_+^{(2B)}}{qn_+^{(1)}\pi_+^{(1)}} + \frac{n_\phi^{(1)}\pi_+^{(2)}}{n_+^{(1)}\pi_+^{(1)}}) \\
& + \eta q\tau B(\frac{n_\phi^{(2B)}\pi_-^{(2B)}}{n_-^{(1)}\pi_-^{(1)}} + \frac{n_\phi^{(1)}\pi_-^{(2)}}{n_-^{(1)}\pi_-^{(1)}})\}^2tN[(1-\eta)(1-q)\pi_+^{(1)} + \eta(1-q)\pi_+^{(0)}] \\
& - \eta(1-q)(1-\tau)q\{A(\frac{N(1-t)(1-q)\pi_+^{(2B)}}{qn_+^{(1)}\pi_+^{(1)}} + \frac{n_\phi^{(1)}\pi_+^{(2)}}{n_+^{(1)}\pi_+^{(1)}})(1-\eta)\eta(1-q) \\
& - B(\frac{n_\phi^{(2B)}\pi_-^{(2B)}}{n_-^{(1)}\pi_-^{(1)}} + \frac{n_\phi^{(1)}\pi_-^{(2)}}{n_-^{(1)}\pi_-^{(1)}})(1-q)(1-\eta)\eta\}^2tN[\tau q\pi_+^{(1)} + (1-\tau)q\pi_+^{(0)}].
\end{aligned}
$$

## 9.6 Type I error $\alpha$ allocation and critical values

For our two-stage enrichment design, we split the overall alpha (e.g, $\alpha = 0.025$) between the two stages, following a similar method described by Lin et al. (2019). In Stage I, a fraction of the overall alpha, $\alpha_1$, is allocated to test the global hypothesis $H_0$, we have

$$
\begin{aligned}
\alpha_1 &= P(\text{Reject } H_0|H_0) \\
&= P_0(Z^{(1)} < -c_1 \text{ or } Z_+^{(1)} < -c_2) \\
&= P_0(Z^{(1)} < -c_1) + P_0(Z^{(1)} \geq -c_1 \text{ or } Z_+^{(1)} < -c_2) \\
&= \alpha_{1a} + \alpha_{1b}.
\end{aligned}
$$

The critical value $c_1$ is obtained by allocating a portion $\alpha_{1a}$ of $\alpha_1$ for testing $H_{0a}$, then $c_2$ can be solved for testing $H_{0+}$ in the above equation.

For Stage II, the overall alpha is left with $\alpha - \alpha_1$, to be used between the mutually exclusive Scenarios $II_A$ and $II_B$. We allocate $\alpha_2$, a fraction of $\alpha - \alpha_1$, for the tests in scenario $II_A$ and the rest $\alpha_{2*} = \alpha - \alpha_1 - \alpha_2$ for scenario $II_B$,

$$
\alpha - \alpha_1 = P(\text{Accept } H_0 \text{ at Stage } I, \text{ Reject } H_0 \text{ at Stage } II \text{ in case of } II_A)
$$

$$
\begin{aligned}
&+ P(\text{Accept } H_0 \text{ at Stage } I\text{, Reject } H_{0+} \text{ at Stage } II \text{ in case of } II_B\ ) \\
&= \alpha_2 + \alpha_{2*}.
\end{aligned}
$$

A futility criterion for marker-negative subjects will be used to determine $\alpha_2$ and $\alpha_{2*}$. This is done through a pre-specified threshold value $c_0$ for the test statistic $Z_-^{(1)}$ through futility probability $\mathcal{F}_p = P_0(Z_-^{(1)} \geq c_0)$. For example, if we want the futility probability to be 75% (50%), then from $P_0(Z_-^{(1)} \geq c_0) = 0.75\ (0.50)$ , $c_0 = -0.6745\ (0.0)$.

In case of scenario $II_A$, the test treatment passes the pre-defined futility threshold value $c_0$, $Z_-^{(1)} < c_0$. The study is continued with both marker-status cohorts and test $H_0$ at Stage II. The alpha is controlled by

$$
\begin{aligned}
\alpha_2 &= P(\text{Accept } H_0 \text{ at Stage I, Reject } H_0 \text{ at Stage } II \text{ in case of II}_A\ ) \\
&= P_0(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z < -b_1) \\
&+ P_0(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z \geq -b_1, Z_+ < -b_2) \\
&= \alpha_{2a} + (\alpha_2 - \alpha_{2a}).
\end{aligned}
$$

That is, the critical value $b_1$ is obtained by allocating a portion $\alpha_{2a}$ of $\alpha_2$, for testing $H_{0a}$. Then $b_2$ can be solved for testing $H_{0+}$ in the above equation.

In case of scenario $II_B$, the test treatment is futile on marker-negative group, i.e., $Z_-^{(1)} \geq c_0$. The study is continued with enriching marker-positive group. We test only $H_{0+}$ at final analysis time under scenario $II_B$. The alpha is controlled by

$$
\begin{aligned}
\alpha_{2*} &= P(\text{Accept } H_0 \text{ at Stage I, Reject } H_{0+} \text{ at Stage } II \text{ in case of } II_B\ ) \\
&= P_0(Z^{(1)} \geq -c_1\ ,\ Z_+^{(1)} \geq -c_2, Z_-^{(1)} \geq c_0, \overline{Z}_+ < -b_3).
\end{aligned}
$$

with the critical value $b_3$ solved numerically using the correlation matrix.

The strategy to allocating $\alpha$ to either $II_A$ or $II_B$ is an important design consideration (Lin et al., 2019). To utilize full $\alpha$ in both $II_A$ and $II_B$ scenarios, we split $\alpha - \alpha_1$ into $II_A$ and $II_B$ as follows. Since the trial will only be in one scenario or the other, it would be ideal to maximize the $\alpha$ in both scenarios. Toward this end, we can first rewrite $\alpha_2$ and

$\alpha_{2*}$, respectively, as

$$\begin{aligned}\alpha_2 = [&P_0(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z < -b_1)|Z_-^{(1)} < c_0) \\ &+ P_0(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z \geq -b_1, Z_+ < -b_2)|Z_-^{(1)} < c_0]P_0(Z_-^{(1)} < c_0)\end{aligned}$$

and

$$\alpha_{2*} = P_0(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, \overline{Z}_+ < -b_3|Z_-^{(1)} \geq c_0)P_0(Z_-^{(1)} \geq c_0).$$

Next, if we split $\alpha - \alpha_1$ into $II_A$ and $II_B$ with the same proportion as the odds of $P_0(Z_0^{(1)} < c_0)$ to $P_0(Z_-^{(1)} \geq c_0)$, ie,

$$\frac{\alpha_2}{\alpha_{2*}} = \frac{P_0(Z_-^{(1)} < c_0)}{P_0(Z_-^{(1)} \geq c_0)} = \frac{1 - \mathcal{F}_p}{\mathcal{F}_p}$$

then we have

$$\begin{aligned}\alpha - \alpha_1 = [&P_0(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z < -b_1)|Z_-^{(1)} < c_0) \\ &+ P_0(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z \geq -b_1, Z_+ < -b_2)|Z_-^{(1)} < c_0]\end{aligned}$$

and

$$\alpha - \alpha_1 = P_0(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, \overline{Z}_+ < -b_3|Z_-^{(1)} \geq c_0).$$

This indicates that with the above alpha allocation strategy, the corresponding conditional type I error is $\alpha - \alpha_1$ for either $II_A$ or $II_B$. When the odds of nonfutility versus futility $\frac{P_0(Z_-^{(1)} < c_0)}{P_0(Z_-^{(1)} \geq c_0)}$ is predetermined, the critical values can be calculated.

Tables 9.1 and 9.2 (information time is 0.3 and 0.5, respectively) show the critical values for some commonly used situations. When the prevalence rate $p$, $\lambda_{sen}$ , and $\lambda_{spec}$ are fixed, changing information time form 0.3 to 0.5 does not change the critical value for overall population ($c_1$) and for marker-positive population ($c_2$) at interim analysis, but change the critical values for overall population ($b_1$) and for marker-positive population ($b_2$) under $II_A$ and the critical values for marker-positive population ($b_3$) under $II_B$ at final analysis time.

Table 9.1: Critical Values When $\alpha = 0.025,\ \alpha_+ = 0.004, \alpha_2 = 0.021, r = 0.5, \mathcal{F}_p = 0.5$

| $\lambda_{sen}$ | $\lambda_{spec}$ | Info | $p$ | $c_1$ | $c_2$ | $b_1$ | $b_2$ | $b_3$ |
|---|---|---|---|---|---|---|---|---|
| 1.0 | 1.0 | 0.3 | 0.3 | -2.878 | -2.866 | -2.517 | -2.251 | -2.023 |
| | | | 0.4 | -2.878 | -2.848 | -2.507 | -2.216 | -2.021 |
| | | | 0.5 | -2.878 | -2.816 | -2.488 | -2.165 | -2.019 |
| 0.95 | 0.95 | 0.3 | 0.3 | -2.878 | -2.871 | -2.517 | -2.214 | -2.070 |
| | | | 0.4 | -2.878 | -2.857 | -2.505 | -2.183 | -2.047 |
| | | | 0.5 | -2.878 | -2.826 | -2.484 | -2.132 | -2.041 |
| 0.9 | 0.9 | 0.3 | 0.3 | -2.878 | -2.875 | -2.517 | -2.166 | -2.171 |
| | | | 0.4 | -2.878 | -2.864 | -2.504 | -2.141 | -2.097 |
| | | | 0.5 | -2.878 | -2.836 | -2.476 | -2.090 | -2.072 |
| 0.85 | 0.85 | 0.3 | 0.3 | -2.878 | -2.877 | -2.517 | -2.108 | -2.262 |
| | | | 0.4 | -2.878 | -2.870 | -2.502 | -2.088 | -2.165 |
| | | | 0.5 | -2.878 | -2.845 | -2.473 | -1.955 | -2.106 |
| 0.8 | 0.8 | 0.3 | 0.3 | -2.878 | -2.878 | -2.517 | -2.036 | -2.289 |
| | | | 0.4 | -2.878 | -2.874 | -2.500 | -2.022 | -2.232 |
| | | | 0.5 | -2.878 | -2.854 | -2.462 | -1.857 | -2.150 |
| 0.75 | 0.75 | 0.3 | 0.3 | -2.878 | -2.878 | -2.518 | -1.952 | -2.289 |
| | | | 0.4 | -2.878 | -2.877 | -2.498 | -1.944 | -2.275 |
| | | | 0.5 | -2.878 | -2.861 | -2.446 | -1.742 | -2.193 |
| 0.7 | 0.7 | 0.3 | 0.3 | -2.878 | -2.878 | -2.519 | -1.860 | -2.288 |
| | | | 0.4 | -2.878 | -2.878 | -2.498 | -1.857 | -2.284 |
| | | | 0.5 | -2.878 | -2.867 | -2.424 | -1.615 | -2.229 |
| 1 | 0.8 | 0.3 | 0.3 | -2.878 | -2.876 | -2.517 | -2.144 | -2.270 |
| | | | 0.4 | -2.878 | -2.865 | -2.503 | -2.131 | -2.156 |
| | | | 0.5 | -2.878 | -2.835 | -2.481 | -2.038 | -2.046 |
| 0.8 | 1 | 0.3 | 0.3 | -2.878 | -2.872 | -2.517 | -2.209 | -2.067 |
| | | | 0.4 | -2.878 | -2.860 | -2.505 | -2.167 | -1.910 |
| | | | 0.5 | -2.878 | -2.834 | -2.482 | -2.051 | -2.094 |

Table 9.2: Critical Values When $\alpha = 0.025,\ \alpha_+ = 0.004, \alpha_2 = 0.021, r = 0.5, \mathcal{F}_p = 0.5$

| $\lambda_{sen}$ | $\lambda_{spec}$ | Info | $p$ | $c_1$ | $c_2$ | $b_1$ | $b_2$ | $b_3$ |
|---|---|---|---|---|---|---|---|---|
| 1.0 | 1.0 | 0.5 | 0.3 | -2.878 | -2.866 | -2.509 | -2.234 | -2.016 |
| | | | 0.4 | -2.878 | -2.848 | -2.503 | -2.196 | -2.012 |
| | | | 0.5 | -2.878 | -2.816 | -2.489 | -2.142 | -2.009 |
| 0.95 | 0.95 | 0.5 | 0.3 | -2.878 | -2.871 | -2.510 | -2.183 | -2.056 |
| | | | 0.4 | -2.878 | -2.857 | -2.503 | -2.150 | -2.038 |
| | | | 0.5 | -2.878 | -2.826 | -2.485 | -2.095 | -2.034 |
| 0.9 | 0.9 | 0.5 | 0.3 | -2.878 | -2.875 | -2.511 | -2.118 | -2.142 |
| | | | 0.4 | -2.878 | -2.864 | -2.502 | -2.090 | -2.084 |
| | | | 0.5 | -2.878 | -2.836 | -2.481 | -2.033 | -2.066 |
| 0.85 | 0.85 | 0.5 | 0.3 | -2.878 | -2.877 | -2.511 | -2.035 | -2.232 |
| | | | 0.4 | -2.878 | -2.870 | -2.502 | -2.013 | -2.145 |
| | | | 0.5 | -2.878 | -2.845 | -2.473 | -1.955 | -2.106 |
| 0.8 | 0.8 | 0.5 | 0.3 | -2.878 | -2.878 | -2.512 | -1.932 | -2.274 |
| | | | 0.4 | -2.878 | -2.874 | -2.501 | -1.915 | -2.208 |
| | | | 0.5 | -2.878 | -2.854 | -2.462 | -1.857 | -2.150 |
| 0.75 | 0.75 | 0.5 | 0.3 | -2.878 | -2.878 | -2.512 | -1.811 | -2.279 |
| | | | 0.4 | -2.878 | -2.877 | -2.500 | -1.798 | -2.255 |
| | | | 0.5 | -2.878 | -2.861 | -2.446 | -1.742 | -2.193 |
| 0.7 | 0.7 | 0.5 | 0.3 | -2.878 | -2.878 | -2.512 | -1.670 | -2.280 |
| | | | 0.4 | -2.878 | -2.878 | -2.500 | -1.665 | -2.270 |
| | | | 0.5 | -2.878 | -2.867 | -2.424 | -1.615 | -2.229 |
| 1 | 0.8 | 0.5 | 0.3 | -2.878 | -2.876 | -2.511 | -2.087 | -2.235 |
| | | | 0.4 | -2.878 | -2.865 | -2.502 | -2.077 | -2.115 |
| | | | 0.5 | -2.878 | -2.835 | -2.481 | -2.038 | -2.046 |
| 0.8 | 1 | 0.5 | 0.3 | -2.878 | -2.872 | -2.510 | -2.176 | -2.071 |
| | | | 0.4 | -2.878 | -2.860 | -2.503 | -2.125 | -2.083 |
| | | | 0.5 | -2.878 | -2.834 | -2.482 | -2.051 | -2.094 |

## 9.7 Global and marginal power

The global power is

$$\begin{aligned} 1 - \beta &= P(\text{Reject } H_0 | H_1) \\ &= P(\text{Reject } H_{0a} \text{ or Reject } H_{0+} | H_1) \end{aligned}$$

$$
\begin{aligned}
&= P_1(Z^{(1)} < -c_1 \text{ or } Z_+^{(1)} < -c_2) \\
&\quad + P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z < -b_1) \\
&\quad + P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z \geq -b_1, Z_+ < -b_2) \\
&\quad + P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} \geq c_0, \overline{Z}_+ < -b_3) \\
&= p_1 + p_{2a} + p_{2+} + \overline{p}_{2+},
\end{aligned}
$$

where

$$
\begin{aligned}
p_1 &= P(\text{Reject } H_{0a} \text{ or Reject } H_{0+} \text{ at Stage } I | H_1) \\
&= P_1(Z^{(1)} < -c_1 \text{ or } Z_+^{(1)} < -c_2),
\end{aligned}
$$

$$
\begin{aligned}
p_{2a} &= P(\text{Accept } H_0 \text{ at Stage } I \text{ and Reject } H_{0a} \text{ at Stage } II_A \ | H_1) \\
&= P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z < -b_1),
\end{aligned}
$$

$$
\begin{aligned}
p_{2+} &= P(\text{Accept } H_0 \text{ at Stage } I, \text{ Accept } H_{0a} \text{ at Stage } II_A \text{ but Reject } H_{0+} \text{ at Stage } II_A \ | H_1) \\
&= P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z \geq -b_1, Z_+ < -b_2),
\end{aligned}
$$

$$
\begin{aligned}
\overline{p}_{2+} &= P(\text{Accept } H_0 \text{ at Stage } I \text{ and Reject } H_{0+} \text{ at Stage } II_B \ | H_1) \\
&= P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} \geq c_0, \overline{Z}_+ < -b_3),
\end{aligned}
$$

and $P_1(\cdot)$ is the probability under alternative.

Power for testing the treatment effect in the overall cohort is

$$
\begin{aligned}
1 - \beta_a &= P(\text{Reject } H_{0a} | H_1) \\
&= P(\text{Reject } H_{0a} \text{ at Stage } I) + \text{P(Reject } H_{0a} \text{ at Stage } II_A) \\
&= P_1(Z^{(1)} < -c_1) + p_{2a} \\
&= p_{1a} + p_{2a}.
\end{aligned}
$$

Power for testing the treatment effect in the marker-positive cohort is:

$$
\begin{aligned}
1-\beta_+ &= P(\text{Reject } H_{0+}|H_1) \\
&= P(\text{Reject } H_{0+} \text{ at Stage } I) + P(\text{Reject } H_{0+} \text{ at } II_A) + P(\text{Reject } H_{0+} \text{ at } II_B) \\
&= P_1(Z_+^{(1)} < -c_2) + P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} < c_0, Z_+ < -b_2) \\
&\quad + P_1(Z^{(1)} \geq -c_1 \ , \ Z_+^{(1)} \geq -c_2, Z_-^{(1)} \geq c_0, \overline{Z}_+ < -b_3).
\end{aligned}
$$

## 9.8 Sample size calculations

Similar to Section 6.9, given the global type I error and power, assuming that we have the estimated prevalence rate $p$, sensitivity $\lambda_{sen}$ and specificity $\lambda_{spec}$ from previous studies, the sample size need to detect the treatment effect can be found based on the formulas in Section 9.6 and Section 9.7 after we specify the design parameters in these sections.

To be more specific, given the design parameters shown in Section 9.6, we find the critical values first based on the formulas in Section 9.6. Notice that the critical values are based on the distributions under null hypothesis and do not depend on the sample size. Next using these critical values and the formulas shown in Section 9.7, we can determine the sample size needed to achieve the specified power of specific type (global, overall or marker positive), through optimization programming algorithms to find the solution and round up the the nearest integer. R-code is developed to calculate the needed sample size (the number of events) and results are illustrated in chapter 10.

# Chapter 10

# Numeric examples

## 10.1 Simulation set-up

Considera total number of subjects $N$=1000 for both Stage I and Stage II, with a prevalence rate $p$ (0.3, 0.4, or 0.5) for biomarker positive ($S+$). Randomization to active treatment $T$ or control treatment $C$ will be stratified by marker-appeared status with a randomization ratio of ($r$ to $1-r$, to $T$ or $C$, respectively). In Stage I, we plan to enroll $tN$ ($t = 0.7$) subjects (i.e., 7 months from the start of the study). A decision is made at interim analysis $T_1$ (information: $info = 0.5$), to continue enrolling biomarker-unselected subjects (under scenario $II_A$) or to enroll only biomarker-appeared positive subjects under scenario $II_B$, with a plan to enroll additional $(1-t)N$ subjects in Stage II. After the recruitment is complete, the study will follow-up to time $T_2$, when the final analysis is performed (the total number of events are observed, with pre-specified events from Stage I enrolled subjects and pre-specified events from Stage II enrolled subjects being observed). Assume subject recruitment follow a uniform distribution. Survival times are exponentially distributed.

The null hypothesis is the hazard rates for treatment and control group are equal. Exponential distributions are assumed to simulate the trials. The hazard rate for treated true marker positive $S+$ group is $\lambda_{+T}$, and hazard rate for control treated true marker positive $S+$ group is $\lambda_{+C}$. The hazard rates for true marker negative group $S-$ are $\lambda_{-T}$ and $\lambda_{-C}$.

We consider 9 different combinations of sensitivity and specificity ($\lambda_{sen} = \lambda_{spec} = 0.7$ to 1.0, $\lambda_{sen} = 1$ and $\lambda_{spec} = 0.8$, or $\lambda_{sen} = 0.8$ and $\lambda_{spec} = 1.0$), and simulate the trials for 3000 times for each combination of sensitivity and specificity.

The nominal and empirical global type I error are shown in Table 10.1. As we can see

from table 10.1, the empirical type I errors are close to the nominal type I error 0.025, across different prevalence rate (0.3, 0.4, 0.5), different sensitivity $\lambda_{sen}$ and specificity $\lambda_{spec}$ combinations from 1.0 to 0.7.

Table 10.1: The Nominal and Observed Type I Error for $H_0$ when $\lambda_{-T} = \lambda_{-C} = 1/10$ and $\lambda_{+T} = \lambda_{+C} = 1/15$, Info=0.5, N=1000, d=750, 3000 runs

| $\lambda_{sen}$ | $\lambda_{spec}$ | $p$ | Nominal global type I | Empirical global type I |
|---|---|---|---|---|
| 1.0 | 1.0 | 0.3 | 0.025 | 0.0173 |
| | | 0.4 | 0.025 | 0.0163 |
| | | 0.5 | 0.025 | 0.0210 |
| 0.95 | 0.95 | 0.3 | 0.025 | 0.0157 |
| | | 0.4 | 0.025 | 0.0167 |
| | | 0.5 | 0.025 | 0.0173 |
| 0.9 | 0.9 | 0.3 | 0.025 | 0.0200 |
| | | 0.4 | 0.025 | 0.0143 |
| | | 0.5 | 0.025 | 0.0143 |
| 0.85 | 0.85 | 0.3 | 0.025 | 0.0233 |
| | | 0.4 | 0.025 | 0.0187 |
| | | 0.5 | 0.025 | 0.0170 |
| 0.8 | 0.8 | 0.3 | 0.025 | 0.0203 |
| | | 0.4 | 0.025 | 0.0203 |
| | | 0.5 | 0.025 | 0.0177 |
| 0.75 | 0.75 | 0.3 | 0.025 | 0.0227 |
| | | 0.4 | 0.025 | 0.0220 |
| | | 0.5 | 0.025 | 0.0207 |
| 0.7 | 0.7 | 0.3 | 0.025 | 0.0240 |
| | | 0.4 | 0.025 | 0.0210 |
| | | 0.5 | 0.025 | 0.0190 |
| 1 | 0.8 | 0.3 | 0.025 | 0.0153 |
| | | 0.4 | 0.025 | 0.0160 |
| | | 0.5 | 0.025 | 0.0180 |
| 0.8 | 1 | 0.3 | 0.025 | 0.0203 |
| | | 0.4 | 0.025 | 0.0197 |
| | | 0.5 | 0.025 | 0.0170 |

## 10.2 Theoretical vs. empirical power under different scenarios

The theoretical and empirical global, overall, and marker positive subgroup powers are shown in Table 10.2. In this simulation, we use $\alpha = 0.025, \alpha_+ = 0.004$, $r = 0.5$ to calculate

the critical values. As we can see from Table 10.2, when there is treatment effect only in biomarker positive cohort and no treatment effect in biomarker negative cohort, the empirical powers (global power, overall power, and positive cohort power) are close to the corresponding theoretical powers, across different prevalence rate (0.3, 0.4, 0.5) under different combination of biomarker sensitivity $\lambda_{sen}$ and specificity $\lambda_{spec}$ (from 1.0 to 0.7). In the current set-up, since there is no treatment effect in biomarker negative cohort, the overall power is low.

Table 10.2: The Theoretical and Empirical Power when $\lambda_{-T} = \lambda_{-C} = \lambda_{+C} = 1/10$ and $\lambda_{+T} = 1/15, \mathcal{F}_p = 0.5$, Info = 0.5, N=1000, d=750, 3000 runs

| | | | Global Power | | Full cohort Power | | Positive Cohort Power | |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{sen}$ | $\lambda_{spec}$ | p | Theoretical | Empirical | Theoretical | Empirical | Theoretical | Empirical |
| 1.0 | 1.0 | 0.3 | 0.842 | 0.851 | 0.066 | 0.061 | 0.829 | 0.841 |
| | | 0.4 | 0.925 | 0.930 | 0.163 | 0.150 | 0.907 | 0.916 |
| | | 0.5 | 0.967 | 0.969 | 0.320 | 0.301 | 0.947 | 0.950 |
| 0.95 | 0.95 | 0.3 | 0.766 | 0.786 | 0.066 | 0.063 | 0.746 | 0.771 |
| | | 0.4 | 0.886 | 0.897 | 0.171 | 0.159 | 0.858 | 0.873 |
| | | 0.5 | 0.950 | 0.958 | 0.343 | 0.334 | 0.916 | 0.926 |
| 0.9 | 0.9 | 0.3 | 0.676 | 0.686 | 0.064 | 0.065 | 0.649 | 0.661 |
| | | 0.4 | 0.834 | 0.838 | 0.177 | 0.171 | 0.792 | 0.803 |
| | | 0.5 | 0.925 | 0.936 | 0.363 | 0.355 | 0.868 | 0.878 |
| 0.85 | 0.85 | 0.3 | 0.572 | 0.580 | 0.059 | 0.060 | 0.539 | 0.548 |
| | | 0.4 | 0.768 | 0.784 | 0.178 | 0.182 | 0.709 | 0.734 |
| | | 0.5 | 0.889 | 0.910 | 0.379 | 0.375 | 0.802 | 0.824 |
| 0.8 | 0.8 | 0.3 | 0.446 | 0.431 | 0.053 | 0.054 | 0.408 | 0.395 |
| | | 0.4 | 0.688 | 0.710 | 0.174 | 0.180 | 0.610 | 0.639 |
| | | 0.5 | 0.842 | 0.872 | 0.386 | 0.401 | 0.713 | 0.733 |
| 0.75 | 0.75 | 0.3 | 0.294 | 0.282 | 0.046 | 0.046 | 0.256 | 0.244 |
| | | 0.4 | 0.589 | 0.605 | 0.164 | 0.177 | 0.493 | 0.509 |
| | | 0.5 | 0.778 | 0.826 | 0.385 | 0.396 | 0.599 | 0.647 |
| 0.7 | 0.7 | 0.3 | 0.184 | 0.183 | 0.038 | 0.040 | 0.149 | 0.144 |
| | | 0.4 | 0.460 | 0.474 | 0.149 | 0.158 | 0.350 | 0.361 |
| | | 0.5 | 0.698 | 0.762 | 0.385 | 0.395 | 0.463 | 0.524 |
| 1 | 0.8 | 0.3 | 0.610 | 0.620 | 0.062 | 0.060 | 0.581 | 0.593 |
| | | 0.4 | 0.817 | 0.824 | 0.178 | 0.174 | 0.774 | 0.785 |
| | | 0.5 | 0.925 | 0.940 | 0.362 | 0.356 | 0.871 | 0.882 |
| 0.8 | 1 | 0.3 | 0.775 | 0.789 | 0.066 | 0.064 | 0.754 | 0.767 |
| | | 0.4 | 0.875 | 0.887 | 0.174 | 0.169 | 0.842 | 0.859 |
| | | 0.5 | 0.938 | 0.947 | 0.359 | 0.359 | 0.888 | 0.896 |

Figures 10.1, 10.2, and 10.3 show the contour plots of power surfaces for global (testing $H_1$), overall population (testing $H_{1a}$) and marker-positive population (testing $H_{1+}$) hypotheses, respectively, across $-0.10 \geq \delta \geq -0.40$ and $-0.10 \geq \delta_+ \geq -0.40$ by $n$, $p$ assuming $\alpha = 0.025, \alpha_1 = 0.004, w_+ = p, w_- = 1 - p, \delta = p\delta_+ + (1-p)\delta_-$, $r = 0.5$. The power increases as $n$, $p$, $\lambda_{sen}$, or $\lambda_{spec}$ increases as well. For example, with $p = 0.5$ and $n = 500$, $\delta = -0.15$, and $\delta_+ = -0.4$, Figure 10.1.A with $\lambda_{sen} = \lambda_{spec} = 1$ shows global power of 75%; however, Figure 10.1.D shows global power only about 50%. We also see the $\lambda_{sen}$ has less impact than $\lambda_{spec}$ on the global power. In addition, from Figure 10.1, the global power increases with increasing treatment effects for overall population and marker positive population (decreasing $\delta$ and/or decreasing $\delta_+$).

From Figure 10.2, the power for overall population increases with increasing treatment effect for overall population when treatment effect for positive population is fixed (decreasing $\delta$).

From Figure 10.3, the power for positive subgroup increases with increasing treatment effect for positive population (decreasing $\delta_+$) but decreases with increasing treatment effect for overall population (decreasing $\delta$).

Tables 10.3 and 10.4 show the sample size needed to achieve specified global and marginal power for marker positive subgroup, respectively. For illustration, assume the same $\alpha$ allocation, $\lambda_{sen} = \lambda_{spec} = 0.8$, and treatment effect $\delta = -0.15$ (the log-hazard ratio for the overall cohort), and $\delta_+ = -0.4$ (the log-hazard ratio for the marker positive cohort). Assuming a target of 90% global power for testing $H_1$ when prevalence is 0.4 and interim analysis is performed at information time of 0.5, a total sample size of 1837 is needed as shown in Table 10.3. Assuming the target of 80% marginal power for testing treatment effect on biomarker positive cohort (testing $H_{1+}$) when prevalence is 0.4 and interim analysis is performed at information time of 0.5, a total sample size of 1652 is needed as shown in Table 10.4.

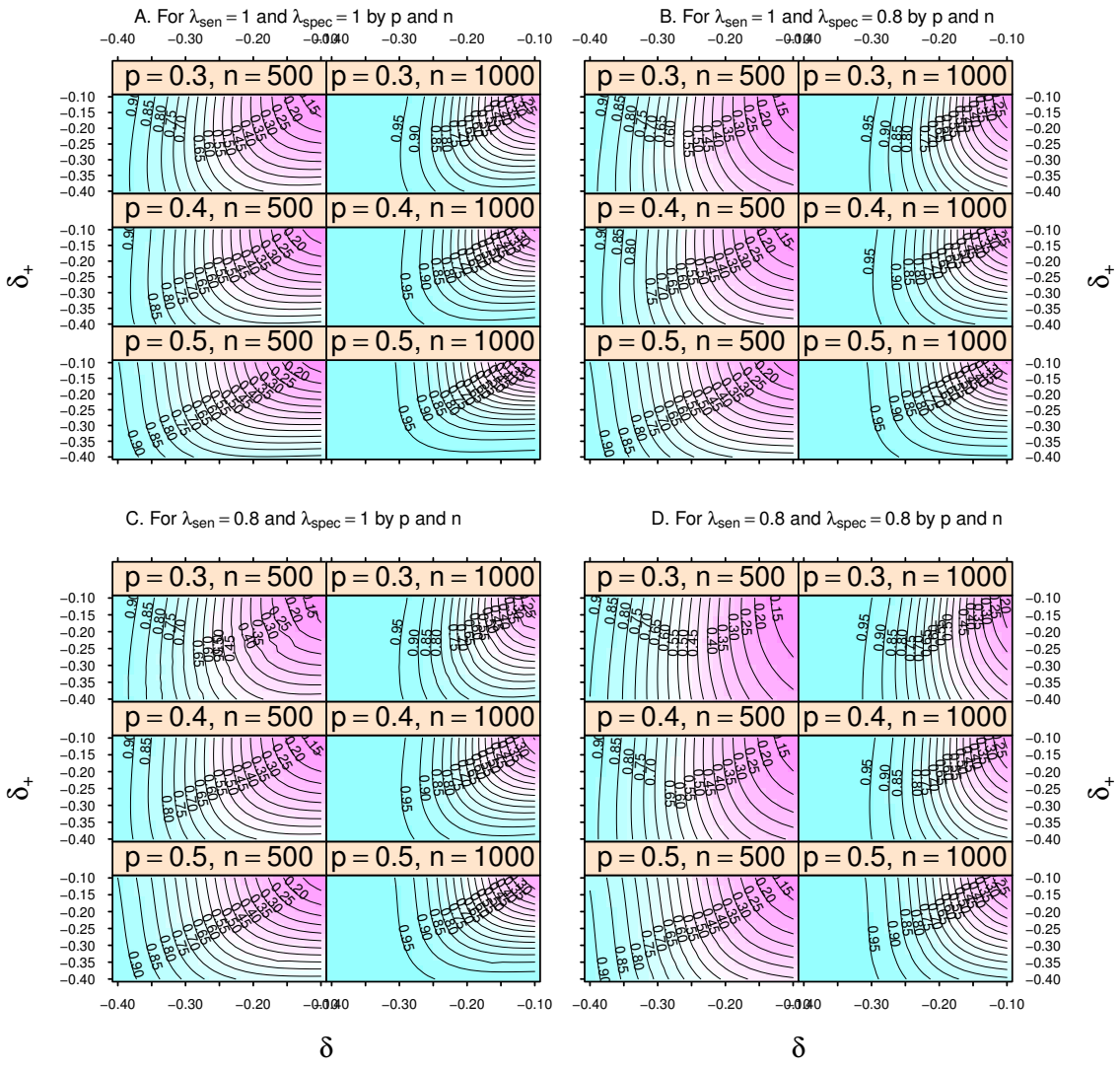



Figure 10.1: Contour plot of global power surface

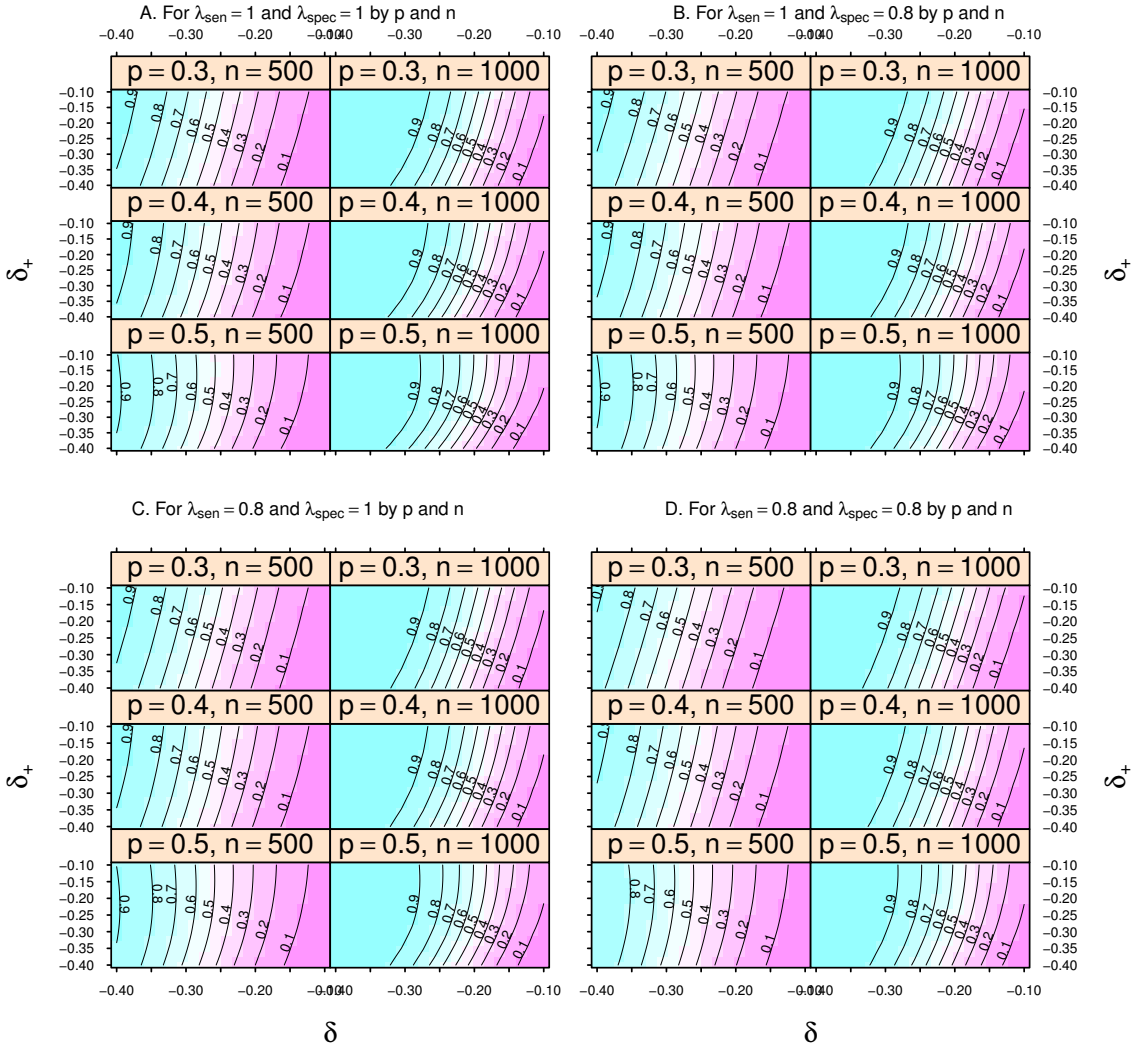



Figure 10.2: Contour plot of power surface for overall cohort effect

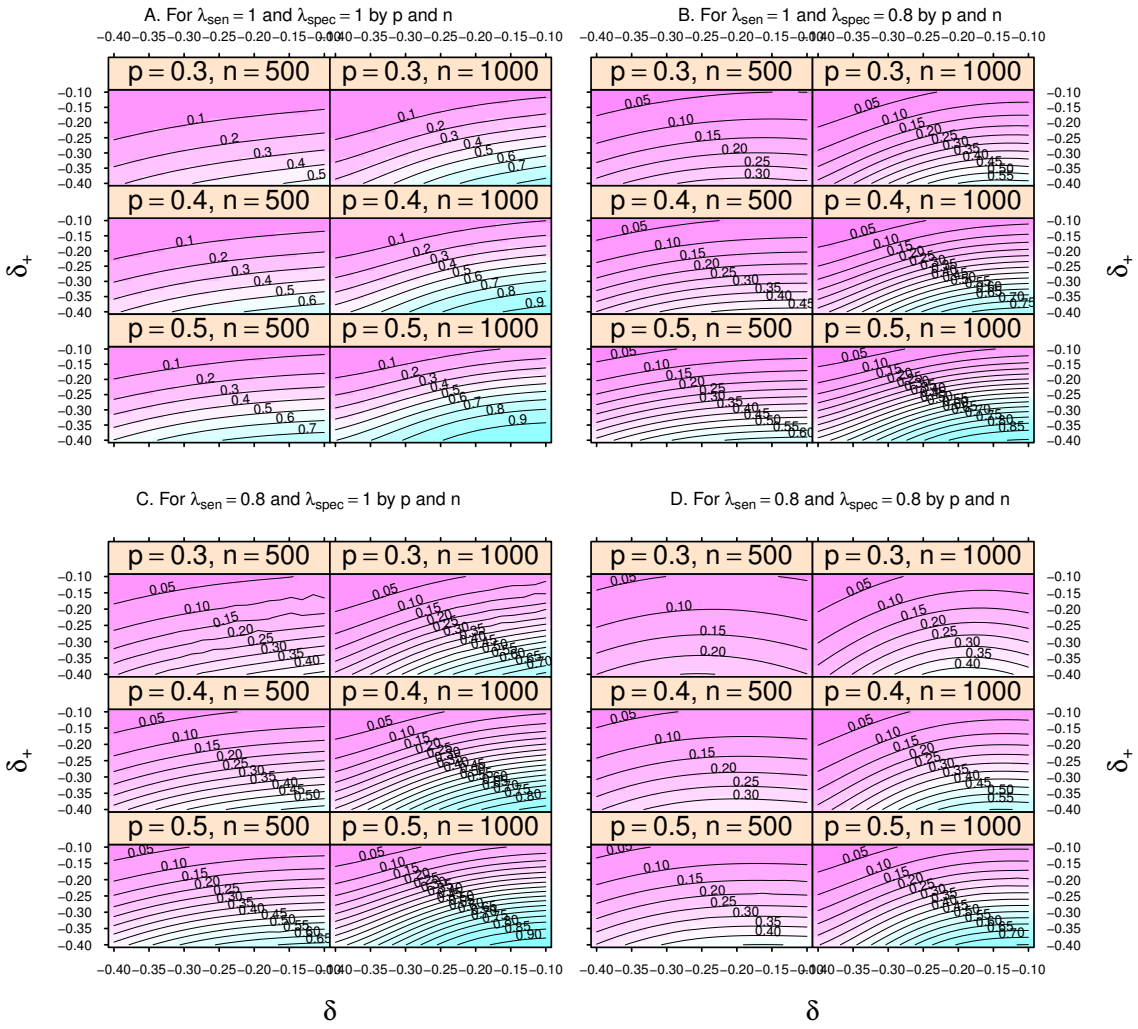



Figure 10.3: Contour plot of power surface for marker positive cohort effect

Table 10.3: Total Sample Size to Achieve Specified Global Power $H_1$ when $\alpha = 0.025, \alpha_+ = 0.004, \mathcal{F}_p = 0.5$, $r = 0.5$, and $\delta = p \log \theta_+ + (1-p) \log \theta_- = -0.15, \delta_+ = \log \theta_+ = -0.4$.

| | | | 90% power | | 80% power | |
|---|---|---|---|---|---|---|
| $p$ | $\lambda_{sen}$ | $\lambda_{spec}$ | Info=0.3 | Info=0.5 | Info=0.3 | Info=0.5 |
| 0.3 | 1 | 1 | 1259 | 1261 | 936 | 935 |
| | 0.95 | 0.95 | 1496 | 1492 | 1116 | 1111 |
| | 0.9 | 0.9 | 1786 | 1768 | 1337 | 1321 |
| | 0.85 | 0.85 | 2165 | 2108 | 1632 | 1582 |
| | 0.8 | 0.8 | 2754 | 2565 | 2112 | 1947 |
| | 0.75 | 0.75 | 4055 | 3375 | 3147 | 2623 |
| | 0.7 | 0.7 | 7750 | 5329 | 5566 | 4117 |
| | 1 | 0.8 | 1984 | 1925 | 1500 | 1448 |
| | 0.8 | 1 | 1498 | 1496 | 1105 | 1102 |
| 0.4 | 1 | 1 | 943 | 931 | 708 | 698 |
| | 0.95 | 0.95 | 1092 | 1078 | 819 | 809 |
| | 0.9 | 0.9 | 1283 | 1266 | 963 | 949 |
| | 0.85 | 0.85 | 1537 | 1510 | 1155 | 1133 |
| | 0.8 | 0.8 | 1891 | 1837 | 1424 | 1382 |
| | 0.75 | 0.75 | 2446 | 2316 | 1853 | 1749 |
| | 0.7 | 0.7 | 3604 | 3185 | 2738 | 2421 |
| | 1 | 0.8 | 1349 | 1327 | 1015 | 997 |
| | 0.8 | 1 | 1132 | 1116 | 845 | 833 |
| 0.5 | 1 | 1 | 754 | 741 | 568 | 558 |
| | 0.95 | 0.95 | 852 | 841 | 643 | 633 |
| | 0.9 | 0.9 | 983 | 972 | 741 | 731 |
| | 0.85 | 0.85 | 1161 | 1147 | 874 | 863 |
| | 0.8 | 0.8 | 1409 | 1391 | 1060 | 1046 |
| | 0.75 | 0.75 | 1771 | 1747 | 1332 | 1312 |
| | 0.7 | 0.7 | 2342 | 2299 | 1761 | 1724 |
| | 1 | 0.8 | 988 | 973 | 745 | 732 |
| | 0.8 | 1 | 912 | 899 | 687 | 677 |

Table 10.4: Total Sample Size to Achieve Specified Marginal Power $H_{1+}$ when $\alpha = 0.025, \alpha_+ = 0.004, \mathcal{F}_p = 0.5$, $r = 0.5$, and $\delta = p \log \theta_+ + (1-p) \log \theta_- = -0.15, \delta_+ = \log \theta_+ = -0.4$.

| | | | 90% power | | 80% power | |
|---|---|---|---|---|---|---|
| $p$ | $\lambda_{sen}$ | $\lambda_{spec}$ | Info=0.3 | Info=0.5 | Info=0.3 | Info=0.5 |
| 0.3 | 1 | 1 | 1397 | 1422 | 1012 | 1024 |
| | 0.95 | 0.95 | 1736 | 1793 | 1242 | 1262 |
| | 0.9 | 0.9 | 2217 | 2362 | 1550 | 1581 |
| | 0.85 | 0.85 | 2976 | 3568 | 1992 | 2032 |
| | 0.8 | 0.8 | 14595 | 8513 | 2733 | 2742 |
| | 0.75 | 0.75 | 24613 | 14404 | 4343 | 4193 |
| | 0.7 | 0.7 | 40364 | 23643 | 30360 | 17725 |
| | 1 | 0.8 | 2517 | 2709 | 1760 | 1768 |
| | 0.8 | 1 | 1766 | 1836 | 1244 | 1267 |
| 0.4 | 1 | 1 | 985 | 975 | 733 | 726 |
| | 0.95 | 0.95 | 1162 | 1159 | 859 | 854 |
| | 0.9 | 0.9 | 1408 | 1414 | 1031 | 1030 |
| | 0.85 | 0.85 | 1769 | 483 | 1274 | 1276 |
| | 0.8 | 0.8 | 2363 | 2539 | 1646 | 1652 |
| | 0.75 | 0.75 | 11109 | 6586 | 2291 | 2315 |
| | 0.7 | 0.7 | 22606 | 13249 | 3790 | 4188 |
| | 1 | 0.8 | 1488 | 1494 | 1090 | 1086 |
| | 0.8 | 1 | 1223 | 1222 | 896 | 892 |
| 0.5 | 1 | 1 | 768 | 755 | 577 | 567 |
| | 0.95 | 0.95 | 878 | 866 | 658 | 649 |
| | 0.9 | 0.9 | 1030 | 1021 | 767 | 760 |
| | 0.85 | 0.85 | 1250 | 1251 | 921 | 919 |
| | 0.8 | 0.8 | 1598 | 1644 | 1153 | 1163 |
| | 0.75 | 0.75 | 2271 | 2667 | 1540 | 1596 |
| | 0.7 | 0.7 | 13912 | 8231 | 2319 | 2801 |
| | 1 | 0.8 | 1033 | 1021 | 769 | 760 |
| | 0.8 | 1 | 950 | 939 | 708 | 700 |

# Chapter 11

# Keytruda trial examples

## 11.1 Misclassification of predictive biomarkers

Immunotherapy is a new paradigm for the treatment of non-small-cell cancer (NSCLC), and targeting the PD-1/PD-L1 pathway is a promising therapeutic option. Pembrolizumab is a new immunotherapy that blocks the PD-1 pathway and restores the body's immune response against cancer cells and allows the immune system to recognize and kill cancer cells. We use the KEYNOTE-10 trial as an example to illustrate our method. This was a (phase 2/3) randomized trial to study Pembrolizumab versus Docetaxel for previously treated, PD-L1-positive, advance NSCLC patients (Herbst et al., 2016). This trial stratified qualified subjects by biomarker's TPS (tumor proportion score $\geq$50% vs 1-49% ), which measured the extent of PD-L1 expression, then randomized subjects with 1:1:1 ratio to three treatment groups within each high and low TPS stratum. The companion diagnostic assay for PD-L1 expression was the Dako EnVision FLEX+HRP-Polymer kit using the 22C3 antibody clone, which was validated in the phase 1 KEYNOTE-001 trial (Garon et al., 2015). Here, we use the phase 1 KEYNOTE-001 data as the basis to "redesign" the KEYNOTE-10 as an "imaginary" two-stage enrichment trial to illustrate our method. Since there was no significant difference between the two test doses of Pembrolizumab, we only look at the Pembrolizumab 2 mg (Pem) versus Docetaxel (Dox), which is the control/standard-of-care.

From the phase 1 KEYNOTE-001 trial, the prevalence rate was about 0.39 for TPS <1%, 0.38 for TPS = 1-49%, and 0.23 for TPS $\geq$50%; $\lambda_{sen} = \lambda_{spec} = 0.8$. Thus, we estimate the prevalence rate of the PD-L1 true "strongly positive" (TPS $\geq$ 50%) among the PD-L1 positive (TPS > 1%) NSCLC patients being $p \approx 0.40$; the appeared PD-L1 "strongly positive" prevalence is $q = P(M = *) = p\lambda_{sen} + (1-p)(1-\lambda_{spec}) = 0.40 \times 0.80 + (1 - 0.40)(1-0.80) = 0.44$. Hence $PPV = \tau = \frac{p\lambda_{sen}}{p\lambda_{sen}+(1-p)(1-\lambda_{spec})} = \frac{p\lambda_{sen}}{q} = \frac{0.4\times0.8}{0.44} = 0.73$, and

$NPV = \eta = \frac{(1-p)\lambda_{spec}}{1-P(M=*)} = \frac{(1-p)\lambda_{spec}}{1-q} = \frac{(1-0.4)\times 0.8}{1-0.44} = 0.86$. The real phase 2/3 KEYNOTE-10 trial had both overall survival and progression-free survival (PFS) as primary end-points, We only use PFS for the "imaginary" trial for illustration purpose. Suppose that the overall (one-sided) $\alpha = 0.025$ and an interim analysis planned at $info = 0.5$, with $\alpha_1 = 0.004$ allocated for Stage I. As stated in the work of Herbst et al. (2016), the study aimed to show a benefit of Pem over Dox in PFS in patients with TPS $\geq$ 50% as well as in the whole TPS $\geq$ 1% cohort, so we allocate $\alpha_{1a} = \alpha_{1b} = \frac{\alpha_1}{2} = 0.002$. Moreover, assume that we set the futility probability for the PD-L1 "weakly positive" (TPS 1-49%) subset to be 50% ( implying $\mathcal{F}_p = 0.5$ and $c_0 = 0$). Then $\alpha_2 = 0.0105$ for Stage $II_A$, and an equal amount of 0.0105 for Stage $II_B$. In the case of $II_A$, we further allocate 0.00525 for testing the overall cohort and equally the test 0.00525 for the PD-L1 strongly positive subset.

From the reports by Herbst et al. (2016), the total enrollment time is 19 months for 688 subjects. We expect a total of 688 $\times$ 0.92= 632 PFS events at the end of the trial. In Stage I, we plan to enroll 70% subjects (13.3 months from the start of the study). A decision is made at interim analysis $T_1$ (information: $info = 0.5$, when $316 = 632 \times 0.5$ PFS is expected), to continue enrolling biomarker-unselected subjects (under scenario $II_A$) or to enroll only biomarker positive subjects under scenario $II_B$, with a plan to enroll additional 30% subjects in Stage II. After the recruitment is complete, the study will follow-up to calendar time $T_2$, when the final analysis is performed (under scenario $II_A$, a total number of 632 PFS are observed, with 461 PFS events from Stage I enrolled subjects and 171 PFS events from Stage II enrolled subjects; under scenario $II_B$, a total number of 133 PFS events from Stage II enrolled marker-appeared positive subjects). The number of PFS events from Stage I and Stage II enrolled subjects were determined, to make sure the number of events arrives approximately at the same calendar time.

To illustrate power calculation, we take the treatment effect information from Herbst et al. (2016). Assume PFS times are exponentially distributed. The hazard rate for treated $S+$ group is $\lambda_{+T} = 9.90$, and hazard rate for control treated $S+$ group is $\lambda_{+C} = 5.85$. The hazard rates for marker negative group $S-$ are $\lambda_{-T} = 5.14$ and $\lambda_{-C} = 5.85$. These design parameters lead to the critical values $(c_1, c_2, b_1, b_2, b_3) = (-2.878, -2.848, -2.503, -2.197, -2.015)$ when $\lambda_{sen} = \lambda_{spec} = 1$; and critical values $(c_1, c_2, b_1, b_2, b_3) =$

$(-2.878, -2.874, -2.501, -1.918, -2.215)$ when $\lambda_{sen} = \lambda_{spec} = 0.8$. With a total of $N = 688$ patients and prevalence rate $p = 0.4$, we expected the power to test global hypothesis $H_1$ to be 99%, the power to test $H_{1+}$ to be 99%, and the power to test $H_{1a}$ to be 3.5%, when $\lambda_{sen} = \lambda_{spec} = 1$. However, when $\lambda_{sen} = \lambda_{spec} = 0.8$, with a total of $N = 688$ patients and prevalence rate $p = 0.4$, we expected the power to test global hypothesis $H_1$ to be 82%, the power to test $H_{1+}$ is 81%, and the power to test $H_{1a}$ to be 4.4%.

To illustrate sample size calculation, with the same $\alpha$ allocation, $\lambda_{sen} = \lambda_{spec} = 0.8$, and treatment effect $\delta =$ the log-hazard ratio (Dox vs Pem)$= -0.128$ for the overall cohort, and $\delta_+ =$ the log-hazard ratio (Dox vs Pem)$= -0.528$ for the marker positive cohort. Trials usually aim a power of either 80% or 90% for the marker positive subset. Assuming a target of 90% power for testing $H_{1+}$, a total sample size of 893 is needed. The global power for the composite hypothesis $H_1$ is 91%, and the power for the overall cohort $H_{1a}$ is 5%. If the target of 80% power is for testing $H_{1+}$, a total sample size of 670 is needed. The global power for the composite hypothesis $H_1$ is 81%, and the power for the overall cohort $H_{1a}$ is 4.4%.

# Chapter 12

## Summary

In part I, a two-stage adaptive enrichment clinical trial with survival outcome is designed, based on a binary predictive biomarker. We assume that the classification of the binary biomarker in part I is perfect.

In part II, misclassification of the binary predictive biomarker is considered. We use the information obtained from both marker appeared-positive strata and marker appeared-negative strata to solve the adjusted log rank statistics for true marker positive and true marker negative group. No additional distributional assumption is needed for the group sequential designs we used in this part.

In part III, misclassification adjustment is extended to a two-stage enrichment designs. In the final analysis time $T_2$, with additional distributional assumption (exponential distribution assumption for survival times), we can use the information obtained from interim analysis time $T_1$, to help obtain the adjusted log rank statistics for the true marker positive group, even though the marker-appeared negative group was discontinued after the interim analysis time $T_1$ and no marker-appeared negative subjects are enrolled in Stage II.

In all three parts above, family-wise type I error rate is controlled by using the correlations between the log rank statistics within and between the stages. R-code is developed to calculate critical values, achieved global power, and marginal powers, and to calculate the sample size needed to achieve specified global power and marginal powers as well.

# Chapter 13

# Appendix — lost to follow up

## 13.1 Simulation results for two-stage stratified design with lost to follow up

### 13.1.1 Simulation set-up

Consider a total number of subjects $N$=1000 for both Stage I and Stage II, with a prevalence rate $p$ (0.3, 0.4, or 0.5) for biomarker positive ($S+$). subjects enrollment is expected to complete in 10 months. Randomization to active treatment $T$ or placebo treatment $C$ will be stratified by marker-appeared status with a randomization of $T$ and $C$ with ratio of $r$ to $1-r$. After the recruitment is complete, the study will follow-up to interim analysis time $T_1$ and final analysis time $T_2$. Assume subject recruitment follow a uniform distribution. Survival times are exponentially distributed. **The time of lost to follow up is exponentially distributed with a rate** $\lambda_{censor} = 0.00427 = 1/234$**.**

The null hypothesis is the hazard rates for treatment and placebo group are equal. Exponential distributions are assumed to simulate the trials. The hazard rate for treated true marker positive $S+$ group is $\lambda_{+T}$, and hazard rate for placebo treated true marker positive $S+$ group is $\lambda_{+C}$. The hazard rates for true marker negative group $S-$ are $\lambda_{-T}$ and $\lambda_{-C}$.

We consider 9 different combinations of sensitivity and specificity ($\lambda_{sen} = \lambda_{spec} = 0.7$ to 1.0, $\lambda_{sen} = 1$ and $\lambda_{spec} = 0.8$, or $\lambda_{sen} = 0.8$ and $\lambda_{spec} = 1.0$), and simulate the trials for 3000 times for each combination of sensitivity and specificity.

The nominal and empirical global type I error rates are shown in Table 13.1. As we can see from Table 13.1, the empirical type I error rates are close to the nominal type I error rate 0.025, across different prevalence rate (0.3, 0.4, 0.5), different sensitivity $\lambda_{sen}$ and

specificity $\lambda_{spec}$ combinations from 1.0 to 0.7.

Table 13.1: The Nominal and Empirical Global Type I Error Ratefor $H_0$ when $\lambda_{-T} = \lambda_{-C} = 1/10$ and $\lambda_{+T} = \lambda_{+C} = 1/15$, Info=0.5, N=1000, d=750, 3000 runs

| | | | Global type I | |
|---|---|---|---|---|
| $\lambda_{sen}$ | $\lambda_{spec}$ | $p$ | Nominal | Empirical |
| 1.0 | 1.0 | 0.3 | 0.025 | 0.0227 |
| | | 0.4 | 0.025 | 0.0223 |
| | | 0.5 | 0.025 | 0.0210 |
| 0.95 | 0.95 | 0.3 | 0.025 | 0.0207 |
| | | 0.4 | 0.025 | 0.0217 |
| | | 0.5 | 0.025 | 0.0267 |
| 0.9 | 0.9 | 0.3 | 0.025 | 0.0193 |
| | | 0.4 | 0.025 | 0.0223 |
| | | 0.5 | 0.025 | 0.0233 |
| 0.85 | 0.85 | 0.3 | 0.025 | 0.0250 |
| | | 0.4 | 0.025 | 0.0200 |
| | | 0.5 | 0.025 | 0.0227 |
| 0.8 | 0.8 | 0.3 | 0.025 | 0.0210 |
| | | 0.4 | 0.025 | 0.0233 |
| | | 0.5 | 0.025 | 0.0217 |
| 0.75 | 0.75 | 0.3 | 0.025 | 0.0240 |
| | | 0.4 | 0.025 | 0.0193 |
| | | 0.5 | 0.025 | 0.0237 |
| 0.7 | 0.7 | 0.3 | 0.025 | 0.0297 |
| | | 0.4 | 0.025 | 0.0203 |
| | | 0.5 | 0.025 | 0.0187 |
| 1 | 0.8 | 0.3 | 0.025 | 0.0197 |
| | | 0.4 | 0.025 | 0.0247 |
| | | 0.5 | 0.025 | 0.0220 |
| 0.8 | 1 | 0.3 | 0.025 | 0.0237 |
| | | 0.4 | 0.025 | 0.0217 |
| | | 0.5 | 0.025 | 0.0207 |

### 13.1.2 Theoretical vs. empirical power under different scenarios

In this simulation, the critical values are based on $\alpha = 0.025, \alpha_1 = 0.004$, $r = 0.5$. The theoretical and empirical global, overall, and positive cohort powers are shown in Table 13.2. As we can see from Table 13.2, when there is treatment effect only in biomarker positive cohort and no treatment effect in biomarker negative cohort, the empirical powers (global power, overall power, and marker positive cohort power) are close to the corresponding theoretical powers, across different prevalence rate (0.3, 0.4, 0.5) under different combinations of biomarker sensitivity $\lambda_{sen}$ and specificity $\lambda_{spec}$ (from 1.0 to 0.7). In the current set-up, since there is no treatment effect in biomarker negative cohort, the overall power is low.

Table 13.2: The Theoretical and Empirical Power when $\lambda_{-T} = \lambda_{-C} = \lambda_{+C} = 1/10$ and $\lambda_{+T} = 1/15, \lambda_{censor} = 1/234, \mathcal{F}_p = 0.5$ , N=1000, d=750, 3000 runs

| | | | Global Power | | Full cohort Power | | Positive Cohort Power | |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{sen}$ | $\lambda_{spec}$ | p | Theoretical | Empirical | Theoretical | Empirical | Theoretical | Empirical |
| 1.0 | 1.0 | 0.3 | 0.773 | 0.767 | 0.099 | 0.103 | 0.755 | 0.747 |
| | | 0.4 | 0.895 | 0.897 | 0.233 | 0.230 | 0.875 | 0.878 |
| | | 0.5 | 0.958 | 0.958 | 0.445 | 0.458 | 0.936 | 0.943 |
| 0.95 | 0.95 | 0.3 | 0.690 | 0.693 | 0.100 | 0.109 | 0.661 | 0.665 |
| | | 0.4 | 0.850 | 0.854 | 0.247 | 0.249 | 0.815 | 0.829 |
| | | 0.5 | 0.938 | 0.932 | 0.481 | 0.496 | 0.898 | 0.903 |
| 0.9 | 0.9 | 0.3 | 0.596 | 0.601 | 0.097 | 0.103 | 0.554 | 0.555 |
| | | 0.4 | 0.792 | 0.786 | 0.258 | 0.267 | 0.733 | 0.739 |
| | | 0.5 | 0.910 | 0.916 | 0.518 | 0.540 | 0.840 | 0.848 |
| 0.85 | 0.85 | 0.3 | 0.497 | 0.508 | 0.092 | 0.097 | 0.442 | 0.456 |
| | | 0.4 | 0.720 | 0.712 | 0.263 | 0.275 | 0.629 | 0.630 |
| | | 0.5 | 0.875 | 0.877 | 0.553 | 0.579 | 0.756 | 0.763 |
| 0.8 | 0.8 | 0.3 | 0.397 | 0.404 | 0.084 | 0.085 | 0.332 | 0.338 |
| | | 0.4 | 0.637 | 0.636 | 0.262 | 0.282 | 0.507 | 0.500 |
| | | 0.5 | 0.833 | 0.835 | 0.584 | 0.605 | 0.640 | 0.653 |
| 0.75 | 0.75 | 0.3 | 0.301 | 0.308 | 0.073 | 0.080 | 0.234 | 0.236 |
| | | 0.4 | 0.546 | 0.551 | 0.254 | 0.274 | 0.377 | 0.375 |
| | | 0.5 | 0.787 | 0.796 | 0.609 | 0.640 | 0.496 | 0.496 |
| 0.7 | 0.7 | 0.3 | 0.213 | 0.225 | 0.073 | 0.064 | 0.153 | 0.162 |
| | | 0.4 | 0.448 | 0.455 | 0.236 | 0.266 | 0.254 | 0.246 |
| | | 0.5 | 0.744 | 0.755 | 0.627 | 0.666 | 0.340 | 0.343 |
| 1 | 0.8 | 0.3 | 0.559 | 0.577 | 0.096 | 0.100 | 0.512 | 0.534 |
| | | 0.4 | 0.782 | 0.787 | 0.259 | 0.263 | 0.719 | 0.737 |
| | | 0.5 | 0.915 | 0.913 | 0.515 | 0.536 | 0.847 | 0.853 |
| 0.8 | 1 | 0.3 | 0.675 | 0.667 | 0.099 | 0.110 | 0.644 | 0.631 |
| | | 0.4 | 0.824 | 0.814 | 0.252 | 0.263 | 0.778 | 0.774 |
| | | 0.5 | 0.917 | 0.918 | 0.509 | 0.534 | 0.856 | 0.853 |

## 13.2 Simulation results for two-stage enrichment design with lost to follow up

### 13.2.1 Simulation set-up

Consider a total number of subjects $N$=1000 for both Stage I and Stage II, with a prevalence rate $p$ (0.3, 0.4, or 0.5) for biomarker positive ($S+$). Randomization to active treatment $T$ or placebo treatment $C$ will be stratified by marker-appeared status with a randomization ratio of $r$ to $1-r$, to $T$ or $C$, respectively. In Stage I, we plan to enroll $tN$ ($t = 0.7$) subjects (i.e., 7 months from the start of the study). A decision is made at interim analysis $t_1$ (information: $info = 0.5$), to continue enrolling biomarker-unselected subjects under scenario $II_A$ or to enroll only biomarker positive subjects under scenario $II_B$, with a plan to enroll additional $(1-t)N$ subjects in Stage II. After the recruitment is complete, the study will follow-up to time $T_2$, when the final analysis is performed. The total number of events are observed, with per-specified events from Stage I enrolled subjects and per-specified events from Stage II enrolled subjects being observed. Assume subject recruitment follow a uniform distribution. Survival times are exponentially distributed. **The time of lost to follow up is exponentially distributed with a rate $\lambda_{censor} = 0.00427 = 1/234$.**

The null hypothesis is the hazard rates for treatment and placebo group are equal. Exponential distributions are assumed to simulate the trials. The hazard rate for treated true marker positive $S+$ group is $\lambda_{+T}$, and the hazard rate for placebo treated in true marker positive $S+$ group is $\lambda_{+C}$. The hazard rates for true marker negative group $S-$ are $\lambda_{-T}$ and $\lambda_{-C}$.

We consider 9 different combinations of sensitivity and specificity ($\lambda_{sen} = \lambda_{spec} = 0.7$ to 1.0, $\lambda_{sen} = 1$ and $\lambda_{spec} = 0.8$, or $\lambda_{sen} = 0.8$ and $\lambda_{spec} = 1.0$), and simulate the trials for 3000 times for each combination of sensitivity and specificity. The nominal and empirical global type I error rates are shown in Table 13.3. As we can see from Table 13.3, the empirical type I error rates are close to the nominal type I error rate 0.025, across different prevalence rate (0.3, 0.4, 0.5), different sensitivity $\lambda_{sen}$ and specificity $\lambda_{spec}$ combinations from 1.0 to 0.7.

Table 13.3: The Nominal and Empirical Global Type I Error Rate for $H_0$ when $\lambda_{-T} = \lambda_{-C} = 1/10$ and $\lambda_{+T} = \lambda_{+C} = 1/10, \mathcal{F}_p = 0.5$, Info=0.3, N=1000, d=750, 3000 runs

| | | | Global type I | |
|---|---|---|---|---|
| $\lambda_{sen}$ | $\lambda_{spec}$ | $p$ | Nominal | Empirical |
| 1.0 | 1.0 | 0.3 | 0.025 | 0.0280 |
| | | 0.4 | 0.025 | 0.0223 |
| | | 0.5 | 0.025 | 0.0280 |
| 0.95 | 0.95 | 0.3 | 0.025 | 0.0297 |
| | | 0.4 | 0.025 | 0.0277 |
| | | 0.5 | 0.025 | 0.0293 |
| 0.9 | 0.9 | 0.3 | 0.025 | 0.0260 |
| | | 0.4 | 0.025 | 0.0243 |
| | | 0.5 | 0.025 | 0.0287 |
| 0.85 | 0.85 | 0.3 | 0.025 | 0.0263 |
| | | 0.4 | 0.025 | 0.0247 |
| | | 0.5 | 0.025 | 0.0290 |
| 0.8 | 0.8 | 0.3 | 0.025 | 0.0267 |
| | | 0.4 | 0.025 | 0.0183 |
| | | 0.5 | 0.025 | 0.0233 |
| 0.75 | 0.75 | 0.3 | 0.025 | 0.0250 |
| | | 0.4 | 0.025 | 0.0240 |
| | | 0.5 | 0.025 | 0.0280 |
| 0.7 | 0.7 | 0.3 | 0.025 | 0.0203 |
| | | 0.4 | 0.025 | 0.0267 |
| | | 0.5 | 0.025 | 0.0233 |
| 1 | 0.8 | 0.3 | 0.025 | 0.0290 |
| | | 0.4 | 0.025 | 0.0243 |
| | | 0.5 | 0.025 | 0.0280 |
| 0.8 | 1 | 0.3 | 0.025 | 0.0327 |
| | | 0.4 | 0.025 | 0.0257 |
| | | 0.5 | 0.025 | 0.0277 |

### 13.2.2 Theoretical vs. empirical power under different scenarios

In this simulation, we use $\alpha = 0.025, \alpha_1 = 0.004$, $r = 0.5$ to calculate the critical values. The theoretical and empirical global, overall, and positive subgroup powers are shown in Table 13.4. As we can see from Table 13.4, when there is treatment effect only in biomarker positive cohort and no treatment effect in biomarker negative cohort, the empirical powers (global power, overall power, and positive cohort power) are close to the corresponding theoretical powers, across different prevalence rate (0.3, 0.4, 0.5) under different combinations of biomarker sensitivity $\lambda_{sen}$ and specificity $\lambda_{spec}$ (from 1.0 to 0.7). In the current set-up, since there is no treatment effect in biomarker negative cohort, the overall power is low.

Table 13.4: The Theoretical and Empirical Power when $\lambda_{-T} = \lambda_{-C} = \lambda_{+C} = 1/10$ and $\lambda_{+T} = 1/15, \lambda_{censor} = 1/234, \mathcal{F}_p = 0.5$ , N=1000, d=750, 3000 runs

| | | | Global Power | | Full cohort Power | | Positive Cohort Power | |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{sen}$ | $\lambda_{spec}$ | p | Theoretical | Empirical | Theoretical | Empirical | Theoretical | Empirical |
| 1.0 | 1.0 | 0.3 | 0.847 | 0.835 | 0.070 | 0.066 | 0.834 | 0.823 |
| | | 0.4 | 0.925 | 0.929 | 0.163 | 0.152 | 0.908 | 0.914 |
| | | 0.5 | 0.968 | 0.972 | 0.320 | 0.371 | 0.947 | 0.949 |
| 0.95 | 0.95 | 0.3 | 0.766 | 0.773 | 0.066 | 0.070 | 0.747 | 0.754 |
| | | 0.4 | 0.886 | 0.899 | 0.172 | 0.174 | 0.859 | 0.874 |
| | | 0.5 | 0.950 | 0.957 | 0.343 | 0.338 | 0.916 | 0.923 |
| 0.9 | 0.9 | 0.3 | 0.678 | 0.702 | 0.064 | 0.067 | 0.650 | 0.675 |
| | | 0.4 | 0.835 | 0.854 | 0.177 | 0.183 | 0.793 | 0.813 |
| | | 0.5 | 0.925 | 0.938 | 0.364 | 0.369 | 0.869 | 0.882 |
| 0.85 | 0.85 | 0.3 | 0.574 | 0.592 | 0.060 | 0.058 | 0.540 | 0.558 |
| | | 0.4 | 0.769 | 0.802 | 0.179 | 0.193 | 0.711 | 0.743 |
| | | 0.5 | 0.890 | 0.916 | 0.380 | 0.386 | 0.803 | 0.830 |
| 0.8 | 0.8 | 0.3 | 0.447 | 0.449 | 0.054 | 0.055 | 0.409 | 0.413 |
| | | 0.4 | 0.689 | 0.727 | 0.175 | 0.198 | 0.612 | 0.645 |
| | | 0.5 | 0.842 | 0.872 | 0.387 | 0.401 | 0.714 | 0.740 |
| 0.75 | 0.75 | 0.3 | 0.295 | 0.303 | 0.046 | 0.049 | 0.256 | 0.263 |
| | | 0.4 | 0.591 | 0.635 | 0.165 | 0.195 | 0.494 | 0.523 |
| | | 0.5 | 0.779 | 0.832 | 0.386 | 0.399 | 0.601 | 0.646 |
| 0.7 | 0.7 | 0.3 | 0.185 | 0.187 | 0.038 | 0.045 | 0.149 | 0.146 |
| | | 0.4 | 0.461 | 0.478 | 0.149 | 0.169 | 0.351 | 0.351 |
| | | 0.5 | 0.699 | 0.775 | 0.386 | 0.411 | 0.464 | 0.531 |
| 1 | 0.8 | 0.3 | 0.612 | 0.643 | 0.062 | 0.060 | 0.583 | 0.613 |
| | | 0.4 | 0.818 | 0.844 | 0.178 | 0.182 | 0.775 | 0.801 |
| | | 0.5 | 0.925 | 0.936 | 0.362 | 0.363 | 0.871 | 0.887 |
| 0.8 | 1 | 0.3 | 0.775 | 0.771 | 0.066 | 0.068 | 0.775 | 0.752 |
| | | 0.4 | 0.876 | 0.882 | 0.175 | 0.181 | 0.842 | 0.846 |
| | | 0.5 | 0.938 | 0.941 | 0.359 | 0.365 | 0.888 | 0.894 |

# Bibliography

S-J Wang, R.T. O'Neill, and HMJ. Hung. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical statistics*, 6:227–244, 2007.

Y. Lin, W. Shih, and S-E Lu. Two-stage enrichment clinical trial design with adjustment for misclassification in predictive biomarkers. *Statistics in Medicine*, pages 5445–5469, 2019.

W.J. Shih and Y. Lin. On study designs and hypotheses for clinical trials with predictive biomarkers. *Contemporary Clinical Trials*, (62):140–145, 2017.

L.A. Renfro, H. Mallick, M-W An, D.J. Sargent, and S.J. Mandrekar. Clinical trial designs incorporating predictive biomarkers. *Cancer Treatment Reviews*, 43:74–82, 2016.

W. Brannath, E. Zuber, M. Brabson, F. Bretz, P. Gallo, M. Posch, and A. Racine-Poon. Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine*, 28:1445–1463, 2009.

M. Jenkins, A. Stonr, and C. Jennison. An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics*, (10):347–356, 2001.

T. Friede, N. Parson, and N. Stallard. A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine*, 31:4309–4320, 2012.

B Spiessens and M Debois. Adjusted significance levels for subgroup analysis in clinical trials. *Contemporary Clinical Trials*, pages 647–656, 2010.

C. Mehta, H. Schafer, H. Daniel, and S. Irle. Biomarker driven population enrichment

for adaptive oncology trials with time to event endpoints. *Statistics in Medicine*, 33: 4515–4531, 2014.

EB Garon, NA Rizvi, R. Hui, and et al. Pembrolizumab for the treatment of non-small-cell lung cancer. *New England Journal of Medicine*, pages 2018–2028, 2015.

RS Herbst, P. Baas, DW Kim, and et al. Pembrolizumab versus doctaxel for previously treated, pd-l1-positive, advanced non-small-cell lung cancer (keynote-010): a randomised controlled trial. *The Lancet*, pages 1540–1550, 2016.

FlemingHarrington. Counting process and survival analysis. *Wiley-Interscience publication*, page 270, 1991.

D. Schoenfeld. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68:316–319, 1981.

D. Schoenfeld. Sample-size formula for the proportional-hazards regression model. *Biometrika*, pages 499–503, 1983.