# Adaptive designs for confirmatory clinical trials in developing molecular targeted therapies and biosimilars

Doctoral dissertation
March 2018

Tokyo University of Science
4415701 Ryuji Uozumi

# Contents

**Preface** **iv**

**Dedication** **vi**

**Acknowledgments** **vii**

**1 Introduction** **1**

1.1 Clinical trial development . . . 1

1.2 Statistical inference for survival endpoints . . . 4

1.2.1 Summarized statistics . . . 4

1.2.2 Log-rank tests . . . 4

1.2.3 Cox proportional hazards model . . . 6

1.2.4 Survival endpoints . . . 7

1.3 Statistical inference for different types of comparisons . . . 8

1.4 Adaptive designs for confirmatory clinical trials . . . 9

**2 Utility-based interim decision rule planning in adaptive designs for population selection** **11**

2.1 Introduction . . . 11

2.2 Setting of adaptive population selection design with two stages using survival endpoints . . . 14

2.3 Types of measures used in the interim decision . . . 15

2.3.1 Simple hazard ratio . . . 15

2.3.2 Predictive power . . . 16

2.3.3 Restricted mean survival time . . . 17

2.4 Formulation of an interim decision rule based on utility functions . . . 18

2.5 Simulation study . . . 20

2.5.1 Simulation setting . . . 20

2.5.2 Simulation results . . . 21

2.6 Summary and discussion . . . 27

3 Interim decision-making strategies in adaptive designs for population selection 29
3.1 Introduction . . . . 29
3.2 Interim decision-making . . . . 31
3.2.1 Procedures for interim decision-making . . . . 31
3.2.2 Simple exponential model for dependence between OS and PFS . . . . 31
3.2.3 The illness-death model for semi-competing risks data . . . . 32
3.2.4 Interim decision rule using predictive power . . . . 33
3.3 Simulation study . . . . 35
3.3.1 Design assumptions and simulation setting . . . . 35
3.3.2 Simulation results . . . . 37
3.4 Summary and discussion . . . . 43

4 State of the art: Adaptive seamless design in developing biosimilars 44
4.1 Introduction . . . . 44
4.2 Motivating example . . . . 48
4.3 Literature review . . . . 49
4.4 Proposed framework . . . . 49
4.4.1 Collecting both PK and efficacy data from first stage . . . . 50
4.4.2 Sample size adjustment for efficacy endpoint of interim analysis . . . . 50
4.4.3 PK equivalence including early stopping for efficacy . . . . 52
4.5 Simulation study . . . . 54
4.5.1 Simulation setting . . . . 54
4.5.2 Simulation results . . . . 55
4.6 Summary and discussion . . . . 59

5 Discussion 60
5.1 Applicability of adaptive population selection designs in oncology . . . . 60
5.2 Multiplicity issues in adaptive population selection designs . . . . 60
5.3 The proportional hazards assumption in molecular targeted therapies . . . . 61
5.4 Applicability of adaptive seamless design in developing biosimilars . . . . 62

6 Conclusion 64

Bibliography 66

Appendices 77

A Derivation of the correlation model between OS and PFS by Fleischer et al. (2009) in Chapter 3 78
A.1 Correlation between OS and PFS with maximal independence . . . . 78
A.2 A more general model considering dependencies between OS and TTP . . . . 79

**B Details of simulation results in Chapter 3** **82**

# Preface

All stakeholders associated with clinical drug developments are longing for breakthrough therapy designation. To satisfy this requirement, clinical trials should be conducted efficiently and biostatisticians need to provide innovative clinical trial designs.

In recent years, adaptive clinical trial designs have been attractive to pharmaceutical sponsors, regulatory agencies, and medical investigators. Adaptive designs in early phase clinical trials got to be applied in real clinical trials in the last ten years owing to their exploratory aspects. In contrast, adaptive designs in late phase clinical trials were rarely applied in real clinical trials. In comparison to the early phase clinical trials, in our opinion, more pragmatic methodologies have been expected because the late phase clinical trials correspond to confirmatory ones and need to satisfy the agreement between pharmaceutical sponsors and regulatory agencies.

In consideration for the current trend in clinical drug developments, we focus on adaptive designs in developing molecular targeted therapies and biosimilars. Although there has been a proliferation of research articles on the adaptive clinical trial designs, little research that is more pragmatic has been done. The aim of this dissertation is to investigate three focal issues of the adaptive designs in confirmatory clinical trials as follows.

### Issue 1: Utility-based interim decision rule planning in adaptive designs for population selection

The use of adaptive population selection designs has spread in response to the emergence of numerous molecular targeted therapies. Such a design provides an opportunity to stop recruitment for a population in midcourse when this population does not benefit from the treatment being tested. However, there are no well-established procedures to setting the thresholds in an interim decision when applying a design to a clinical trial. We propose a novel utility-based approach to guide the construction of the interim decision rule of an adaptive population selection design for the setting of the survival endpoint.

### Issue 2: Interim decision-making strategies in adaptive designs for population selection

As an extension of Issue 1, we consider an interim analysis using overall survival (OS), progression-free survival (PFS), and both OS and PFS, to determine whether the whole population or only the biomarker-positive population should continue into the subsequent stage of the trial, whereas the final decision is made based on OS data only. In order to increase the probability of selecting the most appropriate population at the interim analysis, we propose an interim decision-making strategy in adaptive designs with correlated endpoints considering the post-progression survival (PPS) magnitudes. In our approach, the interim decision is made on the basis of predictive power by incorporating information on OS as well as PFS to supplement the incomplete OS data.

**Issue 3: State of the art: Adaptive seamless design in developing biosimilars**

Recently, numerous pharmaceutical sponsors have expressed a great deal of interest in the development of biosimilars, which requires clinical trials to demonstrate the equivalence of pharmacokinetics (PK) and clinical efficacy. Pharmacodynamics (PD) may be used in evaluating efficacy if there are relevant PD markers available. However, in their absence, it is necessary to design the associated clinical trials to include efficacy measures as the primary endpoint. Hence, we propose a novel adaptive seamless PK and efficacy design with an efficient framework to remedy the risk of misspecification of efficacy parameters and to discontinue the trial evaluating the efficacy for futility based on the PK evaluation. Here, we consider the clinical development of biosimilars including their evaluation in patients rather than healthy volunteers under a situation where both PK and efficacy parameters are required to demonstrate the equivalence. The original idea of the proposed method was to organize a clinical trial that includes the statistical analysis of PK as an interim analysis, with sample size recalculation of the efficacy data.

This doctoral dissertation consists of five chapters. Chapter 1 presents background such as clinical trial development and an outline of the adaptive designs in confirmatory clinical trials. In Chapter 2, we propose a novel utility-based approach to guide the construction of the interim decision rule of an adaptive population selection design for the setting of the survival endpoint (Uozumi and Hamada, submitted). Chapter 3 provides the interim decision-making strategies in adaptive designs with correlated survival endpoints considering the post-progression survival (PPS) magnitudes (Uozumi and Hamada, 2017b). In Chapter 4, we propose an adaptive seamless PK and efficacy design in developing biosimilars (Uozumi and Hamada, 2017a). Finally, Chapter 5 discusses the issues associated with this work, and Chapter 6 provides a conclusion. We wish to accelerate the use of adaptive designs in real clinical trials through this work.

# Dedication

To my parents, Mr. Makoto Uozumi and Mrs. Kazue Uozumi, with my gratitude and love.

# Acknowledgments

First and foremost, I wish to express my deepest gratitude to my supervisor, Professor Chikuma Hamada[†], who taught me the pleasure of studying biostatistics, who inspired me to work as a biostatistician, and who always guided me in the right direction. He had supported me since I was a undergraduate school student. I admire his passion for statistics and his immense knowledge. I will always remember his tolerance and strictness throughout my study. I could not have completed this dissertation without his support and trust. He is the best mentor ever. His memory will remain with me forever. May he rest in peace.

I would like to appreciate Associate Professor Takashi Sozu, who gave valuable advice every time at the seminar. I am also grateful to my fellow graduate student, Mr. Shinjo Yada, who had instructive discussions with me.

I would like to thank Professor Tohru Ikeguchi, Professor Toshihiro Furukawa, Professor Yukinobu Taniguchi, and Professor Takako Akakura, who were my doctoral committee members, for their valuable suggestions and helpful comments that improved the content and presentation of this dissertation. I am grateful to Associate Professor Masataka Taguri of the Yokohama City University for taking time from his busy schedule to serve as my external examiner.

I wish to appreciate Professor Satoshi Morita, a head of Department of Data Science, Institute for Advancement of Clinical and Translational Science, Kyoto University Hospital, who gave me a chance to work in academia, who encouraged me to pursue a doctoral degree, and who provided a research environment and countless opportunities. I wish to thank Professor Atsushi Kawaguchi of the Saga University, who consistently supported and encouraged me throughout research activities. I am grateful to Assistant Professor Hitoshi Koyano of the Tokyo Institute of Technology, who provided very helpful suggestions on the material in this dissertation.

Last but not least, I would like to express my gratitude to my family and friends. With their love, belief, and support, I never felt mentally alone during the entire PhD work.

March 2018

[†] Deceased 21 December 2017.

# Chapter 1

# Introduction

## 1.1 Clinical trial development

Clinical trials are experiments on human subjects to demonstrate the efficacy and safety of new therapies. Typically, clinical trials mainly consist of three distinct phases. If an experimental therapy is considered to be effective through phases I, II, and III, it will be approved by the regulatory agency for use on the pharmaceutical market. Among the clinical trial phases, phase I clinical trials are conducted with healthy volunteers to evaluate the safety of the experimental therapy whereas phase II clinical trials are conducted with patients to evaluate the efficacy. Phases I and II are also called early phase clinical trials. Finally, phase III clinical trials are called confirmatory clinical trials, and generally large randomized, controlled trials, with the endpoint being a direct measure of patient benefit. To conduct the phase III clinical trials, pharmaceutical sponsors have to pay extravagant resources owing to the large number of patients required. According to the report by Martin et al. (2017), the median costs of conducting a study from protocol approval to final clinical trial report were 3.4 million US dollars for phase I trials involving patients, 8.6 million US dollars for phase II trials and 21.4 million US dollars for phase III trials, in the data set collected from 7 major companies on 726 interventional studies conducted in patients from 2010 to 2015. Nevertheless, the proportion of successful phase III clinical trials is the lowest among all phases (DiMasi and Grabowski, 2007). As a more recent report, the proportion of failure in phase III clinical trials is relatively higher in oncology among all therapeutic areas (Arrowsmith and Miller, 2013; Harrison, 2016). This work focuses on the issues in the confirmatory clinical trials in response to expectation for the improvement.

Recently, numerous pharmaceutical sponsors have expressed a great deal of interest in the development of molecular targeted therapies. Molecular targeted therapies investigated particularly in oncology are beneficial only in a subgroup of the overall population. The European Medicines Agency (EMA) has published the regulatory guidance on the investigation of subgroups in confirmatory clinical trials (European Medicines Agency, 2014b). To identify the subgroup, a biomarker is driven to indicate the biological state. For example, the patients with HER2 amplification are approximately 15 to 20% of patients with breast cancer and can benefit from the administration of paclitaxel after adjuvant chemotherapy with doxorubicin plus cyclophosphamide (Hayes et al., 2007). Table 1.1 shows the selected molecular targeted therapies approved in Japan in the last ten years. If a pre-defined biomarker hypothesis from exploratory studies exists, clinical trial designs should be set up considering the heterogeneity of patient subgroups by using the biomarker at the planning stage, as the U. S. Food and Drug Administration (FDA) has provided the enrichment strategies in developing molecular targeted therapies (Food and Drug Administration, 2012).

Table 1.1: Selected molecular targeted therapies approved in Japan

| Drug | Type of cancer | Biomarker | Approved year |
|---|---|---|---|
| Cetuximab | Colorectal cancer | K-ras | 2008 |
| Gefitinib | Non-small-cell lung cancer | EGFR | 2009 |
| Crizotinib | Non-small-cell lung cancer | ALK | 2012 |
| Pertuzumab | Breast cancer | HER2 | 2013 |
| Ipilimumab | Melanoma | CTLA4 | 2015 |
| Regorafenib | Hepatocellular carcinoma | VEGFR | 2017 |

Similarly, developing biosimilars has recently become an important issue for pharmaceutical sponsors and regulatory agencies owing to the anticipation of the impending expiration of a number of patents for biological medicinal products in numerous countries. Table 1.2 cited from Rémuzat et al. (2017) shows the list of best-selling biologics with patent expiry in the years to come. According to the guidelines, a biosimilar is defined as a biological medicinal product that contains an active substance that is similar to that of an original previously authorized biological medicinal product (European Medicines Agency, 2014a; Food and Drug Administration, 2015). Biological medicinal products are large and complex molecules that include vaccines, gene therapies, and cellular therapies. Thus, most are important life-saving products but are extremely expensive, which makes it difficult to reach the general patient population. A reduction in healthcare costs for patients can be expected if a biosimilar is approved by regulators and placed on the market. For example, biosimilar trastuzumab is expected to be priced at a level at which patients who otherwise would not have access to expensive therapies such as trastuzumab could receive needed therapy (Bauchner et al., 2017).

Therefore, this work mainly focuses on two issues in confirmatory clinical trials: the developments of molecular targeted therapies and biosimilars. Since we handle survival data in oncology in the former development, statistical inference for survival endpoints is described in Section 1.2. In Section 1.3, we briefly illustrate statistical inference for superiority, equivalence, and non-inferiority trials since the latter development is typically conducted with equivalence trials rather than superiority trials. Finally, in Section 1.4, we overview and propose adaptive designs applied in confirmatory clinical trials.

Table 1.2: List of best-selling biologics with patent expiry in the years to come (Rémuzat et al., 2017)

| Drug | Brand name | Initial market authorization date | EU patent expiry date | Categories of licensed indications |
|---|---|---|---|---|
| Adalimumab | Humira | 2003 | 2018 | Rheumatoid arthritis/juvenile idiopathic arthritis<br>enthesitis-related arthritis/psoriatic arthritis<br>axial spondyloarthritis/ankylosing spondylitis<br>psoriasis/Crohn’s disease/ulcerative colitis |
| Bevacizumab | Avastin | 2005 | 2019 | Carcinoma of the colon or rectum/carcinoma of the cervix<br>breast cancer/lung cancer/renal-cell cancer<br>Epithelial ovarian, fallopian tube, primary peritoneal cancer |
| Cetuximab | Erbitux | 2004 | 2014 | Colorectal cancer/cancer of the head and neck |
| Darbepoetin alfa | Aranesp | 2001 | 2016 | Anaemia |
| Enoxaparin | Lovenox | 1999 | 2012 | Anticoagulant |
| Interferon beta-1 | Avonex | 1997 | 2015 | Multiple slerosis |
|  | Rebif | 1998 | 2015 | Multiple slerosis |
| Pegfilgrastim | Neulasta | 2002 | 2017 | Neutropenia |
| Ranibizumab | Lucentis | 2007 | 2016 | Age-related macular degeneration/visual impairment |
| Rituximab | MabThera | 1998 | 2013 | Non-Hodgkin’s lymphoma/chronic lymphocytic leukemia |
| Trastuzumab | Herceptin | 2000 | 2014 | Breast cancer/gastric cancer |

## 1.2 Statistical inference for survival endpoints

Statistical analysis for survival endpoints is called survival analysis that summarizes and analyzes the length of time until the occurrence of an event, e.g., death, as time-to-event process. Clinical trials are generally conducted to assess the effectiveness of new therapies. Hence, investigators usually apply survival analysis in clinical trials to compare the risk of events among population groups receiving different therapies.

### 1.2.1 Summarized statistics

For survival analyses, the median is the preferred summary measure of the location of the distribution since a distribution constructed from survival data typically tends to be positively skewed, that is, the distribution has a longer tail to the right of the interval that contains the largest number of observations. Let $t$ denote the actual survival time of an individual. Once the survival function $S(t)$ has been estimated by the Kaplan and Meier (1958) method, it is straightforward to obtain an estimate of the median survival time (MST), i.e., the time beyond which 50% of the individuals in the proportion are expected to survive. In most clinical studies, survival data are summarized using the MST by group.

Suppose that survival data are exponentially distributed with parameter $\lambda$ where $\lambda > 0$. Let $T$ denote a random variable. Then, the probability density function $f(t)$ is given by

$$f(t) = \lambda \exp(-\lambda t), \tag{1.1}$$

where $t > 0$. The expectation and the variance for $T$ can be written as

$$\begin{aligned} E[T] &= \int_0^\infty t \cdot f(t)dt = \frac{1}{\lambda} \\ E[T^2] &= \int_0^\infty t^2 \cdot f(t)dt = \frac{2}{\lambda^2} \\ V[T] &= E[T^2] - E[T]^2 = \frac{1}{\lambda^2}. \end{aligned}$$

Using Equation (1.1), the survival function $S(t)$ is illustrated as follows:

$$\begin{aligned} S(t) &= P(T \geq t) = P(T < t) \\ &= 1 - F(t) = 1 - \int_0^t f(u)du = \exp(-\lambda t). \end{aligned}$$

Note that the MST is given by $\text{MST} = \log 2/\lambda$ when the survival time is assumed to be exponentially distributed.

### 1.2.2 Log-rank tests

In the comparison of two groups of survival data, there are a number of nonparametric tests that can be used to quantify the extent of the coincidence of between-group differences (Ohashi et al., 2016). Among these methods, the log-rank test is most frequently used in clinical trials. In practice, the p-value obtained from the log-rank tests for survival data is reported. The p-value is given by

$$p = \Phi\left(\frac{U_L}{\sqrt{V_L}}\right),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, $U_L$ is the log-rank score, and $V_L$ is the variance estimate of $U_L$ (Collett, 2015). Note that $U_L/\sqrt{V_L}$ has a normal distribution with zero mean and unit variance denoted by $N(0, 1)$ under the null hypothesis that there is no treatment difference.

Furthermore, consider a randomized clinical trial in which patients are assigned to each group stratified according to variables such as age, sex, performance status, and other potential risk factors for the disease under study. In this case, a stratified log-rank test may be employed. Essentially, this involves calculating $U_L$ and $V_L$ for each stratum, and then combining these values over the strata. Let $U_{L,s}$ and $V_{L,s}$ denote the log-rank score and the variance estimate of $U_{L,s}$, respectively, obtained from the $s$th of $S$ strata. The p-value based on the stratified log-rank test is then based on the statistic is given by

$$p = \Phi\left(\frac{\sum_{s=1}^{S} U_{L,s}}{\sqrt{\sum_{s=1}^{S} V_{L,s}}}\right).$$

Note that $\sum_{s=1}^{S} U_{L,s}/\sqrt{\sum_{s=1}^{S} V_{L,s}}$ has a normal distribution with zero mean and unit variance denoted by $N(0, 1)$ under the null hypothesis that there is no treatment difference.

### 1.2.3 Cox proportional hazards model

In addition to the hypothesis testing like the log-rank test, the proportional hazards model (Cox, 1972) is widely used in clinical trials to illustrate the treatment effect. This model is referred as a semi-parametric model since no particular type of probability distribution is assumed for survival data.

In summarizing survival data, the hazard function $h(t)$ is widely used to express the risk or hazard of event at time $t$. Under the assumption of a continuous distribution with differentiable survival function, the hazard function, also known as the ”force of mortality” (Klein et al., 2013), is defined by

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t},$$

where $\Delta t$ is the time interval and $T$ is a random variable. Note that the relationship with $f(t)$ and $S(t)$ is given by

$$h(t) = \frac{f(t)}{S(t)}.$$

Suppose that patients are randomly assigned to either a control ($C$) or an experimental ($E$) arms. Let $h_j(t)$ be the hazard of events at time $t$ for patients on arm $j \in \{C, E\}$. According to a simple model for survival data of the two groups of patients, the hazard at time $t$ for a patient on $E$ is proportional to the hazard at that same time for a patient on $C$. This proportional hazards model can be expressed in the form

$$h_E(t) = HR \cdot h_C(t) \tag{1.2}$$

for any non-negative value of $t$, where $HR$ is a constant. This assumption means that the corresponding true survival functions for individuals on $C$ and $E$ do not cross. Note that the score test obtained from the proportional hazards model is identical to the log-rank test (Collett, 2015).

The value of $HR$ is the hazard ratio at any time for an individual on $E$ relative to an individual on $C$. If $HR < 1$, the hazard is smaller for an individual on $E$, relative to an individual on $C$. That means that $E$ is an improvement on $C$. On the other hand, if $HR > 1$, the hazard is greater for an individual on $E$, and $C$ is superior.

Suppose that survival data are exponentially distributed with parameter $\lambda$ where $\lambda > 0$. Then, $h(t)$ is expressed as $\lambda$. By using the value of MST and transforming Equation (1.2), $HR$ is also given by

$$HR = \frac{h_E(t)}{h_C(t)} = \frac{\lambda_E}{\lambda_C} = \frac{\mathrm{MST}_C}{\mathrm{MST}_E},$$

where $\lambda_j$ is the hazard and $\mathrm{MST}_j$ is the MST on arm $j \in \{C, E\}$.

### 1.2.4 Survival endpoints

Table 1.3 describes survival endpoints employed in clinical trials with progressive cancer patients. According to the Food and Drug Administration (2007) guidelines, OS is defined as the time from randomization to death from any cause and is usually measured in the intent-to-treat population. With respect to the clinical trials to confirm the effectiveness for a new therapy in oncology, the gold standard endpoint is OS. The reason is that it is simple to measure clinical importance, easy to interpret, and measurement is unbiased. OS is accepted as the most reliable cancer endpoint and therefore is preferred by regulatory agencies. However, OS generally requires long-term follow-up after disease progression. Consequently, it will be expensive to conduct clinical trials since we need quite long time and large number of patients when employing and evaluating OS as a primary endpoint. Thus, short-term survival endpoints such as progression-free survival (PFS) are often set as an alternative to OS in some therapeutic areas.

Table 1.3: Survival endpoints in progressive cancer and their definitions

| Endpoint | Definition |
|---|---|
| Overall survival (OS) | Death from any cause |
| Time-to-progression (TTP) | Objective tumor progression<br>It does not include deaths. |
| Progression-free survival (PFS) | Objective tumor progression or death from any cause |
| Time-to-treatment failure (TTF) | Discontinuation of treatment for any reason,<br>including disease progression, treatment toxicity,<br>and death from any cause |

PFS is defined as the time from randomization until objective tumor progression, i.e., time-to-progression (TTP), or death from any cause, whichever occurs first. PFS has recently gained much importance in oncology clinical trials particularly in phase II because objective response rate (ORR) based on the Response Evaluation Criteria in Solid Tumors (RECIST) criteria (Eisenhauer et al., 2009), frequently used in real phase II trials as a primary endpoint, is not surrogate for OS in the development of molecular targeted therapies (Seymour et al., 2010). In Chapter 3, we focus on PFS as a short-term endpoint for OS.

## 1.3 Statistical inference for different types of comparisons

There are various types of comparisons between the experimental ($E$) and control ($C$) arm. Figure 1.1 illustrates each type of comparisons. Typically, in developing new therapies through the clinical trials, the main purpose is to confirm superiority. That means that the experimental arm is better than the control arm. In this case, the experimental arm is also non-inferior to the control arm. However, under the non-inferiority hypothesis, it is unnecessary to confirm superiority. The non-inferiority objective is fulfilled once the experimental arm is inferior to the control arm by less than some prespecified margin (Rothmann et al., 2011). Thus, the choice of the margin has important practical consequences (Food and Drug Administration, 2016b). In contrast, the absolute difference between the experimental and control arms is smaller than a prespecified margin to meet the equivalence objective.

The equivalence objective is usually set in a bioequivalence trial to develop generic drugs (Food and Drug Administration, 2001). A generic drug is a product which has the same qualitative and quantitative composition in active substances and the same pharmaceutical form as a brand-name product (European Medicines Agency, 2010). In addition, the equivalence trial has recently been conducted to develop biosimilars. A biosimilar is defined as a biological medicinal product that contains an active substance that is similar to that of an original previously authorized biological medicinal product (European Medicines Agency, 2014a). It differs from generic products, e.g., with respect to the complexity and heterogeneity of the molecular structure (Berghout, 2011; Chow, 2013). Nowadays, the development of biosimilars is attractive to pharmaceutical sponsors because healthcare costs for patients are expected to be reduced. As shown in the development of biosimilar insulin glargine (Blevins et al., 2015; Rosenstock et al., 2015), non-inferiority trials are also applied in the development of biosimilars. However, most biosimilars have been developed with equivalence trials. In Chapter 4, we focus on the statistical issues in equivalence trials.

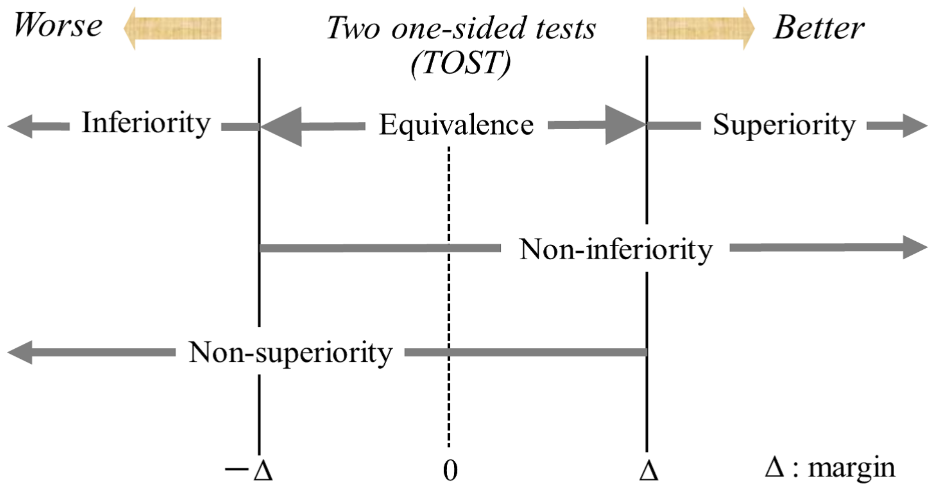

Figure 1.1: Relationship of each type of comparisons

## 1.4 Adaptive designs for confirmatory clinical trials

Randomized controlled clinical trials with two arms are commonly applied in confirmatory clinical trials. In recent years, adaptive clinical trial designs have been attractive to pharmaceutical sponsors, regulatory agencies, and medical investigators, in response to the expectation for conducting clinical trials efficiently. Adaptive designs mean the extension of group sequential designs. Group sequential designs include interim analyses before the formal completion of a trial and possible early stopping for either positive or negative results (Jennison and Turnbull, 2000). In addition to the option for early stopping, adaptive designs also provides decisions based on data accumulated until the interim analyses how to modify design facets without undermining the validity and integrity of the trial (Food and Drug Administration, 2010, 2016a). The aim of the adaptive designs is to increase the likelihood of a successful trial and lower the number of patients exposed to an inferior or harmful treatment (Bretz et al., 2009). However, we have to note that many clinical scientists conceptually misuse or abuse the adaptive design methods in clinical trials (Cheng and Chow, 2010; Wittes, 2010). The relative performance over alternative design options depends on the scenarios, assumptions, and trial objectives (Bretz et al., 2017).

Adaptive designs encompass every phase of clinical trials. In early phase clinical trials, adaptive designs are often applied in real clinical trials in the last ten years owing to their exploratory aspects. Currently, adaptive designs in consideration for biomarker information have been applied in real phase II clinical trials. For example, the I-SPY (Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Analysis) 2 trial using adaptive randomization was reported (Park et al., 2016; Rugo et al., 2016) and the innovative clinical trial designs are expected in a confirmatory stage. Hence, we focus on the adaptive designs for confirmatory clinical trials.

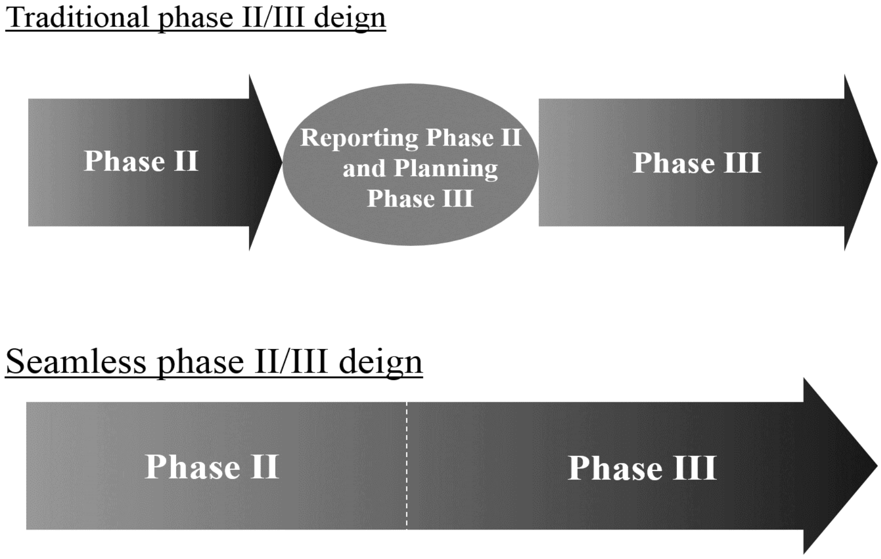
Traditional phase II/III deign
Phase II
Reporting Phase II and Planning Phase III
Phase III
Seamless phase II/III deign
Phase II
Phase III

Figure 1.2: Top: traditional development with two separate phases, Bottom: seamless phase II/III design

Clinical trials with multiple phases can be conducted seamlessly with adaptation in the common protocol when applying adaptive seamless designs in clinical drug developments. With respect to confirmatory clinical trials, adaptive seamless phase II/III designs are increasingly explored in the current research articles to solve the problem for late stage failures and rising costs of phase II/III clinical trials. As shown in Figure 1.2, the adaptive seamless phase II/III designs can minimize the period, white space, between the analysis of phase II data and recruitment of phase III patients, while the seamless approach also allows the flexibility to investigate other crucial issues, without the need for separate trials (Schmidli et al., 2006).

The variety of the adaptive seamless phase II/III designs is roughly divided into an adaptive treatment selection design and an adaptive population selection design. The adaptive treatment selection design could be used for the development comparing a control arm with multiple competing experimental treatments. Any experimental treatments that appear no better than control are quickly rejected in a late phase II trial and those that are significantly better than control are identified in a phase III confirmatory trial. On the other hand, the adaptive population selection design could be used for the development of molecular targeted therapies. any hypotheses for each population are identified in a late phase II trial and confirmed in the selected hypotheses in a phase III trial.

# Chapter 2

# Utility-based interim decision rule planning in adaptive designs for population selection

## 2.1 Introduction

As many molecular targeted therapies whose treatment effect notably differs among subgroups of patients based on biomarker information in oncology have recently been developed by numerous pharmaceutical sponsors (Aggarwal, 2010), various clinical trial designs considering patient heterogeneity have been proposed (Biankin et al., 2015). Often a benefit is found only for a subgroup in the confirmatory clinical trial report without any pre-specification in statistical planning; for example, see Karapetis et al. (2008). To deal with this issue, some clinical trials pre-specify a subpopulation in the statistical analysis plan in addition to the overall population, in accordance with a multiple testing procedure (Goteti et al., 2014). In most instances, a fixed clinical trial design is applied in the setting of targeted therapy development. However, some patients recruited in the development of molecular targeted therapies cannot help being exposed to unfavorable treatments during a trial with such a traditional design. In particular, even worse, some molecular targeted therapies show a positive effect only for the subpopulation but a negative effect otherwise; for example, see Mok et al. (2009). From an ethical viewpoint, patients who are likely to be harmed by the therapy, as identified by biomarker information, must be dropped from the trial, even though such information from the clinical trial is required by regulatory agencies.

In this work, we consider that the full population comprises biomarker-positive and biomarker-negative populations categorized based on a promising biomarker, as shown in Figure 2.1, in a setting wherein a promising biomarker exists and in which the targeted therapy is beneficial only for the biomarker-positive population. If there are a sufficient number of patients assigned to each population, adaptive population selection designs are highly attractive, with the aim of developing molecular targeted therapies. Figure 2.2 illustrates the schematic representation of the adaptive population selection design. In this context, an interim analysis is conducted in midcourse to identify whether the entire population or a subpopulation is benefitting (Food and Drug Administration, 2010). Thus, interim analysis can play a role in determining whether the targeted population is restricted to only the biomarker-positive population owing to futility for the biomarker-negative population or is continued as the full population. For the sake of simplicity, we consider the interim analysis that discontinues the trial early only for futility.

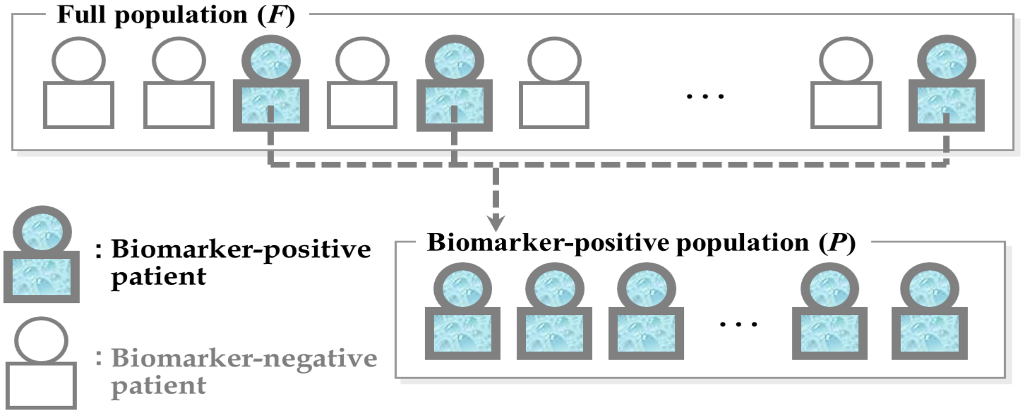
Full population (*F*)
. . .
: Biomarker-positive patient
: Biomarker-negative patient
Biomarker-positive population (*P*)
. . .

Figure 2.1: Assumed populations driven by a biomarker

We consider a design setting in which the endpoint is of the time-to-event type, such as overall survival or progression-free survival. In this context, one of the simple measures used for an interim decision is the hazard ratio for each population. For instance, Jenkins et al. (2011) and Friede et al. (2012) incorporated hazard ratios into the interim decision in their methodology. As an example of a real clinical trial, the futility criteria for the interim analysis were set to be achieved if the estimated hazard ratio at interim exceeded 1.00, and then, futility stopping was consequently fulfilled at the interim analysis, i.e., the estimated hazard ratio was 1.09 (Fujitani et al., 2016). However, a point estimate of a hazard ratio is not reasonable, since the immature sample size at interim could possibly result in a misleading decision. In particular, if this criterion is applied to an interim decision with adaptive population selection, the sample size in the biomarker-positive group is even more inadequate than the overall sample size.

Alternatively, stochastic curtailment, based on the conditional or predictive power, is also frequently used in practice (Jennison and Turnbull, 2000). In the context of adaptive population selection design, Brannath et al. (2009) and Uozumi and Hamada (2017b) considered the use of predictive power in the setting of survival endpoints, whereas Wang et al. (2009) and Götte et al. (2015) evaluated performance using conditional power. Predictive power is relatively preferred as an interim decision feature because conditional power is highly dependent on the interim result, regardless of having an insufficient sample size in midcourse (Spiegelhalter et al., 1986). Even if either statistic is employed in the interim decision, searching the thresholds is not straightforward, and a thorough investigation to confirm the operating characteristic depending on the setting would be required via simulation to pre-specify the thresholds for each population at the planning stage. However, there are no well-established procedures for setting the threshold to select the population at interim in the context of adaptive population selection design.

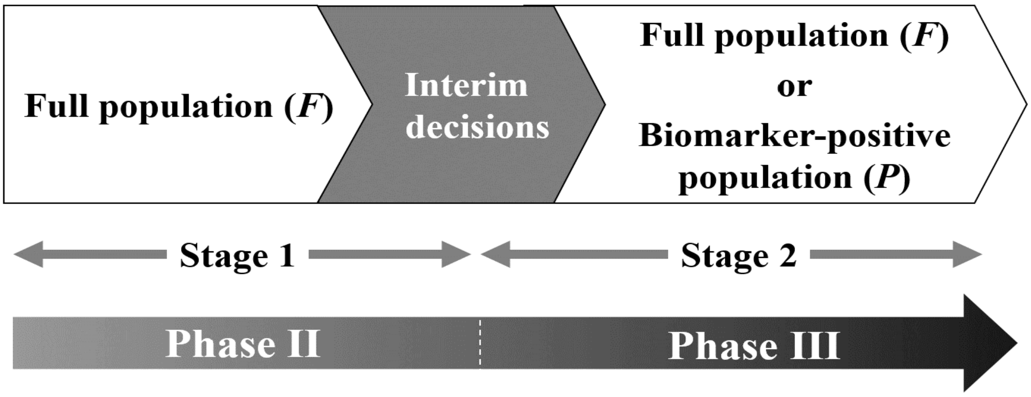

Figure 2.2: Schematic representation of the adaptive population selection design

In this chapter, we propose a novel approach to setting the thresholds in an interim decision rule constructed on the basis of utility functions in the context of adaptive population selection design. The utility functions are motivated by the work of Graf et al. (2015) in which several trial designs for the development of molecular targeted therapies were compared with consideration of the gain for sponsors and public health. We use utility functions that consist of gain and loss functions solely to help in formulating the interim decision rule to determine whether the entire population is continued or only the promising population is selected. Following the formulation of an interim decision rule based on the proposed approach, we consider the situation in which the interim analysis is performed based on the hazard ratios or predictive powers for each population and the final analysis is performed using a log-rank test. We further examine the proposed approach by using the difference and ratio of the restricted mean survival time within the interim and final decisions. To assess the proposed approach, we evaluate the operating characteristics with respect to the probability of selecting each population and the probability of rejecting the hypotheses for each population in the setting in which the targeted therapy is considered.

The remainder of this chapter is structured as follows. In Section 2.2, we introduce the setting of the adaptive design considered in this chapter in the case in which the type of endpoints is time-to-event. In Section 2.3, we specify the measures used within the interim decision rule as follows: hazard ratio, predictive power, and difference and ratio of restricted mean survival times. In Section 2.4, we provide the utility functions and describe how to incorporate those functions into the method for setting the thresholds in the interim decision. Section 2.5 presents a simulation study, and we provide concluding remarks in Section 2.6.

## 2.2 Setting of adaptive population selection design with two stages using survival endpoints

In this chapter, we focus on adaptive population selection design for the development of molecular targeted therapies. We assume, for the sake of simplicity, that the full population ($F$) comprises biomarker-positive ($P$) and biomarker-negative ($N$) populations without any consideration of an unknown group. Let $H_g$ denote the null hypothesis for population $g \in \{F, P\}$. We consider one interim analysis in midcourse to identify whether $F$ or only $P$ is to continue into the subsequent stage. This means that the interim analysis provides an opportunity to stop recruitment for $N$ in the case in which the targeted therapy is deemed as promising only for $P$. Multiple testing problems with respect to the interim analysis do not arise because we consider the interim analysis only for futility.

With respect to the final analysis, we consider two null hypotheses, $H_F$ and $H_P$. Both $H_F$ and $H_P$ are tested if $F$ is determined to be continued. On the other hand, only $H_P$ is tested if only $P$ is determined to be continued and $N$ is terminated at the interim analysis. Let $p_{k,g}$ denote the p-value for stage $k$ ($k = 1, 2$) in population $g \in \{F, P\}$. We use combination methods to perform the hypothesis testing at the final analysis. There are two methods: the Fisher's product combination method (Bauer and Köhne, 1994) and the weighted inverse normal combination method (Lehmacher and Wassmer, 1999). To construct the combined p-value, we employ the weighted inverse combination method,

$$C(p_{1,g}, p_{2,g}) = 1 - \Phi\{w_1\Phi^{-1}(1 - p_{1,g}) + w_2\Phi^{-1}(1 - p_{2,g})\},$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution and $w_k$ denotes the weight for each stage chosen by $w_k = \sqrt{d_k / \sum_j d_j}$, where $0 \leq w_k \leq 1$ and $\sum_{k=1}^{2} w_k^2 = 1$, with pre-specified number of events, $d_k$, at stage $k$ on the assumption that $p_{1,g}$ and $p_{2,g}$ are independent and uniformly distributed under the null hypotheses. The weighted inverse combination method can be extended to the setting with multiple stages.

To solve multiple testing issues for $H_F$ and $H_P$, the use of Simes (1986) procedure is applied to the p-value for the intersection hypothesis $H_{FP} = H_F \cap H_P$ given by $p_{k,FP} = \min[2 \cdot \min(p_{k,F}, p_{k,P}), \max(p_{k,F}, p_{k,P})]$, since the familywise type I error rate is controlled in the situation in which $p_{k,F}$ and $p_{k,P}$ are non-negatively correlated or independent (Samuel-Cahn, 1996; Sarkar and Chang, 1997). Using the weighted inverse combination method, the final decision for each hypothesis is made in accordance with the closure principle (Marcus et al., 1976). That is, $H_F$ is rejected if $C(p_{1,FP}, p_{2,FP}) < \alpha$ and $C(p_{1,F}, p_{2,F}) < \alpha$ are satisfied and $H_P$ is rejected if $C(p_{1,FP}, p_{2,FP}) < \alpha$ and $C(p_{1,P}, p_{2,P}) < \alpha$ are satisfied in case where $F$ is determined to be continued, whereas $H_P$ is rejected if $C(p_{1,FP}, p_{2,P}) < \alpha$ and $C(p_{1,P}, p_{2,P}) < \alpha$ are satisfied in case where only $P$ is determined to be continued and $N$ is terminated at the interim analysis, where $\alpha$ denotes a one-sided significance level. For a one-sided significance level of $\alpha = 0.025$, we set the critical value as $\Phi^{-1}(1 - \alpha) \risingdotseq 1.96$. Note that it is vital to specify the combination function and the design of stage 1 at the planning stage.

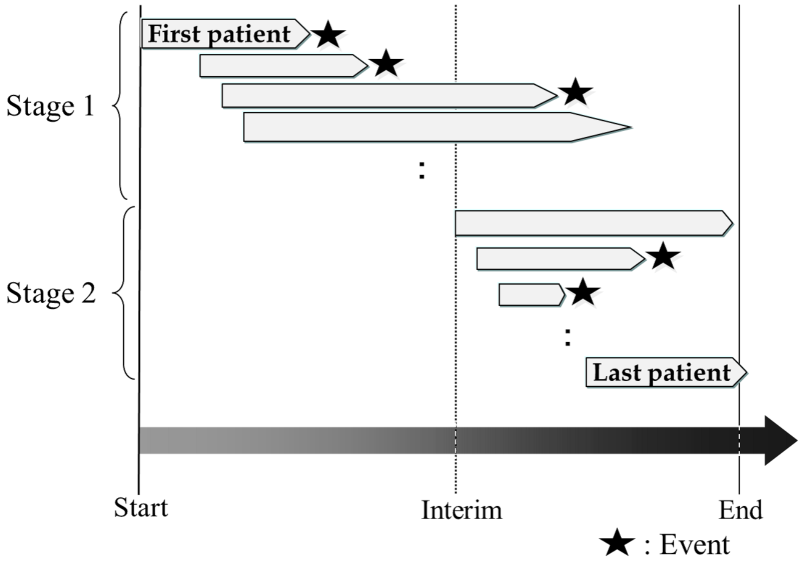

Figure 2.3: Handling survival data with two-stage designs

To handle the setting in which the endpoint is time-to-event, the p-values are derived following the framework of Jenkins et al. (2011) to control the familywise type I error rate. As shown in Figure 2.3, if the patient accrued in the first stage has an event during the second stage, then this event is handled as the p-value for the first stage.

## 2.3 Types of measures used in the interim decision

The aim of this work is to propose a novel utility-based approach to guide the construction of the interim decision rule of an adaptive population selection design in the setting of survival endpoints. This section describes how to assist in constructing the interim decision rule using utility functions. With regard to the final analysis of survival data, the traditional method is to perform a log-rank test and Cox proportional hazard model to show the treatment effect through the hazard ratio and its confidence interval. In particular, the Cox proportional hazard model is popularly applied even in the situation where the proportional hazard assumption is violated (Schemper, 1992).

### 2.3.1 Simple hazard ratio

Because we assume that the final analysis will be performed using the Cox proportional hazard model to show the treatment effect by the hazard ratio and its confidence interval, the midcourse hazard ratio can be used as the measure within the interim decision. Derivation of the hazard ratio was described in Section 1.2.3. Jenkins et al. (2011) and Friede et al. (2012) used an interim decision using the point estimate of the interim hazard ratio $\widehat{HR}_{I,g}$ for population $g \in \{F, P\}$. Let $\eta_g^{HR}$ denote the threshold of the hazard ratio for $g \in \{F, P\}$ at interim. Based on previous work (Friede et al., 2012; Jenkins et al., 2011), we consider the interim decision rule as follows. (i) Continue $F$ if $\widehat{HR}_{I,F} < \eta_F^{HR}$, regardless of the value of $\widehat{HR}_{I,P}$. (ii) Continue only $P$ if $\widehat{HR}_{I,F} \geq \eta_F^{HR}$ and $\widehat{HR}_{I,P} < \eta_P^{HR}$. (iii) Stop the trial for futility if $\widehat{HR}_{I,F} \geq \eta_F^{HR}$ and $\widehat{HR}_{I,P} \geq \eta_P^{HR}$, where the experimental treatment is beneficial against the control if the hazard ratio is less than 1.00. Note that this decision rule is constructed using the point estimate, rather than the confidence interval, of the hazard ratio.

However, an interim decision using the point estimate rather than the confidence limit would become unstable because the sample size at interim is insufficient, particularly for $P$. Therefore, we further consider the interim decision rule based on the upper limit of the confidence interval for the hazard ratio estimated from the interim data denoted by $\widehat{HR}_{I,g}^{U}$ for $g \in \{F, P\}$ as follows. (i) Continue $F$ if $\widehat{HR}_{I,F}^{U} < \eta_F^{HR}$, regardless of the value of $\widehat{HR}_{I,P}^{U}$. (ii) Continue only $P$ if $\widehat{HR}_{I,F}^{U} \geq \eta_F^{HR}$ and $\widehat{HR}_{I,P}^{U} < \eta_P^{HR}$. (iii) Stop the trial for futility if $\widehat{HR}_{I,F}^{U} \geq \eta_F^{HR}$ and $\widehat{HR}_{I,P}^{U} \geq \eta_P^{HR}$, where the experimental treatment is beneficial against the control if the hazard ratio is less than 1.00.

### 2.3.2 Predictive power

Alternatively, stochastic curtailment based on the conditional or predictive power is often used to make an interim decision for early stopping owing to futility in the context of a group sequential design. Since the conditional power must be obtained for a specified treatment effect, which might not be supported by the final data (Jennison and Turnbull, 2000), we focus on the predictive power as an alternative to the hazard ratio. Because we assume that the final analysis will be performed using a log-rank test to derive the p-value, calculation of the predictive power is conducted based on the result of the test statistic by the log-rank test and the assumption that non-informative priors for the treatment effect are employed for the sake of simplicity (Dmitrienko and Wang, 2006). That is, the predictive power for population $g \in \{F, P\}$ is given by

$$PP_g = 1 - \Phi\left[\left(1 - \Pi_g\right)^{-1/2}\left(\Phi^{-1}(1-\alpha) \cdot \sqrt{\Pi_g} - z_g\right)\right] \tag{2.1}$$

where $\Pi_g$ represents the event fraction at the interim analysis and $z_g$ is the observed test statistic based on the log-rank test using stage 1 data.

If the predictive power based on interim results is greater than the threshold for one hypothesis, then the possibility of rejection is high for the corresponding hypothesis when continuing with the corresponding population. In a manner similar to that of the interim decision using the hazard ratio, we use the predictive power $\widehat{PP}_g$ for $g \in \{F, P\}$ in the interim decision rule as follows. (i) Continue $F$ if $\widehat{PP}_F > \eta_F^{PP}$, regardless of the value of $\widehat{PP}_P$. (ii) Continue only $P$ if $\widehat{PP}_F \leq \eta_F^{PP}$ and $\widehat{PP}_P > \eta_P^{PP}$. (iii) Stop the trial for futility if $\widehat{PP}_F \leq \eta_F^{PP}$ and $\widehat{PP}_P \leq \eta_P^{PP}$, where $\eta_g^{PP}$ is the threshold of predictive power for $g \in \{F, P\}$.

### 2.3.3 Restricted mean survival time

As an alternative to hazard ratio for survival data between treatment groups, the measure constructed using the restricted mean survival time is recommended, when the proportional hazard assumption is violated (Royston and Parmar, 2011; Uno et al., 2014). Furthermore, as complementary information on summarizing the survival endpoints, presentation of the restricted mean survival time is currently recommended in oncology clinical trials (A'Hern, 2016). The restricted mean survival time was first proposed by Irwin (1949) and then extended to adjust covariates (Karrison, 1987; Zucker, 1998).

Let $RMD_g$ and $RMR_g$ denote the difference and ratio, respectively, of the restricted mean survival times for $g \in \{F, P\}$. The interim estimates, $\widehat{RMD}_g$ and $\widehat{RMR}_g$, of $RMD_g$ and $RMR_g$, respectively, are given by

$$\widehat{RMD}_g = \int_0^{t^*} \hat{S}_{g,E}(t)dt - \int_0^{t^*} \hat{S}_{g,C}(t)dt \tag{2.2}$$

and

$$\widehat{RMR}_g = \int_0^{t^*} \hat{S}_{g,E}(t)dt / \int_0^{t^*} \hat{S}_{g,C}(t)dt \tag{2.3}$$

where $t^*$ is the time point when the interim decision is made and $\hat{S}_{g,E}$ and $\hat{S}_{g,C}$ are the estimated survival functions for $g \in \{F, P\}$ in experimental ($E$) and control ($C$) groups. As described by Equations (2.2) and (2.3), the restricted mean survival time in each group is calculated by the area under the survival curve up to $t^*$.

In a manner similar to that of the interim decision using the hazard ratio, we use $RMD_g$ and $RMR_g$ for $g \in \{F, P\}$ solely in the interim decision rule. First, using the lower limit of $\widehat{RMD}_g$ denoted by $\widehat{RMD}_g^L$, the interim decision rule is defined as follows. (i) Continue $F$ if $\widehat{RMD}_F^L > \eta_F^{RMD}$, regardless of the value of $\widehat{RMD}_P^L$. (ii) Continue only $P$ if $\widehat{RMD}_F^L \leq \eta_F^{RMD}$ and $\widehat{RMD}_P^L > \eta_P^{RMD}$. (iii) Stop the trial for futility if $\widehat{RMD}_F^L \leq \eta_F^{RMD}$ and $\widehat{RMD}_P^L \leq \eta_P^{RMD}$, where $\eta_g^{RMD}$ is the threshold of $RMD_g$ for $g \in \{F, P\}$. Second, similar to $RMD_g$, the interim decision rule using the lower limit of $\widehat{RMR}_g$ denoted by $\widehat{RMR}_g^L$ is considered as follows. (i) Continue $F$ if $\widehat{RMR}_F^L > \eta_F^{RMR}$, regardless of the value of $\widehat{RMR}_P^L$. (ii) Continue only $P$ if $\widehat{RMR}_F^L \leq \eta_F^{RMR}$ and $\widehat{RMR}_P^L > \eta_P^{RMR}$. (iii) Stop the trial for futility if $\widehat{RMR}_F^L \leq \eta_F^{RMR}$ and $\widehat{RMR}_P^L \leq \eta_P^{RMR}$, where $\eta_g^{RMR}$ is the threshold of $RMR_g$ for $g \in \{F, P\}$.

As noted, we assume that the final analysis will be performed in the traditional manner using a log-rank test and Cox proportional hazard model. However, either $RMD_g$ or $RMR_g$ can be used similarly, instead of the hazard ratio, in the final analysis, when the interim decision is made by the interim $RMD_g$ or $RMR_g$, but the required sample size is increased, because using $RMD_g$ or $RMR_g$ is less powerful than using the hazard ratio (Trinquart et al., 2016).

## 2.4 Formulation of an interim decision rule based on utility functions

In the interim decision rule presented in Section 2.3, sponsors are required to set the thresholds for each measure beforehand. Currently, there are no statistical approaches for setting the thresholds. Hence, sponsors tend to decide each threshold based on thorough simulation results considering the cost and benefit for the development.

Here, we use gain and loss functions as utility functions to construct the optimal decision rule at interim. Let $\eta_g$ denote the general expression of the threshold of a particular measure, such as the hazard ratio for $g \in \{F, P\}$. We provide utility functions in terms of sponsors (Sp) and patients (Pat) given by

$$U^{\zeta}(\eta_F, \eta_P) = G^{\zeta}(\eta_F, \eta_P) - L^{\zeta}(\eta_F, \eta_P)$$

where $G^{\zeta}$ and $L^{\zeta}$ denote the gain and loss functions, respectively, for $\zeta \in \{\mathrm{Sp}, \mathrm{Pat}\}$. We assume that each function is constructed on the basis of pseudo clinical trial results via simulation. With the use of the expected utility, $\mathbb{E}\left[U^{\zeta}(\eta_F, \eta_P)\right]$, the optimal thresholds $\tilde{\eta}_F$ and $\tilde{\eta}_P$ are explored via grid search to satisfy

$$\arg\max_{\eta_F, \eta_P} \mathbb{E}\left[U^{\zeta}(\eta_F, \eta_P)\right] \tag{2.4}$$

for $\zeta \in \{\mathrm{Sp}, \mathrm{Pat}\}$, where $\eta_F$ and $\eta_P$ are subject to the restrictions depending on the measure. For instance, when $\eta_g$ is considered as $\eta_g^{HR}$, i.e., the threshold of the hazard ratio, $\eta_g^{HR}$ is set to be greater than 1.00 in order to prevent the situation in which the decision rule at interim is stricter than that at the final analysis.

Here, we describe the gain and loss functions in terms of sponsors and patients. First, sponsors have a great deal to be concerned about in rejecting $H_g$ for $g \in \{F, P\}$. Thus, each function is shown as

$$\begin{aligned} G^{\mathrm{Sp}}(\eta_F, \eta_P) &= \{\delta_F I(\Omega = F) + \delta_P I(\Omega = P)\} I(r_F = 1) I(r_P = 1) \\ &+ \delta_F I(\Omega = F) I(r_F = 1) I(r_P = 0) + \delta_P I(\Omega = P) I(r_F = 0) I(r_P = 1) \\ &= \delta_F I(\Omega = F) I(r_F = 1) + \delta_P I(\Omega = P) I(r_P = 1) \end{aligned} \tag{2.5}$$

and

$$\begin{aligned} L^{\mathrm{Sp}}(\eta_F, \eta_P) = {} & \delta_F \{I(\Omega = P) + I(\Omega = \phi)\} I(r_F = 1) I(r_P = 1) \\ & + \{\delta_P I(\Omega = P) + \delta_F I(\Omega = \phi)\} I(r_F = 1) I(r_P = 0) \\ & + \{\delta_F I(\Omega = F) + \delta_P I(\Omega = \phi)\} I(r_F = 0) I(r_P = 1) \\ & + \{\delta_F I(\Omega = F) + \delta_P I(\Omega = P)\} I(r_F = 0) I(r_P = 0), \end{aligned} \tag{2.6}$$

respectively, where $I(\cdot)$ is the indicator function that takes the value one if $\cdot$ is true and zero otherwise, $\Omega$ denotes the selected population determined based on the interim analysis, $\delta_g$ denotes the discount parameter for population $g \in \{F, P\}$, $r_g$ denotes an indicator variable such that

$r_g = 1$, if $H_g$ for population $g \in \{F, P\}$ is rejected at the final analysis regardless of the interim results, and $\phi$ indicates that neither $F$ nor $P$ is selected. On the one hand, Equation (2.5) says that the gain function is achieved in the case that the correct population is selected at the interim analysis in the situation in which either $H_F$ or $H_P$ or both $H_F$ and $H_P$ are rejected at the final analysis. With regard to the population, $P$ is obviously narrower than $F$. If the sponsors correspond to a pharmaceutical company, then their interest includes the market size after regulatory approval. Hence, we incorporate the discount parameters as $\delta_F = 1.0$ and $0 \leq \delta_P \leq 1.0$ into each function. Each term of Equation (2.5) shows that gain is accumulated with $\delta_g$ in case where the corresponding population $g \in \{F, P\}$ is selected at the interim analysis, denoted by $\Omega = g$, in the situation in which $H_g$ is rejected at the final analysis, denoted by $r_g = 1$. On the other hand, Equation (2.6) indicates that loss is accumulated, if the appropriate population is terminated at the interim analysis, regardless of the situation in which either $H_F$ or $H_P$ is rejected, or an inappropriate population is continued to the following stage, even though neither $H_F$ nor $H_P$ is rejected. Note that the first term of Equation (2.6), $\delta_F\{I(\Omega = P) + I(\Omega = \phi)\}I(r_F = 1)I(r_P = 1)$, shows that loss is accumulated with $\delta_F$ if only $P$ is determined to be continued or the trial is determined to be discontinued at interim, regardless of the situation in which both $H_F$ and $H_P$ can be rejected at the final analysis. In the same way, the second and third terms of Equation (2.6), $\{\delta_P I(\Omega = P) + \delta_F I(\Omega = \phi)\}I(r_F = 1)I(r_P = 0)$ and $\{\delta_F I(\Omega = F) + \delta_P I(\Omega = \phi)\}I(r_F = 0)I(r_P = 1)$, denote that loss is accumulated if the appropriate population is terminated at interim in the situation in which the hypothesis for the terminated population can be rejected at the final analysis. Finally, the fourth term of (5), $\{\delta_F I(\Omega = F) + \delta_P I(\Omega = P)\}I(r_F = 0)I(r_P = 0)$, indicates that loss is accumulated if either $F$ or $P$ is determined to be continued at interim even though neither hypothesis can be rejected at the final analysis.

Second, patients are concerned with the targeted range of an approved therapy. For instance, the patients whose population is $N$ are harmed in a situation in which marketing approval reaches $F$ since the fact in which the only $P$ is beneficial was not observed in the relevant trial, although the true treatment effect should have been restricted to $P$. This case discounts as loss $\delta_F$ in $L^{\mathrm{Sp}}$ of Equation (2.6). To handle the issue above, the loss function in terms of patients is given by

$$\begin{aligned} L^{\mathrm{Pat}}(\eta_F, \eta_P) = &\{\delta_N I(\Omega = P) + \delta_F I(\Omega = \phi)\}I(r_F = 1)I(r_P = 1) \\ &+ \{\delta_P I(\Omega = P) + \delta_F I(\Omega = \phi)\}I(r_F = 1)I(r_P = 0) \\ &+ \{\delta_N I(\Omega = F) + \delta_P I(\Omega = \phi)\}I(r_F = 0)I(r_P = 1) \\ &+ \{\delta_F I(\Omega = F) + \delta_P I(\Omega = P)\}I(r_F = 0)I(r_P = 0). \end{aligned} \tag{2.7}$$

Note that $L^{\mathrm{Pat}}$ is less than $L^{\mathrm{Sp}}$ because the discount parameter $0 \leq \delta_N \leq 1.0$ is incorporated into Equation (2.7), when $F$ is selected even if only $H_P$ is intrinsically rejected, denoted by $\delta_N I(\Omega = F)I(r_F = 0)I(r_P = 1)$ obtained from the third term of (2.7), or only $P$ is selected even if $H_F$ is rejected as well as $H_P$, denoted by $\delta_N I(\Omega = P)I(r_F = 1)I(r_P = 1)$ obtained from the first term of (2.7). $L^{\mathrm{Pat}}$ is comparable to $L^{\mathrm{Sp}}$, if $\delta_P = \delta_N = 1.0$. With regard to the gain function, $G^{\mathrm{Pat}}$ is identical to $G^{\mathrm{Sp}}$. Therefore, $U^{\mathrm{Pat}}$ is comparable to $U^{\mathrm{Sp}}$ as well, if $\delta_P = \delta_N = 1.0$.

## 2.5 Simulation study

We have described how to optimize the thresholds in an interim decision using utility functions in the context of adaptive population selection designs. We evaluated the performance of the proposed approach to derive thresholds that maximized the utility in an interim decision and to calculate the probability of selecting the population at the interim analysis based on specifying thresholds. The design assumptions and simulation setting we assume for the development of a targeted therapy are illustrated in Section 2.5.1. The operating characteristics based on the simulation results are presented in Section 2.5.2.

### 2.5.1 Simulation setting

First, we consider a two-stage randomized, parallel-group clinical trial with two arms, comprising experimental ($E$) and control ($C$) groups. We assume that the final analysis is to be conducted with a one-sided significance level of 2.5% following the closure principle as described in Section 2.2. In addition, an interim analysis is conducted to determine whether the population continues as $F$ or is restricted to $P$, as described in Section 2.3. The interim analysis is conducted after half of the pre-specified events set as 300 occur among the pre-specified sample size of 400. The median time for $C$ is set as 6.5 months for both $P$ and $N$ and that for $E$ is derived via the hazard ratio, where the distribution is assumed to be exponential.

Let $HR_g$ denote the assumed hazard ratio for $g \in \{F, P\}$. With respect to the treatment effect difference between $E$ and $C$, we employ the hazard ratio for $g \in \{F, P\}$ as Scenarios 1 to 4. We assume that the treatment effect for $P$ is promising across every scenario, i.e., $HR_P = 0.50$. Regarding the treatment effect for $N$, as shown in Figure 2.4, Scenario 1 is assumed to indicate a tendency similar to that of $P$, i.e., $HR_N = 0.50$, whereas Scenario 2 assumes that $E$ is more beneficial in $P$ than in $N$, i.e., $HR_N = 0.90$. Scenario 3 assumes that $E$ is beneficial only in $P$, but not in $N$, i.e., $HR_N = 1.00$. We further assume Scenario 4, in which $E$ is remarkably harmful in $N$, i.e., $HR_N = 1.43$. The hazard ratio for $F$ is handled as $\exp\{\psi \cdot \log HR_P + (1-\psi) \cdot \log HR_N\}$, where $\psi$ denotes the prevalence of $P$ under the assumption that $F$ is categorized as $P$ and $N$, without consideration of an unknown group for simplicity. Even if we employ either $RMD_g$ or $RMR_g$ at the interim analysis, we assume that the hazard ratio derived by using the Cox proportional hazard model is to be used at the final analysis.

Regarding the utility functions, we set $\delta_P = \delta_N = 1.0, 0.5$ in accordance with $\psi$ being assumed to be 50%, in addition to the setting of $\delta_F = 1.0$. Thus, we treated $\delta_P$ as a relative discount of $P$ to $F$ with respect to the setting of $\delta_P = 0.5$. In the case that there are multiple values satisfying Equation (2.4), the mean value was used as a threshold in the interim decision. In the interim decision, each threshold, viz., $\eta_g^m$, where $m \in \{HR, PP, RMD, RMR\}$, of $\widehat{HR}_{I,g}^U$, $\widehat{PP}_g$, $\widehat{RMD}_g^L$, and $\widehat{RMR}_g^L$ for $g \in \{F, P\}$, has the restrictions $\eta_g^{HR} \geq 1.00$, $0 \leq \eta_g^{PP} \leq 0.80$, $\eta_g^{RMD} \leq 0$, and $\eta_g^{RMR} \leq 1.00$ during the grid search on the basis of Equation (2.4). If we assume the treatment effect for $P$ under null hypothesis, i.e., $HR_P = 1.00$, we confirmed that the optimal thresholds are found with lower or upper limit of aforementioned restrictions. Hence, we solely assume that the treatment effect for $P$ is promising across every scenario.

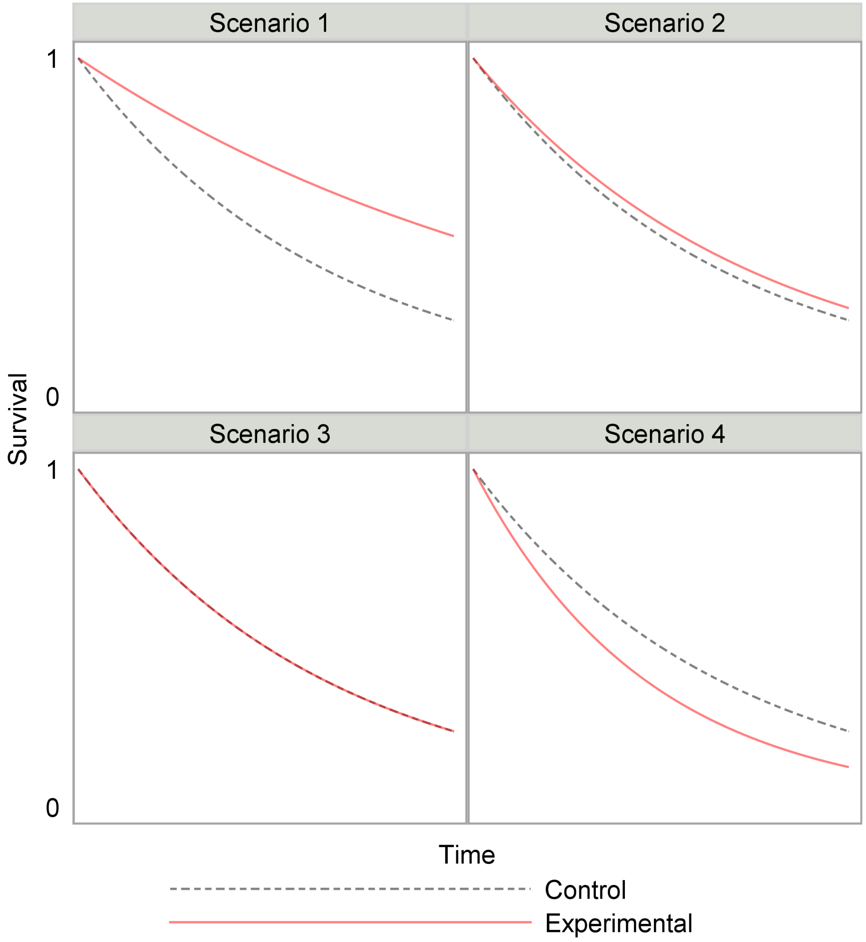

Figure 2.4: Assumed scenarios of survival distribution comparing $E$ (solid) and $C$ (dashed) in $N$

### 2.5.2 Simulation results

Calculation was performed based on 10,000 simulations to obtain all subsequent results per scenario.

First, Tables 2.1 to 2.4 show the results for the optimal thresholds that maximized Equation (2.4) by grid search under the restrictions described in Section 2.5.1, wherein $\tilde{\eta}_g^m$ using $m \in \{HR, PP, RMD, RMR\}$ is the optimal thresholds for $g \in \{F, P\}$, and the interim decision was conducted with $\tilde{\eta}_g^m$. There is an opposite tendency when comparing $\tilde{\eta}_g^{HR}$ with $\tilde{\eta}_g^{PP}$, $\tilde{\eta}_g^{RMD}$, and $\tilde{\eta}_g^{RMR}$ since the direction of positive or negative effects of $\widehat{HR}_{I,g}$ is the reverse of that of $\widehat{PP}_g$, $\widehat{RMD}_g$, and $\widehat{RMR}_g$. Note that in our rule at the interim decision, $H_P$ is at any rate tested, even if $F$ is continued at the interim analysis. Hence, the optimal threshold for $P$ was stricter than that for $F$, regardless of the difference in the sample size in each population in midcourse, especially in Scenario 1, in which we assumed that the treatment effect was observed in both $F$ and $P$. Moreover, the optimal threshold for $F$ was stricter than that for $P$, when the treatment effect in $F$ was diluted in Scenarios 2 to 4. In particular, the optimal threshold for $F$ was searched as the strictest value under the restriction as mentioned in Section 5.1 when $\delta_P$ is set to 1.0 in Scenario 4, in which there is actually a harmful effect in $N$. However, the optimal threshold for $F$ was de-escalated when $\delta_P$ was set to 0.5, since we considered the discount for $P$ by $\delta_P$, wherein $P$ is evidently narrower than $F$, affecting both the sponsors who pay the trial cost and who attempt to obtain a benefit and the patients who are exposed to the targeted therapy. With regard to the type of utility functions, it can be confirmed that the thresholds constructed using utility functions between sponsors and patients are comparable, since $U^{\text{Pat}}$ is identical to $U^{\text{Sp}}$ when $\delta_P = \delta_N = 1.0$.

Table 2.1: Results of optimal thresholds of interim hazard ratio, $\tilde{\eta}_g^{HR}$, for $g \in \{F, P\}$, full ($F$) and biomarker-positive ($P$) populations, derived by grid search when the interim decision rule is constructed from utility functions in terms of sponsors (Sp) and patients (Pat)

| Scenario | $HR_N$ | $\tilde{\eta}_F^{HR}$ | | | $\tilde{\eta}_P^{HR}$ | | |
|---|---|---|---|---|---|---|---|
| | | Sp and Pat | Sp | Pat | Sp and Pat | Sp | Pat |
| | | ($\delta_P = 1.0$) | ($\delta_P = 0.5$) | ($\delta_P = 0.5$) | ($\delta_P = 1.0$) | ($\delta_P = 0.5$) | ($\delta_P = 0.5$) |
| 1 | 0.50 | 1.68 | 1.68 | 1.68 | 1.50 | 1.50 | 1.50 |
| 2 | 0.90 | 1.25 | 1.35 | 1.35 | 1.83 | 1.63 | 1.65 |
| 3 | 1.00 | 1.20 | 1.25 | 1.25 | 1.75 | 1.60 | 1.78 |
| 4 | 1.43 | 1.00 | 1.05 | 1.05 | 1.60 | 1.60 | 1.60 |

$HR_N$ is the assumed hazard ratio for biomarker-negative ($N$) population; $HR_P$ is the assumed hazard ratio for $P$ and is set to 0.50; $\delta_P$ is the discount parameter for $P$; $\delta_N$ is equal to $\delta_P$ since we assume that the prevalence of $P$, $\psi$, is set to 50%.

Table 2.2: Results of optimal thresholds of predictive power, $\tilde{\eta}_g^{PP}$, for $g \in \{F, P\}$, full ($F$) and biomarker-positive ($P$) populations, derived by grid search when the interim decision rule is constructed from utility functions in terms of sponsors (Sp) and patients (Pat)

| Scenario | $HR_N$ | $\tilde{\eta}_F^{PP}$ | | | $\tilde{\eta}_P^{PP}$ | | |
|---|---|---|---|---|---|---|---|
| | | Sp and Pat | Sp | Pat | Sp and Pat | Sp | Pat |
| | | ($\delta_P = 1.0$) | ($\delta_P = 0.5$) | ($\delta_P = 0.5$) | ($\delta_P = 1.0$) | ($\delta_P = 0.5$) | ($\delta_P = 0.5$) |
| 1 | 0.50 | 0.03 | 0.03 | 0.03 | 0.40 | 0.40 | 0.40 |
| 2 | 0.90 | 0.20 | 0.05 | 0.05 | 0.00 | 0.00 | 0.00 |
| 3 | 1.00 | 0.25 | 0.15 | 0.15 | 0.00 | 0.00 | 0.00 |
| 4 | 1.43 | 0.80 | 0.70 | 0.70 | 0.05 | 0.05 | 0.05 |

$HR_N$ is the assumed hazard ratio for biomarker-negative ($N$) population; $HR_P$ is the assumed hazard ratio for $P$ and is set to 0.50; $\delta_P$ is the discount parameter for $P$; $\delta_N$ is equal to $\delta_P$ since we assume that the prevalence of $P$, $\psi$, is set to 50%.

Table 2.3: Results of optimal thresholds of interim difference of restricted mean survival times, $\tilde{\eta}_g^{RMD}$, for $g \in \{F, P\}$, full ($F$) and biomarker-positive ($P$) populations, derived by grid search when the interim decision rule is constructed from utility functions in terms of sponsors (Sp) and patients (Pat)

| Scenario | $HR_N$ | $\tilde{\eta}_F^{RMD}$ | | | $\tilde{\eta}_P^{RMD}$ | | |
|---|---|---|---|---|---|---|---|
| | | Sp and Pat | Sp | Pat | Sp and Pat | Sp | Pat |
| | | ($\delta_P = 1.0$) | ($\delta_P = 0.5$) | ($\delta_P = 0.5$) | ($\delta_P = 1.0$) | ($\delta_P = 0.5$) | ($\delta_P = 0.5$) |
| 1 | 0.50 | −5.38 | −5.38 | −5.38 | −4.50 | −4.50 | −4.50 |
| 2 | 0.90 | −1.25 | −1.50 | −1.50 | −6.25 | −6.10 | −6.25 |
| 3 | 1.00 | −0.75 | −1.25 | −1.25 | −2.50 | −2.50 | −2.50 |
| 4 | 1.43 | 0.00 | −0.25 | −0.25 | −2.50 | −2.50 | −2.50 |

$HR_N$ is the assumed hazard ratio for biomarker-negative ($N$) population; $HR_P$ is the assumed hazard ratio for $P$ and is set to 0.50; $\delta_P$ is the discount parameter for $P$; $\delta_N$ is equal to $\delta_P$ since we assume that the prevalence of $P$, $\psi$, is set to 50%.

Table 2.4: Results of optimal thresholds of interim ratio of restricted mean survival times, $\tilde{\eta}_g^{RMR}$, for $g \in \{F, P\}$, full ($F$) and biomarker-positive ($P$) populations, derived by grid search when the interim decision rule is constructed from utility functions in terms of sponsors (Sp) and patients (Pat)

| | | $\tilde{\eta}_F^{RMR}$ | | | $\tilde{\eta}_P^{RMR}$ | | |
|---|---|---|---|---|---|---|---|
| Scenario | $HR_N$ | Sp and Pat | Sp | Pat | Sp and Pat | Sp | Pat |
| | | ($\delta_P = 1.0$) | ($\delta_P = 0.5$) | ($\delta_P = 0.5$) | ($\delta_P = 1.0$) | ($\delta_P = 0.5$) | ($\delta_P = 0.5$) |
| 1 | 0.50 | 0.65 | 0.65 | 0.65 | 0.75 | 0.75 | 0.75 |
| 2 | 0.90 | 0.85 | 0.85 | 0.85 | 0.58 | 0.58 | 0.58 |
| 3 | 1.00 | 0.90 | 0.85 | 0.85 | 0.58 | 0.61 | 0.58 |
| 4 | 1.43 | 1.00 | 0.95 | 0.95 | 0.75 | 0.75 | 0.75 |

$HR_N$ is the assumed hazard ratio for biomarker-negative ($N$) population; $HR_P$ is the assumed hazard ratio for $P$ and is set to 0.50; $\delta_P$ is the discount parameter for $P$; $\delta_N$ is equal to $\delta_P$ since we assume that the prevalence of $P$, $\psi$, is set to 50%.

Second, Table 2.5 illustrates the probabilities of selecting each corresponding population at the interim analysis, where the interim decision rule was constructed in terms of sponsors and patients, respectively, and where $\delta_P$ is set to 1.0 and 0.5, respectively. For $\delta_P = \delta_N = 1.0$, comparable results were obtained regardless of the type of utility function for constructing the interim decision rule. The more diluted the treatment effect in $F$ was, the higher the probability of selecting only $P$ was. However, the probability of continuing $F$ was higher, when $\delta_P$ was set to 0.5, especially in Scenarios 2 and 3, regardless of the type of utility function, since the utility for $P$ was discounted considering the population size. Furthermore, with respect to the interim decision rule by either $\widehat{RMD}_g$ or $\widehat{RMR}_g$, the probability of continuing $F$ was less than that using either $\widehat{HR}_{I,g}$ or $\widehat{PP}_g$, notably in Scenarios 2 and 3. If the treatment effect in $P$ is weaker, this probability would be de-escalated, owing to the augmentation for the probability of stopping the trial for futility.

Finally, Table 2.6 represents the probabilities of rejecting each hypothesis at the final analysis, where the interim decision rule was constructed in terms of sponsors and patients, respectively, and where $\delta_P$ is set to 1.0 and 0.5, respectively. Note that the probability of rejecting the hypothesis $H_F \cup H_P$ shows the power. Prior to the description of Table 2.6, we confirmed that the familywise type I error rate was controlled at less than 2.5% across all scenarios, regardless of every interim measure and the type of utility function, where calculation did not include the possibility of stopping for futility at the interim analysis. Here, the probability of rejecting $H_F$ was greater than that of rejecting $H_P$ and comparable to the power notably in Scenario 1, whereas the probability of rejecting $H_P$ was greater than that of rejecting $H_F$ and comparable to the power when the treatment effect in $F$ was diluted in Scenarios 2 to 4. Similar to the results of the interim analysis, the probabilities of rejecting $H_F$ when $\delta_P$ was set to 0.5 were higher than those when $\delta_P$ was set to 1.0.

Table 2.5: Probabilities of each decision at the interim analysis based on utility functions in terms of sponsors (Sp) and patients (Pat) when each measure ($HR_I$, hazard ratio; $PP$, predictive power; $RMD$, difference of restricted mean survival times; $RMR$, ratio of restricted mean survival times) for full ($F$) and biomarker-positive ($P$) populations is used at the interim analysis

| Scenario | $HR_N$ | Measure | Sp and Pat ($\delta_P = 1.0$) | | | Sp ($\delta_P = 0.5$) | | | Pat ($\delta_P = 0.5$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\Omega \in F$ | $\Omega \in P$ | $\Omega \in \phi$ | $\Omega \in F$ | $\Omega \in P$ | $\Omega \in \phi$ | $\Omega \in F$ | $\Omega \in P$ | $\Omega \in \phi$ |
| 1 | 0.50 | $HR_I$ | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| | | $PP$ | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| | | $RMD$ | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| | | $RMR$ | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| 2 | 0.90 | $HR_I$ | 0.95 | 0.05 | 0.00 | 0.98 | 0.02 | 0.00 | 0.98 | 0.02 | 0.00 |
| | | $PP$ | 0.93 | 0.07 | 0.00 | 0.98 | 0.02 | 0.00 | 0.98 | 0.02 | 0.00 |
| | | $RMD$ | 0.97 | 0.03 | 0.00 | 0.99 | 0.01 | 0.00 | 0.99 | 0.01 | 0.00 |
| | | $RMR$ | 0.98 | 0.02 | 0.00 | 0.98 | 0.02 | 0.00 | 0.98 | 0.02 | 0.00 |
| 3 | 1.00 | $HR_I$ | 0.86 | 0.14 | 0.00 | 0.91 | 0.09 | 0.00 | 0.91 | 0.09 | 0.00 |
| | | $PP$ | 0.84 | 0.16 | 0.00 | 0.90 | 0.10 | 0.00 | 0.90 | 0.10 | 0.00 |
| | | $RMD$ | 0.83 | 0.17 | 0.00 | 0.95 | 0.05 | 0.00 | 0.95 | 0.05 | 0.00 |
| | | $RMR$ | 0.86 | 0.14 | 0.00 | 0.96 | 0.05 | 0.00 | 0.96 | 0.05 | 0.00 |
| 4 | 1.43 | $HR_I$ | 0.09 | 0.91 | 0.00 | 0.15 | 0.85 | 0.00 | 0.15 | 0.85 | 0.00 |
| | | $PP$ | 0.08 | 0.91 | 0.01 | 0.13 | 0.87 | 0.01 | 0.13 | 0.87 | 0.01 |
| | | $RMD$ | 0.13 | 0.87 | 0.00 | 0.21 | 0.79 | 0.00 | 0.21 | 0.79 | 0.00 |
| | | $RMR$ | 0.12 | 0.88 | 0.00 | 0.27 | 0.73 | 0.00 | 0.27 | 0.73 | 0.00 |

$HR_N$ is the assumed hazard ratio for biomarker-negative ($N$) population; $HR_P$ is the assumed hazard ratio for $P$ and is set to 0.50; $\Omega$ is the selected population determined based on the interim analysis; $\phi$ indicates that neither $F$ nor $P$ is selected; $\delta_P$ is the discount parameter for $P$; $\delta_N$ is equal to $\delta_P$ since we assume that the prevalence of $P$, $\psi$, is set to 50%.

Table 2.6: Probabilities of rejecting the hypothesis $H_F$ ($r_F = 1$), $H_P$ ($r_P = 1$), or $H_F \cup H_P$ $\{(r_F = 1) \cup (r_P = 1)\}$ at the final analysis, when each measure ($HR_I$, hazard ratio; $PP$, predictive power; $RMD$, difference of restricted mean survival times; $RMR$, ratio of restricted mean survival times) for full ($F$) and biomarker-positive ($P$) populations is used at the interim analysis and when the interim decision rule is constructed from utility functions in terms of sponsors (Sp) and patients (Pat)

| | | | Sp and Pat ($\delta_P = 1.0$) | | | Sp ($\delta_P = 0.5$) | | | Pat ($\delta_P = 0.5$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | $HR_N$ | Measure | $r_F = 1$ | $r_P = 1$ | $(r_F = 1)$ | $r_F = 1$ | $r_P = 1$ | $(r_F = 1)$ | $r_F = 1$ | $r_P = 1$ | $(r_F = 1)$ |
| | | | | | $\cup (r_P = 1)$ | | | $\cup (r_P = 1)$ | | | $\cup (r_P = 1)$ |
| 1 | 0.50 | $HR_I$ | 1.00 | 0.98 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | 0.98 | 1.00 |
| | | $PP$ | 1.00 | 0.98 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | 0.98 | 1.00 |
| | | $RMD$ | 1.00 | 0.98 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | 0.98 | 1.00 |
| | | $RMR$ | 1.00 | 0.98 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | 0.98 | 1.00 |
| 2 | 0.90 | $HR_I$ | 0.90 | 0.97 | 0.98 | 0.92 | 0.97 | 0.98 | 0.92 | 0.97 | 0.98 |
| | | $PP$ | 0.88 | 0.97 | 0.98 | 0.91 | 0.97 | 0.98 | 0.91 | 0.97 | 0.98 |
| | | $RMD$ | 0.91 | 0.97 | 0.98 | 0.92 | 0.97 | 0.98 | 0.92 | 0.97 | 0.98 |
| | | $RMR$ | 0.91 | 0.97 | 0.98 | 0.91 | 0.97 | 0.98 | 0.91 | 0.97 | 0.98 |
| 3 | 1.00 | $HR_I$ | 0.77 | 0.97 | 0.97 | 0.80 | 0.97 | 0.97 | 0.80 | 0.97 | 0.97 |
| | | $PP$ | 0.76 | 0.97 | 0.97 | 0.79 | 0.97 | 0.97 | 0.79 | 0.97 | 0.97 |
| | | $RMD$ | 0.75 | 0.97 | 0.97 | 0.81 | 0.97 | 0.97 | 0.81 | 0.97 | 0.97 |
| | | $RMR$ | 0.77 | 0.97 | 0.97 | 0.82 | 0.97 | 0.97 | 0.82 | 0.97 | 0.97 |
| 4 | 1.43 | $HR_I$ | 0.07 | 0.96 | 0.96 | 0.11 | 0.96 | 0.96 | 0.11 | 0.96 | 0.96 |
| | | $PP$ | 0.07 | 0.96 | 0.96 | 0.10 | 0.96 | 0.96 | 0.10 | 0.96 | 0.96 |
| | | $RMD$ | 0.10 | 0.96 | 0.96 | 0.14 | 0.96 | 0.96 | 0.14 | 0.96 | 0.96 |
| | | $RMR$ | 0.10 | 0.96 | 0.96 | 0.17 | 0.96 | 0.96 | 0.17 | 0.96 | 0.96 |

$HR_N$ is the assumed hazard ratio for biomarker-negative ($N$) population; $HR_P$ is the assumed hazard ratio for $P$ and is set to 0.50; $\delta_P$ is the discount parameter for $P$; $\delta_N$ is equal to $\delta_P$ since we assume that the prevalence of $P$, $\psi$, is set to 50%; $r_g$ is the indicator variable such that $r_g = 1$, if the hypothesis $H_g$ for population $g \in \{F, P\}$ is rejected at the final analysis.

## 2.6 Summary and discussion

In this chapter, we proposed a novel approach to set the thresholds within the interim decision rule in the context of adaptive population selection design. The proposed approach used gain and loss functions as utility functions in terms of sponsors and patients, solely to assist in formulating the interim decision rule to determine whether the entire population is continued or only the promising population is selected.

Simulation results revealed that the proposed approach can guide the setting of thresholds considering the benefits and risks for sponsors and patients. Under the situation assumed at the planning stage, this can minimize the possibility of terminating the appropriate population at the interim analysis, regardless of the scenario in which either hypothesis can be rejected, and the possibility of continuing the inappropriate population to the following stage even though neither hypothesis can be rejected. We further considered the population size that represents the market size for sponsors and the number of patients who are benefited or harmed by the targeted therapy by incorporating the discount parameters into the utility functions.

In our simulation study, we considered that the targeted therapy is beneficial only in the biomarker-positive population, but is harmful in the biomarker-negative population. In this setting, the treatment effect for the full population is diluted. Note that we would hardly apply the setting for the thresholds when the therapy is considerably harmful, i.e., Scenario 4 in Section 2.5.1, since the optimal thresholds based on pseudo clinical trial results via simulation were critically searched to be strict, and the interim thresholds were close to the boundaries at the final analysis, even if we employed the discount parameters for the biomarker-positive population that were less than one. In implementing the clinical trial that also includes the biomarker-negative population with belief that the therapy might be harmful for that population, the assumption that it is not beneficial rather than harmful, i.e., Scenarios 2 and 3 in Section 2.5.1, would be plausible. Hence, setting thresholds must be carefully considered on the basis of the results of using the proposed approach, as Gallo et al. (2014) pointed out that the thresholds for stopping the trial for futility should be viewed as guidelines, rather than rules, in the context of group sequential designs. Moreover, in this chapter, we solely considered the scenario in which the therapy is unalterably beneficial in $P$ and the sample size to reject the null hypothesis for $P$ is sufficient. If an insufficient sample size is considered, the strict thresholds for the interim decision rule would be obtained. These results would guide whether the sample size at interim or final analysis is sufficient or not when considering the utility in terms of sponsors or patients, although, in this case, the thresholds should not be applied without doubt.

With respect to the use of the utility functions in the development of molecular targeted therapies, Graf et al. (2015) used the utility functions in terms of the views of sponsors and public health to compare the fixed design, enrichment design, and adaptive population selection design. Krisam and Kieser (2014) used a utility function to construct the decision rule, even in the situation in which the population is considered on the basis of an imperfect biomarker. On the other hand, we incorporated utility functions constructed from gain and loss functions into the method for setting the thresholds within the interim decision. Moreover, with regard to the decision making based on the utility functions, this would be controversial to apply to the final decision making of a clinical trial. However, we solely incorporate the utility functions into only the interim decision for selecting the population in order to assist in setting the thresholds. As far as it relates to the interim decision, as possible future work, the proposed approach would be extended to the use of informative priors herein if any knowledge were obtained preceding the planning for the trial. Note that we do not encourage the use of utility functions for the decision of the final analysis.

# Chapter 3

# Interim decision-making strategies in adaptive designs for population selection

## 3.1 Introduction

In the development of molecular targeted therapies for progressive cancer patients, both progression-free survival (PFS) and overall survival (OS) have been shown in clinical trial reports. For instance, both OS and PFS significantly improve in patients who have advanced colorectal cancer, receive cetuximab, and have the wild-type K-ras gene, whereas no improvement is seen in patients with mutated K-ras tumors (Karapetis et al., 2008). Several authors have discussed the issues involved in using PFS (Booth and Eisenhauer, 2012; Fleming et al., 2009; Kay et al., 2011; Saad et al., 2010). Figure 3.1 describes the definition of OS and PFS. PFS is defined as the minimum time from randomization until tumor progression or death from any cause, namely, time-to-progression (TTP) or OS, while PPS is defined as the time from progression to death (Food and Drug Administration, 2007; Saad and Buyse, 2012). Although OS is the most commonly used endpoint required in phase III trials by regulatory agencies, PFS is also frequently used in phase II trials, especially those conducted for evaluation of molecular targeted therapies. However, OS generally requires a long follow-up period after tumor progression. Therefore, a long study and a large number of patients are required, making it expensive to conduct clinical trials using OS as the primary endpoint measure.

This chapter also investigates an adaptive design for population selection when using correlated survival endpoints. Although we did not clarify any specific survival endpoints in Chapter 2, we handle both OS and PFS in this chapter. Brannath et al. (2009) presented an adaptive design method for population selection by using a single survival endpoint. Following this, Jenkins et al. (2011) proposed a similar method using correlated survival endpoints; OS was used as the final outcome and PFS as a short-term endpoint. Friede et al. (2012) then demonstrated a more powerful method for a final analysis by using a conditional error function approach (Müller and Schäfer, 2001). Subsequently, Stallard et al. (2014) compared the method of Friede et al. (2012) with that of Jenkins et al. (2011) so as to investigate the properties of the adaptive population selection design methods.

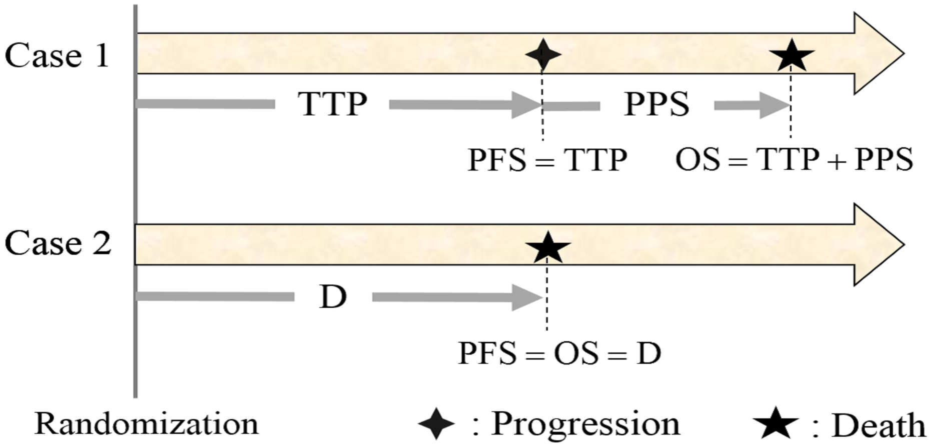

Figure 3.1: Definition of OS and PFS

As shown in Figure 3.2, the interim analysis involves the use of PFS data from stage 1 patients only, whereas the final analysis is conducted based on OS data from each stage, assuming that OS is set as the primary endpoint at the end of stage 2 (Jenkins et al., 2011). However, the method requires the strong assumption that OS and PFS are strongly correlated. For instance, PFS significantly improved in patients who had pulmonary adenocarcinoma, tested positive for epidermal growth factor receptor (EGFR) inhibitor, and received gefitinib, whereas no improvement was seen in OS (Mok et al., 2009). The inconsistent results are derived from OS traits in which OS generally requires long-term follow-up to accumulate events because it can be affected by post-progression survival (PPS) magnitudes.

In this chapter, we propose an interim decision-making strategy in adaptive designs for population selection. We extend the previous methods (Brannath et al., 2009; Jenkins et al., 2011) in two aspects. First, the interim analysis is conducted by incorporating information on PFS as well as OS. Second, we consider a scenario in which OS is calculated based on PPS, if the progression is observed before death. The combination test approach will be applied with respect to final decision-making in a manner similar to Section 2 in Chapter 2. We use the weighted inverse normal combination method for the OS data since OS is a primary endpoint.

The rest of this chapter is structured as follows. In Section 3.2, we discuss the interim decision-making strategies using correlated survival endpoints. Section 3.3 presents a simulation study. Finally, Section 3.4 provides concluding remarks.

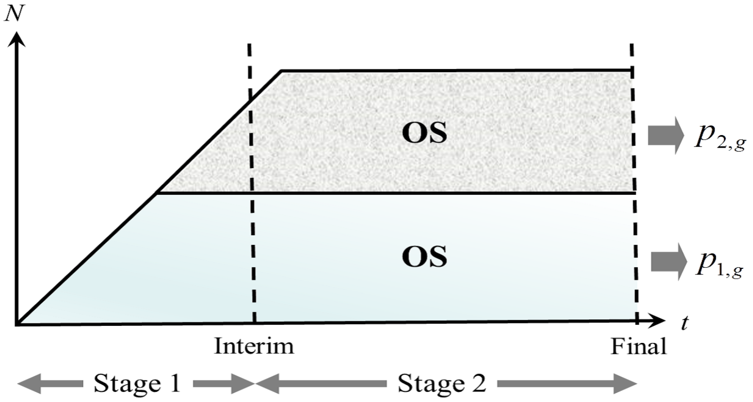

Figure 3.2: Schematic representation of the p-values from each stage

## 3.2 Interim decision-making

The aim of study is to extend the recent methods (Brannath et al., 2009; Jenkins et al., 2011) for interim decision-making in two aspects. First, both OS data and PFS data are incorporated into the interim analysis in the adaptive population selection design. Generally, OS requires a long follow-up period after tumor progression and a long study and a large number of patients, and results in making it expensive to conduct clinical trials using OS as the primary endpoint measure. Hence, it would be practical to use PFS data for interim decision-making when deemed as a short-term intermediate survival endpoint, because it can be observed in a shorter period of time than OS. Furthermore, it would also be pragmatic to consider the impact based on PPS data; therefore, we assume a scenario in which OS is calculated considering PPS after tumor progression, if progression is observed before death.

### 3.2.1 Procedures for interim decision-making

An interim analysis is conducted to identify whether the full population $F$ or only the predefined biomarker-positive population $P$ would benefit from a given treatment based on stage 1 only. In addition, interim analysis can be used to determine whether or not it is worth continuing a clinical trial; the trial can be discontinued early only for futility, when interim analysis deems that the success of the trial is unpromising. This enables sponsors and investigators to optimize the investment of resources. For interim decisions, the sponsor has to be blinded to any results at the interim stage and the Independent Data Monitoring Committee (IDMC) makes the recommendation based on an interim decision rule.

### 3.2.2 Simple exponential model for dependence between OS and PFS

Among several methods available to measure the correlation between time-to-event variables such as OS and PFS, we use the statistical models developed by Fleischer et al. (2009) to account for correlation with censoring. Fleischer et al. (2009) have demonstrated some possible applications for the proposed models. According to the example from a phase II trial with two arms for patients with non-small-cell lung cancer (NSCLC), the estimated survival function for OS based on the data given versus the predicted survival function for OS based on the proposed models show quite a good agreement with 95 patients, particularly in the period until 3 months.

Let D denote the survival endpoint, i.e., the time to death without tumor progression. They used exponential models for each survival endpoint based on the assumption that TTP and D are completely independent and that PFS is given by the minimum values of TTP and OS. Then, OS is calculated as follows:

$$\mathrm{OS} = \begin{cases} \mathrm{PFS} & if \quad \mathrm{PFS} \neq \mathrm{TTP} \\ \mathrm{TTP} + \mathrm{PPS} & otherwise \end{cases} .$$

Suppose that each survival endpoint $a \in \{\mathrm{TTP}, \mathrm{PPS}, \mathrm{D}\}$ is exponentially distributed with parameter $\lambda_a$ where $\lambda_a > 0$. Then, the Pearson correlation coefficient between OS and PFS is given by

$$\rho = \mathrm{Corr}(\mathrm{OS}, \mathrm{PFS}) = \frac{\lambda_{\mathrm{PPS}}}{\sqrt{\lambda_{\mathrm{TTP}}^2 + 2\lambda_{\mathrm{TTP}}\lambda_{\mathrm{D}} + \lambda_{\mathrm{PPS}}^2}}. \tag{3.1}$$

The derivation of Equation (3.1) is provided in Appendix A. Note that $\rho = 1.0$ if no tumor progression occurs before death; in other words, PFS = OS.

### 3.2.3 The illness-death model for semi-competing risks data

The model proposed by Fleischer et al. (2009) for each survival endpoint $a \in \{\mathrm{TTP}, \mathrm{PPS}, \mathrm{D}\}$ assumes an exponential distribution, which could be simplified regardless of the complicated outcomes between survival endpoints. In addition, it is necessary to consider that $\lambda_a$ is limited owing to $\lambda_a > 0$; for example, when $\lambda_{\mathrm{TTP}} = 0.347$ and $\rho = 0.5$, $\lambda_{\mathrm{PPS}} \gtrsim 0.200$ is required to satisfy $\lambda_{\mathrm{D}} > 0$ based on Equation (3.1). Thus, we further consider a semi-competing risks framework (Fine et al., 2001). Each endpoint, TTP and D, is considered to be a nonterminal and terminal event, respectively. On the upper wedge, i.e., TTP $\leq$ D, Fine et al. (2001) describe the copula model developed by Clayton (1978) with the joint survival function expressed as

$$S(\mathrm{TTP}, \mathrm{D}) = (S_{\mathrm{TTP}}^{-\theta} + S_{\mathrm{D}}^{-\theta} - 1)^{-1/\theta}, \tag{3.2}$$

where $S_{\mathrm{TTP}}$ and $S_{\mathrm{D}}$ denote the marginal survival function for TTP and D, respectively, and $\theta \geq 0$ represents a parameter measuring the correlation. The parameter $\theta$ is approximately related to Kendall's $\tau$ as $\tau = \theta/(\theta + 2)$.

Similarly, the approach by Xu et al. (2010) considers a class of illness-death models, as shown in Figure 3.3, with a shared frailty as follows:

$$\lambda_{\mathrm{TTP}|\gamma} = \gamma\lambda_{0,\mathrm{TTP}},$$
$$\lambda_{\mathrm{D}|\gamma} = \gamma\lambda_{0,\mathrm{D}},$$
$$\lambda_{\mathrm{PPS}|\gamma} = \gamma\lambda_{0,\mathrm{PPS}},$$

where $\lambda_{0,a}$ is a baseline hazard for $a \in \{\text{TTP}, \text{PPS}, \text{D}\}$ and $\gamma$ denotes the gamma frailty with mean 1 and variance $1/\theta$. Under the assumption $\lambda_{0,\text{PPS}} = \lambda_{0,\text{D}}$, the restricted illness-death model is intrinsically equivalent to the semi-competing risks framework.

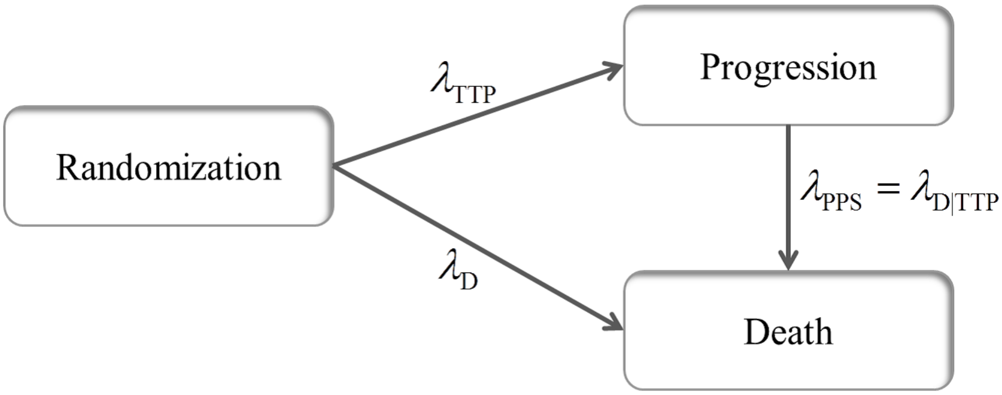

Figure 3.3: Illustration of illness-death models

### 3.2.4 Interim decision rule using predictive power

In this study, the decision tool applied at the interim analysis relies on the predictive power approach. Assume non-informative priors for the treatment effect and the interim $e \in \{\text{OS}, \text{PFS}\}$ data can be obtained. As an extension of the predictive power as described in Equation (2.1) in Chapter 2, the predictive power, $PP_l^{\{e\}}$, for each population $l \in \{F, P, N\}$ is given by

$$PP_l^{\{e\}} = 1 - \Phi\left[\left(1 - \Pi_l^{\{e\}}\right)^{-1/2}\left(\Phi^{-1}(1-\alpha)\cdot\sqrt{\Pi_l^{\{e\}}} - z_l^{\{e\}}\right)\right]$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, $\Pi_l^{\{e\}}$ represents the event fraction at the interim analysis, $\alpha$ is a one-sided significance level at the final analysis, and $z_l^{\{e\}}$ is the observed test statistic based on the log-rank test using stage 1 data. In contrast to Equation (2.1) in Chapter 2, we consider two endpoints: OS and PFS. Further we evaluate the predictive power for $N$.

In the context of adaptive population selection designs, Brannath et al. (2009) have demonstrated the decision rule by using the predictive power at the interim analysis, whereas Jenkins et al. (2011) have proposed the use of the rule based on the estimated hazard ratios. For the clinical development of molecular targeted therapies, when we consider a scenario in which the experimental treatment is beneficial for $P$ but is actually harmful for $N$, the problem of crossing hazard rates might be observed due to the violation of the proportional hazard assumption. Hence, we provide an interim decision rule using predictive power by extending the rule demonstrated by Brannath et al. (2009) for a single endpoint, and by considering the use of multiple endpoints at the interim analysis. The following decision rule is defined for the interim stage:

$$\text{(Case 1):} \quad \bigcap_{l=F,P,N,e=\text{OS},\text{PFS}} PP_l^{\{e\}} > \kappa_e \cdot \eta_{1,l} \quad \rightarrow \quad \text{Go with } F$$

$$\text{(Case 2):} \quad \text{if Case 1 is not met,} \quad \bigcap_{e=\text{OS},\text{PFS}} PP_P^{\{e\}} > \kappa_e \cdot \eta_{2,P} \quad \rightarrow \quad \text{Go with } P$$

$$\text{(Case 3):} \quad \text{if both Case 1 and Case 2 are not met,} \quad \text{Otherwise.} \quad \rightarrow \quad \text{Stop for futility}$$

where $\eta_{i,l}$ denotes the threshold of the predictive power for each population $l \in \{F, P, N\}$ in Case $i$, and $\kappa_e$ denotes the relative importance assigned to the corresponding endpoint $e \in \{\text{OS}, \text{PFS}\}$. Because of the nature that PFS data are more quickly observed than OS data, we should reflect the expected number of events accrued up to the interim analysis. For adaptive treatment selection design, Di Scala and Glimm (2011) combined the predictive probabilities using weights similar to $\kappa_e$, such as $\kappa_{\text{OS}} \cdot PP_l^{\{\text{OS}\}} + \kappa_{\text{PFS}} \cdot PP_l^{\{\text{PFS}\}}$, and have shown the simulation results. However, we evaluate multiple endpoints, i.e., OS and PFS, separately in considering the inconsistency as well as the correlation between OS and PFS due to the impact of PPS. To justify the heuristic decision rule, we consider the development of a targeted therapy: for example, when $(PP_F^{\{\text{PFS}\}}, PP_P^{\{\text{PFS}\}}, PP_N^{\{\text{PFS}\}}) = (0.94, 0.99, 0.11)$, $\kappa_{\text{PFS}} = 1$, and $(\eta_{1,F}, \eta_{1,P}, \eta_{1,N}, \eta_{2,P}) = (0.10, 0.05, 0.05, 0.20)$, $F$ is selected at the interim, whereas when $(PP_F^{\{\text{OS}\}}, PP_P^{\{\text{OS}\}}, PP_N^{\{\text{OS}\}}) = (0.37, 0.98, 0.01)$ and $\kappa_{\text{OS}} = 1/2$, $F$ is not selected, because $PP_N^{\{\text{OS}\}} > \kappa_{\text{OS}} \cdot \eta_{1,N}$ is not met, and $P$ is selected. This numerical example could be observed when the correlation between OS and PFS is not very strong. In addition, when $PP_N^{\{\text{PFS}\}}$ and $PP_N^{\{\text{OS}\}}$ are combined, as in the method of Di Scala and Glimm (2011), $F$ is selected despite the targeted therapy. In practice, a series of simulations is needed to set the thresholds to be used by the IDMC.

## 3.3 Simulation study

In this section, we describe a simulation study to show the operating characteristics of adaptive designs for population selection based on the interim decision-making strategies presented in Section 3.2. The design assumptions and simulation setting are presented in Section 3.3.1. Furthermore, each probability at the interim or the final decision is shown in Section 3.3.2.

### 3.3.1 Design assumptions and simulation setting

As with the simulation study in Chapter 2, we shall consider a randomized parallel-group clinical trial with two arms: experimental ($E$) and control ($C$). Assume that patients classified as either biomarker-positive ($P$) or biomarker-negative ($N$) are included in the trial in order to consider the efficacy of a targeted therapy. A single one-sided null hypothesis $H_g$ is assumed for each population $g \in \{F, P\}$. The prevalence of $P$ among $F$, $\psi$, is set to 50%. Let $HR_P^{\{a\}}$ and $HR_N^{\{a\}}$ denote the hazard ratio using the interim $a \in \{\text{TTP}, \text{PPS}, \text{D}\}$ data for $P$ and $N$, respectively. We consider several scenarios for the treatment effect:

$$\begin{aligned}
&(\text{Scenario 1}): \quad HR_P^{\{a\}} = 0.50 \ \text{ and } \ HR_N^{\{a\}} = 0.90\\
&(\text{Scenario 2}): \quad HR_P^{\{a\}} = 0.50 \ \text{ and } \ HR_N^{\{a\}} = 1.00\\
&(\text{Scenario 3}): \quad HR_P^{\{a\}} = 0.50 \ \text{ and } \ HR_N^{\{a\}} = 1.11\\
&(\text{Scenario 4}): \quad HR_P^{\{a\}} = 0.50 \ \text{ and } \ HR_N^{\{a\}} = 1.43
\end{aligned}$$

where a hazard ratio less than 1 indicates an increased benefit from $E$. Scenarios 1 to 4 represent those in which the experimental treatment is extremely beneficial for $P$, i.e., $HR_P^{\{a\}} = 0.50$, and the hazard ratios for $P$ for TTP, PPS, and D are considered to be similar for the sake of simplicity. In Scenario 1, $E$ is more beneficial for $P$ than it is for $N$. Scenario 2 is the scenario in which $E$ is beneficial for $P$ but not for $N$. In Scenarios 3 and 4, $E$ is beneficial for $P$ but is actually harmful for $N$. In particular, Scenario 4 is that in which $P$ is notably counterproductive for $N$. Note that Scenarios 1 to 4 are roughly set based on the motivating examples outlined in Table 3.1. For instance, Scenario 2 above is roughly based on an actual trial conducted for patients with advanced colorectal cancer that were receiving cetuximab (Karapetis et al., 2008). The result showed that the hazard ratio among patients with the wild-type K-ras gene was $HR_P^{\{\text{OS}\}} = 0.55$, whereas the hazard ratio among patients with mutated K-ras tumors was $HR_N^{\{\text{OS}\}} = 0.98$. In calculating the predictive power in Section 3.2.4, it is necessary to take into account $N$ as well as $F$ and $P$ in these scenarios. Furthermore, the hazard ratio for $F$ is considered as $HR_F^{\{a\}} = \exp\{\psi \cdot \log HR_P^{\{a\}} + (1-\psi) \cdot \log HR_N^{\{a\}}\}$.

The clinical trial consists of two stages with an interim analysis. We assume that the final analysis is performed after 300 OS events occur in reference to the example (Karapetis et al., 2008). An interim analysis is conducted after OS events reach 50% of the pre-planned OS events. Here, we assume that the overall number of patients is 400.

Table 3.1: Motivating examples for the scenarios assumed in the simulation

| Scenario | Reference | Definition of each population | Hazard ratio |
|---|---|---|---|
| 1 | Gregorc et al. (2014) | Patients with non-small-cell lung cancer classified by the proteomic test | $HR_P^{\{OS\}} = 0.58$, $HR_N^{\{OS\}} = 0.94$ |
| 2 | Karapetis et al. (2008) | Patients with advanced colorectal cancer classified by the K-ras gene | $HR_P^{\{OS\}} = 0.55$, $HR_N^{\{OS\}} = 0.98$ |
| 3 | Satoh et al. (2014) | Patients with HER2-amplified advanced gastric cancer classified by immunohistochemistry | $HR_P^{\{OS\}} = 0.59$, $HR_N^{\{OS\}} = 1.07$ |
| 4 | Mok et al. (2009) | Patients with advanced pulmonary adenocarcinoma classified by EGFR mutation status | $HR_P^{\{OS\}} = 0.78$, $HR_N^{\{OS\}} = 1.38$ |

*P*, biomarker-positive population; *N*, biomarker-negative population.

In terms of interim decision-making, the thresholds that needed to be pre-specified at the planning stage are roughly set as $(\eta_{1,F}, \eta_{1,P}, \eta_{1,N}, \eta_{2,P}) = (0.10, 0.05, 0.05, 0.20)$ with weights of $\kappa_{\mathrm{OS}} = 1/2$ and $\kappa_{\mathrm{PFS}} = 1$, respectively, because OS is a primary endpoint whereas PFS data are more rapidly observed than OS data.

With respect to data generation within the simulation, we used the models from Equations (3.1) and (3.2), respectively. Assume that the median TTP for $C$ is 2 months. Then, let us consider the following four models (a) to (d): (a) Equation (3.1), where the correlation between OS and PFS is $\rho = (0.9, 0.7, 0.5)$, and the corresponding median PPS for $C$ is $(0.5, 0.5, 0.5)$ or $(0.5, 2.0, 3.0)$ months for small or large effects of PPS, respectively. (b−d) Equation (3.2), where the correlation of TTP and D is $\tau = (0.9, 0.7, 0.5)$, and the corresponding median PPS for $C$ is $(0.5, 0.5, 0.5)$ or $(1.0, 5.0, 10.0)$ months for small or even larger effects of PPS, respectively, using (b) an exponential baseline hazard distribution ($\lambda_{0,a} = \log(2)/median(a)$) or (c, d) a Weibull baseline hazard distribution ($\lambda_{0,a} = \log(2)/(median(a))^{\nu}$) with the shape parameter $\nu$, where (c) $\nu = 0.5$ and (d) $\nu = 2.0$, for $a \in \{\mathrm{TTP}, \mathrm{PPS}, \mathrm{D}\}$. Note that $\tau$ is set for reference only and does not correspond to $\rho$.

Representative results for the simulation are presented in Figures 3.4 to 3.7. Details of the simulation results are shown in Tables B.1 to B.5 of Appendix B. Furthermore, the comparison of approaches that use OS data only or PFS data only at the interim analysis is also given.

### 3.3.2 Simulation results

First, we confirmed that the familywise type I error rate is controlled at less than 2.5% across all scenarios, regardless of the model type, based on the probabilities of rejecting at least one null hypothesis for $F$ or $P$. Calculation of the familywise type I error rate did not include the possibility of stopping for futility in the interim decision rule. The assumptions of the Simes' procedure and the weighted inverse normal combination method were met since the correlation of p-values between $F$ and $P$ was positive and the independence between stage 1 and stage 2 was also confirmed under all scenarios. All subsequent results were obtained based on 10,000 simulation replications per scenario.

Figures 3.4 and 3.5 show the probabilities of selecting each corresponding population at the interim analysis outlined in Scenario 2. A lower probability is better with respect to the probabilities of selecting $F$ in Figure 3.4, whereas a higher probability is better with respect to the probabilities of selecting $P$ in Figure 3.5, because of the simulation setting in which the targeted therapy is considered. With model (a), a greater probability of selecting each population is expected when using both OS and PFS under the assumption that the effect of PPS is considerable, particularly when the correlation between OS and PFS is not very strong, i.e., $\rho = (0.7, 0.5)$. Likewise, greater probabilities were also observed when evaluating model (b), under the assumption that the effect of PPS is even more considerable. These points indicate that, rather than the correlation between OS and PFS, PPS has a greater effect on misspecification of the population. On the other hand, each probability is similar under the assumption that the effect of PPS is small. These insights can also be seen for other scenarios as illustrated in Appendix B, regardless of the model type, especially in Scenarios 1 to 3. In addition, the probability of discontinuing the trial early owing to futility was less than 10% under every scenario outlined in Table A.3. Although this would be valid in terms of the targeted therapy setting for simulations, it nonetheless depends on the pre-specified thresholds.

Finally, Figures 3.6 and 3.7 show the probabilities of rejecting the null hypothesis, i.e., power, for each population at the final analysis in Scenario 2. The weighted inverse normal combination method was used for this calculation. Note that lower power is better with respect to the probabilities of selecting $F$ in Figure 3.6, because this indicates a benefit only for $P$ under every scenario considered herein. Therefore, incorporation of information for both OS and PFS results in good performance with no dependence on the type of models employed, as illustrated in Appendix B.

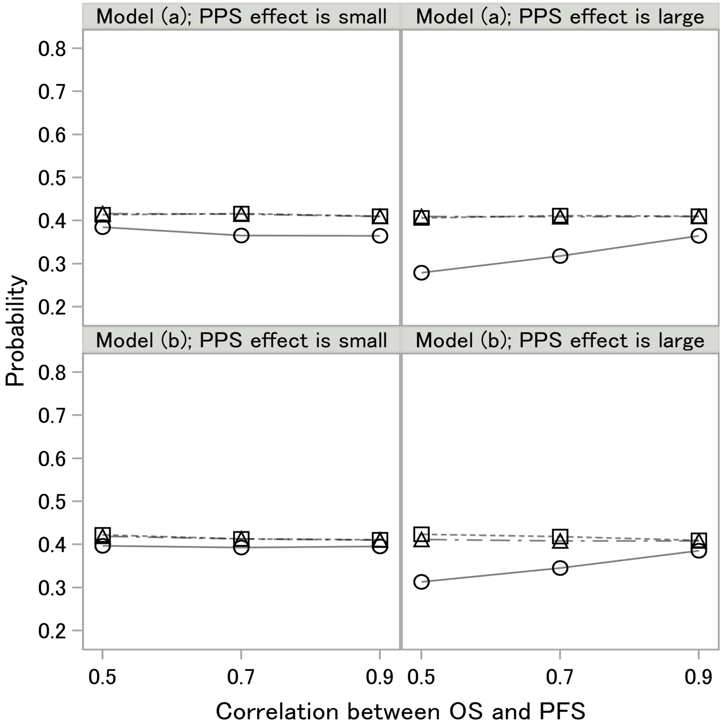

Figure 3.4: Probabilities of selecting $F$ at the interim analysis using PFS only (squares), OS only (triangles), or OS and PFS (circles) under the assumption that the effect of PPS is small [median PPS for $C$ is $(0.5, 0.5, 0.5)$ months (left panel)] and large [median PPS for $C$ is $(0.5, 2.0, 3.0)$ (upper-right panel) or $(1.0, 5.0, 10.0)$ (lower-right panel) months] based on 10,000 simulation replications in Scenario 2.

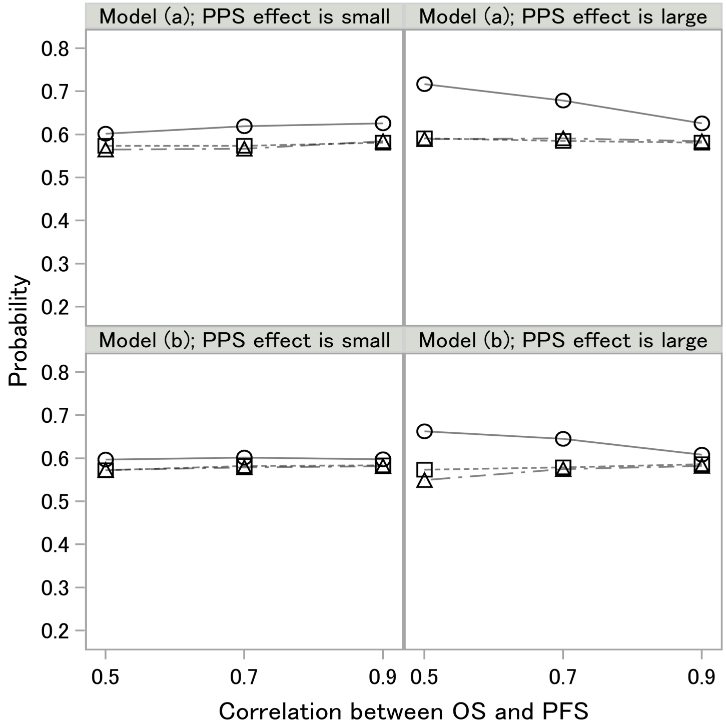

Figure 3.5: Probabilities of selecting *P* at the interim analysis using PFS only (squares), OS only (triangles), or OS and PFS (circles) under the assumption that the effect of PPS is small [median PPS for *C* is $(0.5, 0.5, 0.5)$ months (left panel)] and large [median PPS for *C* is $(0.5, 2.0, 3.0)$ (upper-right panel) or $(1.0, 5.0, 10.0)$ (lower-right panel) months] based on 10,000 simulation replications in Scenario 2.

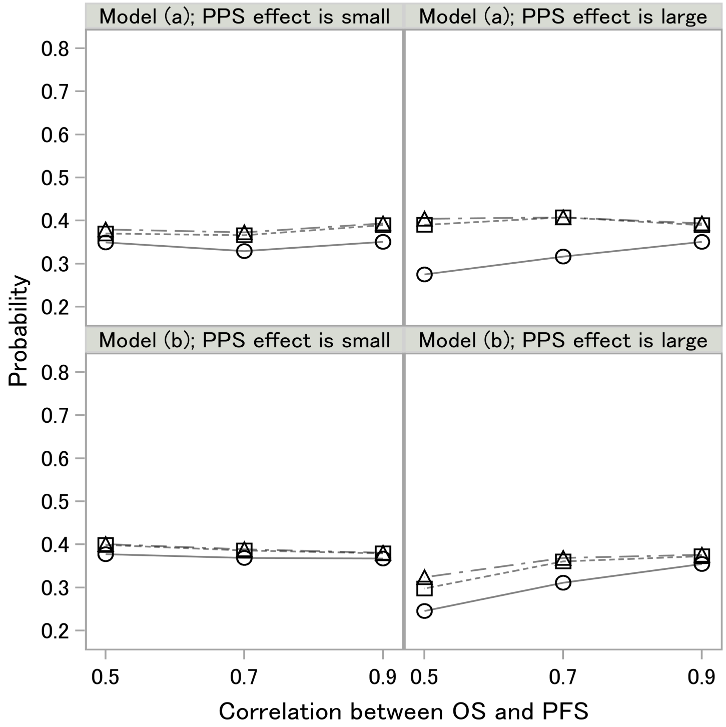

Figure 3.6: Probabilities of rejecting $H_F$ at the final analysis using PFS only (squares), OS only (triangles), or OS and PFS (circles) at the interim analysis under the assumption that the effect of PPS is small [median PPS for $C$ is $(0.5, 0.5, 0.5)$ months (left panel)] and large [median PPS for $C$ is $(0.5, 2.0, 3.0)$ (upper-right panel) or $(1.0, 5.0, 10.0)$ (lower-right panel) months] based on 10,000 simulation replications in Scenario 2.

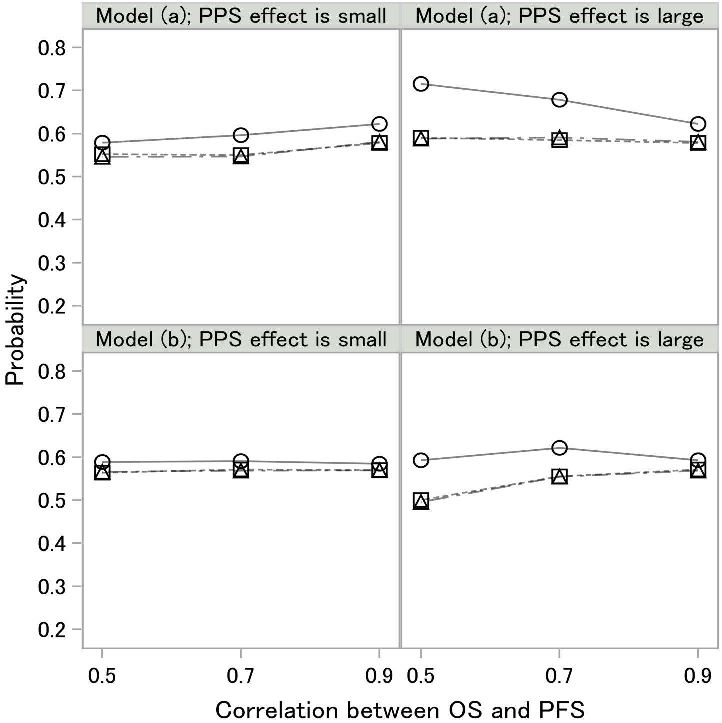

Figure 3.7: Probabilities of rejecting $H_P$ at the final analysis using PFS only (squares), OS only (triangles), or OS and PFS (circles) at the interim analysis under the assumption that the effect of PPS is small [median PPS for $C$ is $(0.5, 0.5, 0.5)$ months (left panel)] and large [median PPS for $C$ is $(0.5, 2.0, 3.0)$ (upper-right panel) or $(1.0, 5.0, 10.0)$ (lower-right panel) months] based on 10,000 simulation replications in Scenario 2.

## 3.4 Summary and discussion

The use of interim decision-making strategies in adaptive designs shows good performance for selecting the most appropriate population in developing molecular targeted therapies when using OS as well as PFS. This is particularly relevant under a scenario in which PPS affects the correlation between OS and PFS, within the context of a simple exponential model. Although this model is relatively simple because it assumes an exponential distribution in spite of complicated outcomes between survival endpoints, models developed under a semi-competing risks framework within the illness-death model and those assuming a Weibull distribution also performed well. A restriction in our simulation results is that sample size calculations were not considered for the sake of simplicity. As shown by Boessen et al. (2013), sample size estimation is required for an adaptive population selection design. In practice, this is determined based on the expected treatment effect for both $F$ and $P$.

In oncology, whether or not the use of PFS instead of OS is acceptable as a primary endpoint in a given treatment evaluation for marketing approval will depend on the specific disease setting. Nevertheless, there are many situations for which a regulatory agency will require the use of OS as a primary endpoint by regulatory agencies. Our procedure could also be applied in these settings. Furthermore, we considered the length of PPS when using the correlation between OS and PFS. Consequently, the probability of selecting the proper population at the interim analysis was improved for situations in which a relatively long PPS is expected. However, as Zhang et al. (2013) have mentioned, the OS benefit, given the PFS benefit, also largely depends on the crossover rate, because treatment crossover from the control to the experimental group frequently occurs after tumor progression in real trials. Regarding the correlation of multiple survival endpoints, it would be worthwhile to consider the effect of crossover rates in addition to the magnitude of PPS. In addition, to overcome the problem of influencing the stage 1 test statistic via adaptation at the interim analysis, we handled the additional follow-up during stage 2 for patients that are accrued in stage 1 by contributing stage 1 p-values, as suggested by Jenkins et al. (2011). Although both interim progression and death are regarded as predictive events for future death, and this information is used to deliberately affect the number of events observed in stage 1 patients (Bauer and Posch, 2004), the sponsor is blinded to any interim results and instead receives the recommendation for the interim decisions made by the IDMC. Hence, our approach of using both OS and PFS would be valid under a situation in which there is no sample size modification at the interim.

# Chapter 4

# State of the art: Adaptive seamless design in developing biosimilars

## 4.1 Introduction

In this chapter, we investigate adaptive seamless designs in the setting in which the aim of clinical trials is to declare the equivalence between experimental and reference products. In particular, equivalence trials are applied in the development of genetic products, which are called bioequivalence trials. To confirm bioequivalence, pharmacokinetics (PK) parameters such as area under the concentration-time curve (AUC) and maximum serum concentration (Cmax) are typically evaluated in healthy volunteers using two one-sided tests (TOST) (Schuirmann, 1987) with log-transformed values. Although some methods like the adaptive sample size recalculation have been proposed for bioequivalence trials (Montague et al., 2012; Potvin et al., 2008; Xu et al., 2016; Zheng et al., 2015), these are rarely used in the actual clinical setting. Since such trials are conducted by recruiting generally between 20 and 50 volunteers, most of them are not required to include interim analyses in the trial.

While numerous pharmaceutical sponsors have expressed interest in bioequivalence trials, the considerate interest in developing biosimilars is also growing (Chow, 2013). There is no unified definition for biosimilars; however, according to the guideline of European Medicines Agency (2014a), a biosimilar is defined as a biological medicinal product that contains an active substance similar to that of the original previously authorized biological medicinal product. Biosimilars differ from generic chemical products, e.g., with respect to the complexity and heterogeneity of the molecular structure (Berghout, 2011; Chow, 2013). A reduction in healthcare costs for patients can be expected if a biosimilar is approved by regulators and placed on the market. However, characteristically, a larger number of subjects would be required to investigate and clinically develop a biosimilar than that required for the development of a generic product because there are regulatory requirements that encourage sponsors to provide pharmacodynamics (PD) or efficacy data in addition to PK data (European Medicines Agency, 2014a; Food and Drug Administration, 2015). In practice, as described in Table 4.1, most biosimilars have been developed with PK/PD trials (Wang and Chow, 2009). The same can be said of the biosimilar developments in Japan as illustrated in Table 4.2 (Nagasaki and Ando, 2014). That is, there must be no clinically meaningful difference between the experimental and reference products. Moreover, as far as the PK data goes, a PK trial is sometimes conducted using a parallel-group clinical trial design instead of a crossover design owing to a high risk of immunogenicity (European Medicines Agency, 2014a). In this instance, a large number of patients would be required to declare PK equivalence. For instance, a trial that was primarily aimed at confirming the PK equivalence was conducted with 250 recruited patients based on the assumption that the coefficient of variation (CV) was 50% (Park et al., 2013).

Here, two main trials served as motivating examples for confirming PK and efficacy in establishing the equivalence of a biosimilar of the innovator infliximab (Park et al., 2013; Yoo et al., 2013). These trials indicated that a clinical trial to establish the efficacy equivalence is often required in case any relevant PD markers are unavailable and that a PK trial could be conducted in patients rather than healthy volunteers. Although these trials were conducted separately, they each had a primary endpoint that was intrinsically set to determine PK and efficacy equivalence, respectively.

Based on this motivating example, we consider the clinical development of biosimilars with emphasis on the importance and necessity of demonstrating the equivalence between experimental and reference products by including both PK and efficacy as primary endpoints. We assume that patients who have the same disease conditions are targeted to provide the equivalence data for the PK and efficacy. Methods using adaptive seamless designs, which allow sample size re-calculation based on interim data, could be applied in this setting. The adaptive seamless PK and efficacy design, which incorporates trials to establish both PK and efficacy equivalence, allows trials to be more efficient than classical trial designs.

This chapter is structured as follows. First, we introduce a motivating example for the development of one biosimilar in Section 4.2 and review the current statistical methods for assessing biosimilarity in Section 4.3. Then, we lay the frameworks of our proposed designs as an adaptive seamless PK and efficacy design in Section 4.1 while Section 4.5 presents a simulation study. Finally, in Section 4.6 we conclude this chapter with a discussion.

Table 4.1: Summary of PK/PD studies in the European Assessment Reports of biosimilar products (Wang and Chow, 2009)

| Product | INN | Dose | N | Acceptance criteria (90% CI) | PK endpoints | PD endpoints |
|---|---|---|---|---|---|---|
| Abseam | Epoetin alfa | IV | 76 | PK: [80-125%] | AUC (×) | Hb-AUEC |
| | | | | PD: [97-103%] | Cmax | |
| | | SC | 74 | PK: [80-125%] | AUC | Hb-AUEC |
| | | | | PD: [97-103%] | Cmax | |
| Retacrit | Epoetin zeta | IV | 24 | AUC: [80-125%] | AUC | - |
| | | | | Cmax: [70-143%] | Cmax | |
| | | SC | 48 | AUC: [80-125%] | AUC | - |
| | | | | Cmax: [70-143%] | Cmax | |
| Zartio | Filgrastim | IV | 26 | PK: [80-125%] | AUC | ANC-AUEC |
| | | | | PD: [80-125%] | Cmax | |
| | | SC / Low dose | 24 | PK: [80-125%] | AUC | ANC-AUEC |
| | | | | PD: [80-125%] | | |
| | | SC / Middle doses | 28 × 2 | PK: [80-125%] | AUC | ANC-AUEC |
| | | | | PD: [80-125%] | Cmax | CD34+-AUEC |
| | | SC / High dose | 40 | PK: [80-125%] | AUC | ANC-AUEC |
| | | | | PD: [80-125%] | Cmax | CD34+-AUEC |
| Biograstim | Filgrastim | SC & IV | 36 × 4 | PK: [80-125%] | AUC | ANC: |
| | | | | PD: | Cmax | AUEC, ANCmax |
| | | | | (ANC) [80-125%] | | CD34+ |
| | | | | (CD34+) [70-143%] | | AUEC |
| | | IV | 28 × 2 | PK: [80-125%] | AUC | ANC: |
| | | | | PD: [80-125%] | Cmax | AUEC, ANCmax |
| | | | | | | CD34+ |
| | | | | | | AUEC |
| Omnitrope | Somatropin | SC | 24 | AUC: [80-125%] | AUC | Qualitative comparison |
| | | | | Cmax: [80-125%] | Cmax | |
| Valtropin | Somatropin | SC | 24 | AUC: [70-143%] | AUC | - |
| | | | | Cmax: [70-143%] | Cmax | |

INN, International Nonproprietary Name; CI, confidence interval; SC, subcutaneous; IV, intravenous; AUC, area under the concentration-time curve; Cmax, maximum serum concentration; ANC, absolute neutrophil count; Hb, hemoglobin; AUEC, area under the effect-time curve.

Table 4.2: Clinical data package for biosimilar products in Japan (Nagasaki and Ando, 2014)

| Product | Approved year | Purpose of study | Dose | Subjects | Primary endpoints | Location |
|---|---|---|---|---|---|---|
| Somatropin | 2009 | PK | SC | Healthy | AUC, Cmax | Japan |
| | | Efficacy | SC | Children with growth failure | Height, HSDS, growth speed, HVSDS, projected final height | Foreign countries |
| Epoetin Kappa | 2009 | PK | IV | Patients | AUC | Japan |
| | | PK | SC | Healthy | AUC, Cmax | Japan |
| | | Efficacy | IV | Patients | Hb concentration | Japan |
| Filgrastim 1 | 2012 | PK/PD | SC | Healthy | PK: AUC, Cmax | Japan |
| | | | | | PD: ANC Cmax, ANC tmax, CD34+ Cmax, CD34+ tmax | Japan |
| | | PD | SC<br>Multiple dose | Healthy | CD34+ Cmax, CD34+ tmax | Japan |
| | | PK | IV | Healthy | AUC | Japan |
| Filgrastim 2 | 2013 | PK | IV | Healthy | AUC, Cmax | Japan |
| | | PK | SC<br>Low dose | Healthy | AUC, Cmax | Japan |
| | | PK | SC<br>High dose | Healthy | AUC, Cmax | Japan |
| | | PD | SC | Healthy | ANC AUEC, ANC Cmax | Japan |
| | | PD | SC<br>Multiple dose | Healthy | CD34+ AUEC, CD34+ Cmax | Japan |

SC, subcutaneous; IV, intravenous; AUC, area under the concentration-time curve; Cmax, maximum serum concentration; HSDS, height standard deviation score; HVSDS, height velocity standard deviation score, ANC, absolute neutrophil count; Hb, hemoglobin; AUEC, area under the effect-time curve.

## 4.2 Motivating example

First, we shall introduce a motivating example of the development of a biosimilar to the innovator infliximab, which is a monoclonal antibody against tumor necrosis factor (TNF)-$\alpha$ and used to treat patients with active rheumatoid arthritis who have shown an inadequate response to methotrexate (Maini et al., 1999). As the biosimilar of the innovator monoclonal antibody, biosimilar infliximab was recently first launched in several markets across numerous countries (Schellekens et al., 2015). Two clinical trials were performed, which were PK and phase III studies that demonstrated the equivalence using PK and efficacy endpoints, respectively, in developing the biosimilar. Both trials were designed as randomized, double-blind, parallel-group clinical trials and their primary endpoints were the Cmax as well as AUC and the American College of Rheumatology 20% improvement (ACR20) (Felson et al., 1995) for the PK and efficacy trials, respectively.

For the PK study, the required sample size was calculated as 196 based on the equivalence margin of 80% to 125%, power of 90%, and TOST with a significance level of 5% under the assumption of a geometric mean ratio of 1.00 and CV of 50%. In addition to PK, efficacy and safety were also compared in this study. In contrast, the sample size of the phase III study was 468, which was required to achieve 80% power to meet the equivalence margin within ±15% for ACR20 at a specific time point under the TOST with a significance level of 2.5%, assuming an expected response rate of 50% in both groups. As secondary endpoints, additional efficacy, immunogenicity, safety, PK, and PD were assessed although no adjustments for multiplicity were performed.

Consequently, the equivalence for the primary PK and efficacy endpoints was established in both trials, respectively. Each confidence interval (CI) was in the range of the corresponding equivalence margins. For the PK parameters, the geometric mean ratios (90% CI) were 1.05 (0.94 to 1.16) and 1.02 (0.95 to 1.09) for the AUC and Cmax, respectively. As an efficacy parameter, the ACR20 response rates for each group were 60.9 and 58.6%, and the difference in the response rates (95% CI) was 2% (−6% to 10%) for intention-to-treat population while the values were 73.4 and 69.7% with a difference of 4% (−4% to 12%) for the per-protocol population.

Note that each between product group difference observed varied from the pre-specified settings. The observed geometric mean ratio of the AUC especially deviated from 1.00, whereas the observed response rates for each group in the per-protocol population differed greatly from 50%. Even if a sponsor in one country designs a trial based on trials conducted in other countries, the observed values from that trial are not necessarily consistent with those of the other countries owing to reasons such as possible race- or measurement technique-related differences between countries (Takeuchi et al., 2015; Yoo et al., 2013). Whether or not that was applicable, these misspecifications have been accounted for by recruiting more patients, which allows for more drop-out or exclusion of population sets than required. In addition, the observed response rates for each group differentially deviated from the expected value of 50% and this was a conservative setting since the range of the CIs tends to widen. Owing to these remedies, the targeted sample sizes for both trials were in effect set to approximately 1.15 to 1.20 times the required sample size. Instead, to compensate for this risk of the failure to demonstrate equivalence, a sponsor could consider increasing the sample size in midcourse, that is, sample size re-calculation within adaptive seamless design.

## 4.3 Literature review

In anticipation of the impending expiration of a number of patents for biological medicinal products in numerous countries, some statistical methods have been developed for evaluating biosimilarity since 2010. Methods for assessing biosimilarity with respect to variability between the experimental and reference products have been investigated (Belleli et al., 2015; Hsieh et al., 2009; Yang et al., 2013; Zhang et al., 2014, 2013).

As a biosimilarity measure, a biosimilar index based on the concept of reproducibility probability has been proposed and discussed (Chow et al., 2013; Hsieh et al., 2013; Yang and Lai, 2014). Chiu et al. (2014) discussed the use of a Bayesian method that uses prior information. Pan et al. (2017) proposed a Bayesian group sequential design that incorporates information adaptively using a calibrated power prior. Chow et al. (2009) proposed methods for assessing biosimilarity based on the assumption that a biomarker is predictive of the clinical outcome. Li et al. (2013) proposed a biosimilarity trial design for evaluating clinical efficacy with asymmetrical margins. Liao and Darken (2013) developed a method for assessing biosimilarity by comparability of critical quality attributes. Furthermore, a three-arm parallel design, which consists of one experimental and two reference products from two different batches, was provided to investigate biosimilarity (Kang and Chow, 2013). When the three-arm parallel design is employed, the approach with the use of the frequency estimator criterion was also proposed to assess biosimilarity (Lu et al., 2014).

In summary, most methods currently focus on one specified trial. To the best of our knowledge, little methodology that enables the performance of multiple trials seamlessly has been developed.

## 4.4 Proposed framework

In this section, we shall consider the clinical development of biosimilars under a randomized parallel-group design with two-arms, experimental ($E$) and reference ($R$) products. Similar to the motivating examples of the biosimilar infliximab (Park et al., 2013; Yoo et al., 2013), we will assume that a crossover design is not a feasible option for the trial design because of the associated problem of carry-over effects, although crossover designs are often used in the development of biosimilars where recruitment for the study population is targeted at healthy volunteers (Nagasaki and Ando, 2014; Wang and Chow, 2009). Based on the motivating example involving two trials, a single trial is conducted for the evaluations of both the PK and efficacy, and their confirmations constitute the interim and final analyses, respectively. To compensate for the misspecification of parameters at the planning stage, an interim analysis will be performed for the final efficacy confirmation that requires more patients than the PK confirmation does. Figure 4.1 shows the proposed framework of the adaptive seamless PK and efficacy design. Note that the objective of the trial is to determine the equivalence of both PK and efficacy.

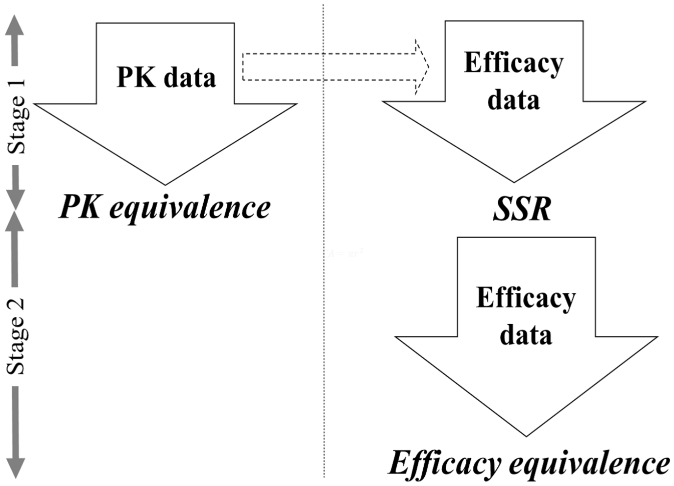

Figure 4.1: Proposed adaptive seamless pharmacokinetics (PK) and efficacy design
SSR, sample size re-calculation.

### 4.4.1 Collecting both PK and efficacy data from first stage

Here, we considered that not only PK data but also efficacy data could be obtained from the PK trial, and to achieve this, we assumed that both the PK and efficacy trials should be conducted with patients who have common diseases and are receiving similar dosage regimens.

The strength of this study is that it constitutes a clinical trial that includes the statistical analysis of PK as an interim analysis of the efficacy data. Therefore, we consider a two-stage design based on the assumption that both the PK data and efficacy data are obtainable from the first stage, as described in Figure 4.1. The data from the first stage can subsequently be used to test the PK equivalence and arrive at an interim decision for the efficacy equivalence, whereas only the efficacy data are obtainable from the second stage to test the efficacy equivalence.

### 4.4.2 Sample size adjustment for efficacy endpoint of interim analysis

The interim analysis conducted in an unblinded fashion gives the option to re-calculate sample size for the subsequent stage if the interim result shows a potential benefit for sample size re-calculation. Note that the interim decision for the subsequent stage is only based on the efficacy data of the interim analysis.

Let $Z_1$ and $Z_2$ denote the test statistics until the interim and final analysis, respectively. Without any interim analysis from the first stage data, the hypothesis testing would be conducted conventionally, with $Z_2 > z_\alpha$ as the one-sided significance level $\alpha$ where $z_\alpha$ is the critical value for a superiority clinical trial. In addition, the sample size is estimated to achieve $1-\beta$ power under an alternative hypothesis.

Mehta and Pocock (2011) have proposed methods in which the sample size is increased if the interim results are promising, and the decision is made based on the conditional power. Recently, a thorough investigation of this approach has been conducted and reported in several articles in the context of a superiority trial (Chen et al., 2015; Jennison and Turnbull, 2015). In this study, we consider this approach in evaluating the equivalence of an efficacy endpoint. The conditional power given by with the observed difference between the groups derived from the interim efficacy data is expressed using the following equation (Jennison and Turnbull, 2000):

$$\begin{aligned} CP(z_1, \tilde{N}_2) &= \Pr(Z_2 > z_\alpha \mid Z_1 = z_1) \\ &= 1 - \Phi\left( \frac{z_\alpha \sqrt{N_2} - z_1 \sqrt{N_1}}{\tilde{N}_2} - \frac{z_1 \sqrt{\tilde{N}_2}}{N_1} \right) \end{aligned}$$

where $N_1$ and $N_2$ are the planned sample sizes until the interim and final analysis, respectively, and $\tilde{N}_2$ is the increment in the sample size during the second stage denoted by $\tilde{N}_2 = N_2 - N_1$.

The promising zone approach arrives at the interim decision by defining the region that represents the promising zone as follows:

Case 1 (Promising): $cp_L \leq CP(z_1, \tilde{N}_2) < 1 - \beta$ → Increase the sample size to $N_2^*$ ;

Case 2 (Otherwise, i.e., favorable or unfavorable): → Continue to the planned $N_2$ .

where $cp_L$ is the pre-specified lower probability of the conditional power. If the interim conditional power is deemed promising, the sample size is increased to $N_2^* = \min(N_2'(z_1), N_{\max})$ where $N_2'(z_1)$ consists of the sum of $N_1$ and $\tilde{N}_2'(z_1)$, which is the increment in the sample size when the sample size is increased from the planned $N_2$. To satisfy the $CP(z_1, \tilde{N}_2) = 1-\beta$ on the condition of the promising zone, the increased sample size is derived as:

$$\tilde{N}_2'(z_1) = \left(\frac{N_1}{z_1^2}\right)\left[\frac{z_\alpha \sqrt{N_2} - z_1 \sqrt{N_1}}{N_2 - N_1} - z_\beta\right]^2$$

with the restriction that the maximum sample size increase is $N_{\max}$. Then, the critical value for the final analysis, in exchange for $z_\alpha$, can be adjusted to

$$z'(z_1, N_2^*) = \frac{1}{\sqrt{N_2^*}}\left[\sqrt{\frac{N_2^* - N_1}{\tilde{N}_2}}(z_\alpha \sqrt{N_2} - z_1 \sqrt{N_1}) + z_1 \sqrt{N_1}\right]$$

which holds that $\Pr\left\{Z_2^* > z'(z_1, N_2^*)\right\} = \alpha$, where $Z_2^*$ is the test statistic for the final analysis using $N_2^*$ instead of $N_2$ (Gao et al., 2008). When the sample size is increased to $N_2^*$, the power for performing the final analysis is greater under the critical value $z'(z_1, N_2^*)$ than it is under the $z_\alpha$.

The final analysis would also be conducted using $Z_2^* > z_\alpha$ similar to the conventional hypothesis testing in line with the rule that the sample size is only increased if the interim conditional power is deemed promising. Note that the inflation of type I error rate would occur if the decision rule is not adhered to by the promising zone approach (Proschan and Hunsberger, 1995).

With an equivalence trial, the interim result, $z_1$, for efficacy is constructed using the following null hypotheses for the efficacy endpoint:

$$H_0^{\text{Eff}} : H_{0,U}^{\text{Eff}} \cup H_{0,L}^{\text{Eff}},$$

that is,

$$H_{0,U}^{\text{Eff}} : q_E - q_R \geq \Delta_{\text{Eff}} \quad \textit{and} \quad H_{0,L}^{\text{Eff}} : q_E - q_R \leq -\Delta_{\text{Eff}}$$

where $q_g$ is the response rate for the group $g \in \{E, R\}$ and $\Delta_{\text{Eff}}$ is a pre-specified equivalence margin of efficacy. Moreover, the alternative hypotheses for the efficacy are shown by

$$H_A^{\text{Eff}} : H_{A,U}^{\text{Eff}} \cap H_{A,L}^{\text{Eff}},$$

that is,

$$H_{A,U}^{\text{Eff}} : q_E - q_R < \Delta_{\text{Eff}} \quad \textit{and} \quad H_{A,L}^{\text{Eff}} : q_E - q_R > -\Delta_{\text{Eff}}.$$

### 4.4.3 PK equivalence including early stopping for efficacy

We shall begin this section by describing the hypotheses for the PK evaluation. For the PK equivalence, the null hypotheses for PK are constructed as follows:

$$H_0^{\text{PK}} : H_{0,U}^{\text{PK}} \cup H_{0,L}^{\text{PK}},$$

that is,

$$H_{0,U}^{\text{PK}} : X_E - X_R \geq \Delta_{\text{PK}} \quad \textit{and} \quad H_{0,L}^{\text{PK}} : X_E - X_R \leq -\Delta_{\text{PK}}$$

where $X_g$ is the log-transformed mean of the PK parameters such as the AUC or Cmax for group $g \in \{E, R\}$ while $\Delta_{\text{PK}}$ is the pre-specified equivalence margin of the PK endpoint. Thus, the alternative hypotheses for the PK are as follows:

$$H_A^{\text{PK}} : H_{A,U}^{\text{PK}} \cap H_{A,L}^{\text{PK}},$$

that is,

$$H_{A,U}^{\text{PK}} : X_E - X_R < \Delta_{\text{PK}} \quad \textit{and} \quad H_{A,L}^{\text{PK}} : X_E - X_R > -\Delta_{\text{PK}}.$$

In the bioequivalence trial, the $\Delta_{\text{PK}}$ is set as the log-transformed 1.25, which corresponds to a range of 80% to 125% under the $H_A^{\text{PK}}$ (Food and Drug Administration, 2001). In this study, we consider a biosimilar development process where it is necessary to demonstrate the equivalence of both the PK and efficacy endpoints. Hence, the null and alternative hypotheses $H_0$ and $H_A$, respectively are expressed as

$$H_0 : H_0^{\text{PK}} \cup H_0^{\text{Eff}} \quad vs. \quad H_A : H_A^{\text{PK}} \cap H_A^{\text{Eff}}.$$

Note that the $H_0$ is rejected only if the equivalence for both the PK and efficacy is established. This meets the requirement based on several guidelines in which both the PK and efficacy are required to show equivalence when developing biosimilars (European Medicines Agency, 2014a; Food and Drug Administration, 2015).

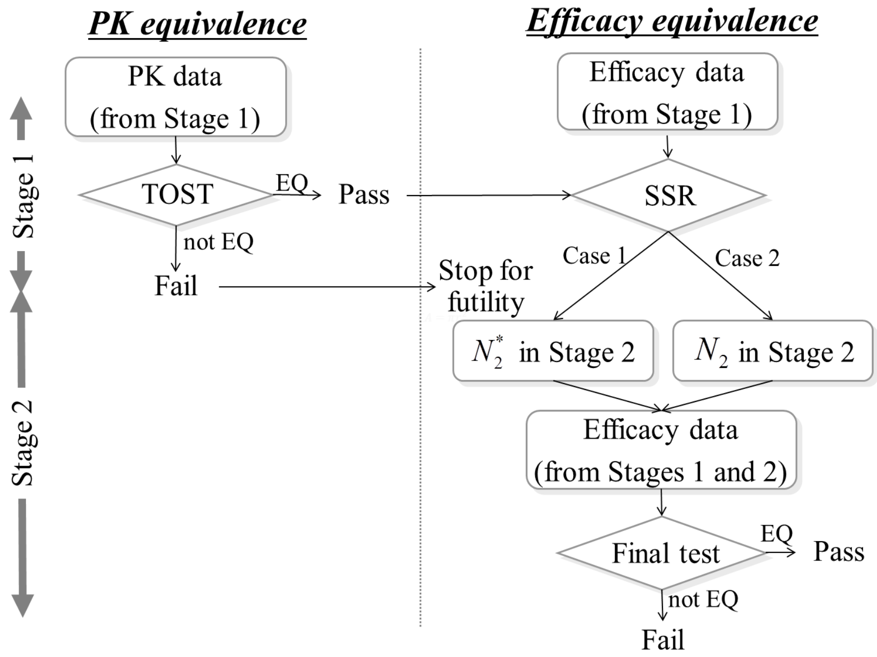

Figure 4.2: Framework that determines equivalence of both pharmacokinetic (PK) and efficacy. ”EQ” and ”not EQ” denote where equivalence is declared and not declared, respectively. SSR, sample size re-calculation; TOST, two one-sided tests.

Figure 4.2 shows the detailed framework used to declare the equivalence of both the PK and efficacy within the adaptive seamless PK and efficacy design as described in Figure 4.1. With respect to controlling the type I error rate for PK and efficacy equivalence under each significance level, a fixed sequence testing procedure (Wiens, 2003) is incorporated into this framework. That is, the efficacy equivalence is tested only if the PK equivalence is declared. As shown in Figure 4.1, it organizes a clinical trial that includes the statistical analysis of PK as an interim analysis of the efficacy data because we considered that both PK data and efficacy data could be obtained from the sample size $N_1$ in the PK trial. If PK equivalence fails to be declared, subsequent collection of efficacy data is discontinued for futility. On the other hand, collecting efficacy data is subsequently continued if PK equivalence is declared. Thereafter, sample size re-calculation from $N_2$ to $N_2^*$ is conducted following the promising zone approach in Case 1, as shown in Figure 2, while the enrollment is continued to the planned $N_2$ in Case 2.

## 4.5 Simulation study

Here, We evaluated the efficiency of the adaptive seamless design under each framework proposed in this study. This is to enable to presentation of their operating characteristics based on the design parameters such as the geometric mean ratio and CV in the PK trial, and response rates for each group in the efficacy trial. The assumptions made for the PK and efficacy trial designs as well as the simulation setting are presented in Section 4.5.1. In Section 4.5.2, we demonstrate the effectiveness of the adaptive seamless PK and efficacy design with the frameworks proposed based on the power and expected sample size.

### 4.5.1 Simulation setting

First, we shall consider a randomized, parallel-group clinical trial with two arms, which are the experimental ($E$) and reference ($R$) product arms. We assume that $E$ is set as the biosimilar, and the sample size re-calculation for the efficacy endpoint under the adaptive seamless PK and efficacy design proposed is conducted with an interim analysis, which plays a role in the statistical analysis of the PK data.

Based on the examples (Park et al., 2013; Yoo et al., 2013), we suppose that the overall planned sample size is set as $N_2 = 480$, and the interim analysis is conducted after $N_1 = 200$ are recruited in the trial. The set of $N_2 = 480$ and $N_1 = 200$ corresponds to 80% power for efficacy assuming the 50% response rates expected and 90% power for the PK, assuming that the geometric mean ratio and CV are 1.00 and 50%, respectively (Park et al., 2013; Yoo et al., 2013). The equivalence margins are set as ±15% for efficacy, which is binary data and a range of 80% to 125% for the PK using a one-sided significance level of 2.5% and 5% for efficacy and PK, respectively.

To assess the sample size re-calculation for efficacy equivalence, the lower probability of the conditional power $cp_L$ is set at 50% or 33% (Chen et al., 2004; Mehta and Pocock, 2011). We further assume that the magnitude of the sample size increase is set as $R_{\max}$, which denotes the ratio of $N_{\max}$ to $N_2$. The assessment of the power of the efficacy is performed using the true difference between the response rates of the product groups denoted as $\pi_{\text{Diff}}$ of 0% to 5%.

For reference, the power and expected sample size are also illustrated when a fixed design is used in exchange for the adaptive seamless PK and efficacy design. Regarding the expected sample size using the fixed design, we consider that the PK and efficacy trials are conducted separately, suggesting that the efficacy data are not available for the PK confirmation. That is, the PK trial is assumed to be conducted and the efficacy trial is subsequently conducted separately under the fixed design. Hence, if the PK equivalence fails to be declared, the efficacy trial is not implemented.

### 4.5.2 Simulation results

We evaluated the efficiency of the adaptive seamless PK and efficacy design proposed based on the type I error rate, power, and expected sample size. The calculation was based on 500,000 simulations with respect to the type I error rate and 10,000 with respect to power and expected sample size to display the operating characteristics.

#### Controlling type I error

Table 4.3 shows the probabilities of rejecting the null hypotheses $H_0^{\text{PK}}$, $H_0^{\text{Eff}}$, or $H_0 : H_0^{\text{PK}} \cup H_0^{\text{Eff}}$, i.e., the type I error rates where the geometric mean ratio is 1.25 for the PK and the difference in the response rates between the groups is 15% for the efficacy. As shown in Table 4.3, the type I error rates of the PK and efficacy endpoints are controlled under the significance level of 5% and 2.5%, respectively. In particular, the type I error rate of the efficacy endpoint is confirmed to be controlled even when it includes a sample size re-calculation as long as multiplicity adjustments are included as described in Section 4.4.2, in contrast with that without multiplicity adjustments. Therefore, the probability of rejecting the null hypothesis $H_0 : H_0^{\text{PK}} \cup H_0^{\text{Eff}}$ is also controlled owing to the fixed sequence testing procedure for PK and efficacy equivalence, as described in Section 4.4.3.

Table 4.3: Type I error rates

| $cp_L$ | $N_1$ | Rejecting $H_0^{\text{PK}}$ | Rejecting $H_0^{\text{Eff}}$ with multiplicity adjustments | Rejecting $H_0^{\text{Eff}}$ without multiplicity adjustments |
|---|---|---|---|---|
| 0.33 | 200 | 0.050 | 0.024 | 0.029 |
| | 120 | 0.050 | 0.023 | 0.027 |
| 0.50 | 200 | 0.050 | 0.023 | 0.029 |
| | 120 | 0.050 | 0.023 | 0.027 |

Calculation was performed under the assumption that $N_2 = 480$, $R_{\max} = 2.0$, and expected response rate of 50% as a function of lower probability of conditional power ($cp_L$), planned sample size until interim analysis ($N_1$).

#### Power comparisons

The comparison of powers between the fixed design and the adaptive seamless PK and efficacy design in the final analysis is shown in Table 4.4. These powers are defined as the probability of rejecting $H_0 : H_0^{\text{PK}} \cup H_0^{\text{Eff}}$ when using the adaptive seamless PK and efficacy design. The result revealed that the design is more powerful when using the adaptive seamless PK and efficacy design owing to the incorporation of sample size re-calculation. Note that the adaptive seamless PK and efficacy design would minimize the risk of misspecification of the pre-specified parameters of the efficacy endpoint. In particular, it is possible to compensate for the power using this design under the situation where the conditional power falls within the range of the promising zone in the interim analysis of the efficacy and sample size is increased.

Table 4.4: Powers of adaptive seamless pharmacokinetic (PK) and efficacy design and fixed designs

| $cp_L$ | $N_1$ | $\pi_{\text{Diff}}$ | Fixed design | Adaptive seamless PK and efficacy design |
|---|---|---|---|---|
| 0.33 | 200 | 0.00 | 0.740 | 0.771 |
| | | 0.01 | 0.731 | 0.763 |
| | | 0.02 | 0.700 | 0.740 |
| | | 0.03 | 0.658 | 0.701 |
| | | 0.04 | 0.599 | 0.645 |
| | | 0.05 | 0.530 | 0.575 |
| | 120 | 0.00 | 0.523 | 0.547 |
| | | 0.01 | 0.519 | 0.544 |
| | | 0.02 | 0.499 | 0.528 |
| | | 0.03 | 0.467 | 0.489 |
| | | 0.04 | 0.425 | 0.454 |
| | | 0.05 | 0.376 | 0.399 |
| 0.50 | 200 | 0.00 | 0.740 | 0.761 |
| | | 0.01 | 0.731 | 0.755 |
| | | 0.02 | 0.700 | 0.726 |
| | | 0.03 | 0.658 | 0.681 |
| | | 0.04 | 0.599 | 0.627 |
| | | 0.05 | 0.530 | 0.552 |
| | 120 | 0.00 | 0.523 | 0.528 |
| | | 0.01 | 0.519 | 0.529 |
| | | 0.02 | 0.499 | 0.511 |
| | | 0.03 | 0.467 | 0.475 |
| | | 0.04 | 0.425 | 0.434 |
| | | 0.05 | 0.376 | 0.385 |

Calculation was performed under the assumption that $N_2 = 480$, $R_{\text{max}} = 2.0$ and expected response rate of 50% as a function of lower probability of conditional power ($cp_L$), planned sample size until interim analysis ($N_1$), and difference between response rates ($\pi_{\text{Diff}}$).

### Expected sample size comparisons

We showed that the power is improved by the use of the frameworks for PK and the sample size re-calculation in the promising zone approach for the efficacy of the adaptive seamless PK and efficacy design. The gain of power is attributable to the increase in sample size. Table 4.5 shows the expected sample size for achieving the corresponding power illustrated in Table 4.4. Note that the expected sample sizes for the fixed design do not exactly correspond to the sum of the interim and final sample size (i.e., $N_1$ and $N_2$) because of the failure to declare the PK equivalence in the PK trial preceding the efficacy trial. It is obvious that the approach that uses the adaptive seamless PK and efficacy design is more efficient than the approach that uses the fixed design where $N_1$ is set at 200 and the expected sample size is decreased because the evaluations for both the PK and efficacy are conducted separately in the latter design. Hence, it is necessary to set a sample size that has sufficient power. Furthermore, the increase in power and expected sample size is higher when $cp_L$ is set at a lower value.

Table 4.5: Expected sample sizes corresponding to powers in adaptive seamless pharmacokinetic (PK) and efficacy design and fixed designs

| $cp_L$ | $N_1$ | $\pi_{\text{Diff}}$ | Fixed design | Adaptive seamless PK and efficacy design |
|---|---|---|---|---|
| 0.33 | 200 | 0.00 | 635.6 | 530.4 |
| | | 0.01 | 635.6 | 529.8 |
| | | 0.02 | 635.6 | 530.4 |
| | | 0.03 | 635.6 | 531.7 |
| | | 0.04 | 635.6 | 527.5 |
| | | 0.05 | 635.6 | 524.1 |
| | 120 | 0.00 | 427.7 | 406.7 |
| | | 0.01 | 427.7 | 405.0 |
| | | 0.02 | 427.7 | 406.4 |
| | | 0.03 | 427.7 | 404.1 |
| | | 0.04 | 427.7 | 403.0 |
| | | 0.05 | 427.7 | 401.6 |
| 0.50 | 200 | 0.00 | 635.6 | 501.0 |
| | | 0.01 | 635.6 | 500.3 |
| | | 0.02 | 635.6 | 500.9 |
| | | 0.03 | 635.6 | 499.7 |
| | | 0.04 | 635.6 | 498.8 |
| | | 0.05 | 635.6 | 495.5 |
| | 120 | 0.00 | 427.7 | 372.6 |
| | | 0.01 | 427.7 | 371.5 |
| | | 0.02 | 427.7 | 371.4 |
| | | 0.03 | 427.7 | 371.1 |
| | | 0.04 | 427.7 | 370.6 |
| | | 0.05 | 427.7 | 369.5 |

Calculation was performed under the assumption that $N_2 = 480$, $R_{\text{max}} = 2.0$ and expected response rate of 50% as a function of lower probability of conditional power ($cp_L$), planned sample size until interim analysis ($N_1$), and difference between response rates ($\pi_{\text{Diff}}$).

## 4.6 Summary and discussion

The aim of this study was to propose a novel adaptive seamless PK and efficacy design for establishing the equivalence of both the PK and efficacy in clinical trial phases for the development of biosimilars. The proposed design, which allows sponsors to develop biosimilars with shorter periods, leading to additional cost savings and fewer patients required for trials would be an appealing strategy for implementing trials efficiently. This enhanced process would consequently accelerate product approval by regulatory agencies.

Our proposed framework is an attractive option with respect to the total trial period as mentioned in Section 4.4. For efficacy equivalence, sample size re-calculation was incorporated in the adaptive seamless PK and efficacy design to compensate for the risk of misspecification of the efficacy parameters. The study is subject to early termination in the efficacy part if PK equivalence fails to be declared. The power was improved as shown in Table 4.4, but not dramatically increased with sample size re-calculation for the efficacy part because we considered that the pre-planned sample size was adequate to achieve the target level power of 80%. Note that the trial should be planned carefully to estimate the sample size and should not be set up deliberately with an underestimated sample size with insufficient power solely dependent on the sample size re-calculation. With the planned sample size with insufficient power, the expected sample size using the adaptive seamless PK and efficacy design is larger than that using a fixed design: however, it is smaller when a sample size with sufficient power is used. The promising zone approach also enables the trial to avoid implementing further support when the interim result deems it obviously unpromising. This would reallocate and optimize the additional investment of resources. However, a downside of sample size re-calculation is that the statisticians associated with the sponsor can grasp the interim conditional power based on the additional sample size to be enrolled during the subsequent stage.

In conclusion, our study proposed a novel method for developing biosimilars using an adaptive seamless design that enables sample size re-calculation based on interim data and incorporates trials to establish both PK and efficacy equivalence. Furthermore, the newly proposed design allows clinical trials to be more efficiently conducted than conventionally designed methods, thereby reducing costs, saving time, and providing an attractive option for pharmaceutical sponsors.

# Chapter 5

# Discussion

## 5.1 Applicability of adaptive population selection designs in oncology

In Chapters 2 and 3, we focused on the the adaptive designs with population selection using survival endpoints. If the source of the potentially pre-defined subgroup regarded as biomarker-positive is known, we can consider the situation in which the full-population comprises biomarker-positive and biomarker-negative populations. Note that the adaptive population selection design meets the regulatory requirement as it provides efficacy results for *N*. However, a large number of patients would be required when using the adaptive population selection designs in the situation in which the experimental treatment shows a benefit for *P* but has a negative effect on *N*, particularly in Scenario 4 of Sections 2.5.1 and 3.3.1. As evaluated by Graf et al. (2015), the judgment that would be preferred among a fixed design, an enrichment design, and the adaptive population selection design would depend on the situation regarding the prevalence of the biomarker-positive patients. As an example in which the adaptive population selection design is unfavorable, particularly in the situation in which the prevalence of *P* is small, the enrichment design rather than the classical fixed design is applied in the development of crizotinib, since there are approximately 5% of patients who present the rearrangement of the anaplastic lymphoma kinase (ALK) gene among all of the patients with non-small-cell lung cancer (Solomon et al., 2014). Therefore, the approach described in this work can be applied when there is an available biomarker prior to designing a clinical trial and in the situation in which an adaptive population selection design is preferred.

## 5.2 Multiplicity issues in adaptive population selection designs

Regarding the interim analysis for population selection described in Chapters 2 and 3, it was unnecessary to perform any multiple testing procedures since we assumed, for the sake of simplicity, that the interim analysis incorporates the early termination only for futility. With respect to multiplicity issues in which we assumed two hypotheses for the full population and biomarker-positive population at the final analysis, we used Simes' procedure to control the familywise type I error rate. However, as presented in Section 3.3.2, these results were conservative because of the assumptions of an asymptotic bivariate normal distribution and positive correlation. Although the method based on the conditional error function approach, described by Friede et al. (2012), would be more powerful for the final analysis, we primarily focused on the interim decision so as to improve the probability of selecting the most appropriate population. In addition, as another type of correction, the Spiessens and Debois (2010) procedure can be applied to control the familywise type I error rate for the final analysis, since we considered the multiplicity under a situation where the test statistics between $F$ and $P$ are positively correlated. Furthermore, as a more complicated situation, this work can be extended to a clinical trial that has more than two arms: adaptive designs for treatment and population selection. Although adaptive designs for treatment selection are often scrutinized under a multi-arm setting, most cases are assumed to be selected at the interim analysis and to conduct the final analysis for two arms: experimental and control. Thus, as a practical setting for population selection under multiple arms, more complex corrections for the familywise type I error rate would be needed for adaptations for both treatment and population selection conducted at the interim analysis.

## 5.3 The proportional hazards assumption in molecular targeted therapies

In Chapter 2, we tackled the current issue in the development of molecular targeted therapies for which the proportional hazard assumption is violated for the full population. The interim decision rule based on the estimated hazard ratios mentioned by Jenkins et al. (2011) may be preferred, when using the adaptive population selection design, in terms of consistency between the interim and final analyses; however, we used predictive power only for the interim analysis owing to the violation of the proportional hazard assumption in developing molecular targeted therapies. As an alternative to the hazard ratio, we further considered the use of the restricted mean survival time at the interim analysis. Between the hazard ratio and the restricted mean survival time as a measure in the interim decision, the difference or ratio of the restricted mean survival time is preferred owing to the characteristic that the proportional hazard assumption has collapsed. However, this is considering the property with respect to the statistical power, where it is less powerful than that using the hazard ratio if the restricted mean survival time is employed at the final analysis similarly (Trinquart et al., 2016). Admittedly, the characteristic of the loss of power for the restricted mean survival time would lead to the results of the probabilities of selecting the population at the interim analysis in our simulation studies. Nevertheless, we proved that the use of the restricted mean survival time can be conducted in our proposed approach and that this would perform fairly if the sample size is increased to satisfy the desired power for at the planning stage in practice. Note that this would achieve the current need for the restricted mean survival time instead of the hazard ratio, if the proportional hazard assumption is violated. Future research need to improve the loss of power for the restricted mean survival time.

Our simulation studies performed in Chapters 2 and 3 showed that the use of measures in midcourse between the confidence interval of the estimated hazard ratio and the predictive power are comparable, as long as the predictive power is derived from the test statistic using a log-rank test and the non-informative priors for the treatment effect are employed. However, the predictive power is flexible and preferred, if the proposed approach is extended to some disease setting in the sense that the predictive power can be derived based on other non-parametric test statistics. For instance, the Fleming and Harrington (1991) test is often used in the development of cancer vaccines and immunotherapies (Hasegawa, 2014, 2016). In the development of molecular targeted therapies, the predictive power may be preferred. A possible future investigation would be to consider how external information, or any other knowledge obtained prior to the phase II trial, might be incorporated into the best use of informative priors and for setting the thresholds for predictive power in the interim decision rule.

## 5.4 Applicability of adaptive seamless design in developing biosimilars

In Chapter 4, we proposed a novel adaptive seamless PK and efficacy design with efficient frameworks for establishing the equivalence of both the PK and efficacy in clinical trial phases for the development of biosimilars. Note that there is still a controversy about statistical analysis for biosimilar development, even though related guidelines (European Medicines Agency, 2014a; Food and Drug Administration, 2015) have been issued from the regulatory agencies. For instance, choosing the margin, primary endpoint, and primary time point for efficacy represent the issues and challenges with respect to biosimilar development. Hence, consultation with regulatory agencies must be required before applying the proposed design, which has originality specific to biosimilar development and offers benefits even considering the issues and challenges.

It is noteworthy to mention that this work was limited to a specific situation where there are no relevant PD markers for measuring the efficacy in clinical trials. In addition, we propose the adaptive seamless PK and efficacy design with the restriction that both PK and efficacy trials are required to be conducted with patients. This is because the premise of this study is based on the characteristic of biosimilar development trials that are often conducted in patients rather than in healthy volunteers (Nagasaki and Ando, 2014). Although healthy volunteers are used in most applications for biosimilars (Wang and Chow, 2009), the development of biosimilars has a greater possibility of targeting patients hereafter owing to the high molecular complexity of biosimilars. For example, the biosimilar of trastuzumab (Herceptin), which is similar to the biosimilar infliximab because it is a monoclonal antibody, has been developed (Stebbing et al., 2017). In addition, there would be a controversial issue with respect to the assumption that patients who have common diseases and are being treated with similar dosage regimens are targeted to provide equivalence data for PK and efficacy. Note that, in the motivating example described in Section 4.2 in Chapter 4, the targeted patients for the PK study had ankylosing spondylitis (Park et al., 2013), whereas those for the efficacy study had rheumatoid arthritis (Yoo et al., 2013). However, there is also an example in which targeted patients were rheumatoid arthritis even in the PK study (Takeuchi et al., 2015). Moreover, our framework was constructed using parallel-group clinical trial designs based on the two main trials performed in the development of Remsima (Park et al., 2013; Yoo et al., 2013). Because crossover designs for PK trials are often used, the clinical trials that consist of PK evaluations using crossover designs and efficacy trials using parallel-group designs could be extended for use with the adaptive seamless PK and efficacy design. In this case, the adaptive sequential design used for PK confirmation (Montague et al., 2012; Potvin et al., 2008; Xu et al., 2016; Zheng et al., 2015) would be an additional option for the PK trial. As a further and practical consideration, a multiple testing issue for multiple PK endpoints would be needed in addition to the fixed sequence testing procedure considered between PK and efficacy endpoints because two PK endpoints, i.e., AUC and Cmax, are often evaluated in practice in the PK trial (Hua et al., 2015). In addition, several types of AUC are often set as primary endpoints. For instance, AUCs from time zero to predicted infinity and from time zero to the last measurable concentration were assessed in addition to Cmax as primary endpoints in the PK study within the development of the biosimilar adalimumab (Wynne et al., 2016). Further, other PK parameters, such as tmax, volume of distribution, and half-life, should be set as secondary PK endpoints, whereas AUC and Cmax are frequently set as primary PK endpoints (European Medicines Agency, 2014a). In the motivating PK trial (Park et al., 2013), nine parameters were set as secondary endpoints, whereas AUC and Cmax were set as primary endpoints. Although multiplicity for secondary PK endpoints was not usually addressed and only primary PK endpoints were required to demonstrate equivalence statistically, providing these secondary PK parameters is necessary to conclude biosimilarity in practice. Furthermore, we assumed that the primary efficacy endpoint in a binary type. In the development of therapies for patients with rheumatoid arthritis, continuous data are also set as a primary endpoint, i.e., co-primary endpoints. In this case, the sample size estimation for co-primary endpoints, which are both continuous and binary endpoints to be evaluated, should pragmatically be performed in consideration of a correlation between endpoints (Sozu et al., 2012).

# Chapter 6

# Conclusion

Recently, regulatory agencies as well as numerous pharmaceutical sponsors have expressed a great deal of interest in the development of molecular targeted therapies and biosimilars. In this dissertation, we addressed the adaptive designs to contribute to conducting confirmatory clinical trials more efficiently.

First, Issue 1 proposed a novel utility-based approach to guide the construction of the interim decision rule of an adaptive population selection design for the setting of the survival endpoint. In Chapter 2, We introduced utility functions constructed from gain and loss functions into the method for setting the thresholds within the interim decision. In our simulation studies, we considered the hazard ratio, predictive power, and difference and ratio of restricted mean survival time as interim measures. Simulation results revealed that the proposed approach can guide the setting of thresholds considering the benefits and risks for sponsors and patients. However, the assumption that it is not beneficial rather than harmful would be plausible in implementing the clinical trial that also includes the biomarker-negative population with belief that the therapy might be harmful for that population. Therefore, setting thresholds must be carefully considered based on the results of using the proposed approach.

Second, Issue 2 was motivated by the development of two molecular targeted therapies: gefitinib and cetuximab. We improved the interim decision rule in the setting in which we consider the phase II/III trials with progressive cancer patients using correlated survival endpoints: OS and PFS. In our approach, the interim decision was made by incorporating information on OS as well as PFS to supplement the incomplete OS data. The use of interim decision-making strategies proposed in Chapter 3 showed good performance for selecting the most appropriate population in developing molecular targeted therapies when using OS as well as PFS. This is particularly relevant under a scenario in which PPS affects the correlation between OS and PFS.

Finally, Issue 3 developed a novel adaptive seamless PK and efficacy design in response to the current situation in which little methodology that enables the performance of multiple trials seamlessly has been developed. In Chapter 4, we considered the clinical development of biosimilars including their evaluation in patients rather than healthy volunteers under a situation where both PK and efficacy parameters are required to demonstrate the equivalence. The original idea of the proposed method was to organize a clinical trial that includes the statistical analysis of PK as an interim analysis, with sample size recalculation of the efficacy data. Our simulation study indicated that the proposed design would allow trials to be more efficient than with the classical design. Therefore, this proposal provides appealing advantages, such as a shorter period, additional cost savings, and a smaller number of patients required.

In summary, the outcome of this study will contribute to the development of two types of state of the art therapies: molecular targeted therapies and biosimilars.

# Bibliography

Aggarwal, S. (2010). Targeted cancer therapies. *Nature Reviews Drug discovery* 9(6): 427–428.

A'Hern, R. P. (2016). Restricted mean survival time: an obligatory end point for time-to-event analysis in cancer trials? *Journal of Clinical Oncology* 34(28): 3474–3476.

Arrowsmith, J., Miller, P. (2013). Phase ii and phase iii attrition rates 2011-2012. *Nature Reviews Drug Discovery* 12(8): 569–570.

Bauchner, H., Fontanarosa, P. B., Golub, R. M. (2017). Scientific evidence and financial obligations to ensure access to biosimilars for cancer treatment. *Journal of the American Medical Association* 317(1): 33–34.

Bauer, P., Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics 50*: 1029–1041.

Bauer, P., Posch, M. (2004). Letter to the editor: Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine* 23(8): 1333–1334.

Belleli, R., Fisch, R., Renard, D., Woehling, H., Gsteiger, S. (2015). Assessing switchability for biosimilar products: modelling approaches applied to children's growth. *Pharmaceutical Statistics* 14(4): 341–349.

Berghout, A. (2011). Clinical programs in the development of similar biotherapeutic products: rationale and general principles. *Biologicals* 39(5): 293–296.

Biankin, A. V., Piantadosi, S., Hollingsworth, S. J. (2015). Patient-centric trials for therapeutic development in precision oncology. *Nature* 526(7573): 361–370.

Blevins, T., Dahl, D., Rosenstock, J., et al. (2015). Efficacy and safety of LY2963016 insulin glargine compared with insulin glargine (lantus®) in patients with type 1 diabetes in a randomized controlled trial: the ELEMENT 1 study. *Diabetes, Obesity and Metabolism* 17(8): 726–733.

Boessen, R., Baan, F., Groenwold, R., et al. (2013). Optimizing trial design in pharmacogenetics research: comparing a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. *Pharmaceutical Statistics* 12(6): 366–374.

Booth, C. M., Eisenhauer, E. A. (2012). Progression-free survival: meaningful or simply measurable? *Journal of Clinical Oncology* 30(10): 1030–1033.

Brannath, W., Zuber, E., Branson, M., et al. (2009). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* 28(10): 1445–1463.

Bretz, F., Gallo, P., Maurer, W. (2017). Adaptive designs: The swiss army knife among clinical trial designs? *Clinical Trials* 14(5): 417–424.

Bretz, F., Koenig, F., Brannath, W., Glimm, E., Posch, M. (2009). Adaptive designs for confirmatory clinical trials. *Statistics in Medicine* 28(8): 1181–1217.

Chen, Y., DeMets, D. L., Gordon Lan, K. (2004). Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine* 23(7): 1023–1038.

Chen, Y. J., Li, C., Lan, K. G. (2015). Sample size adjustment based on promising interim results and its application in confirmatory clinical trials. *Clinical Trials* 12(6): 584–595.

Cheng, B., Chow, S. C. (2010). On flexibility of adaptive designs and criteria for choosing a good one –a discussion of fda draft guidance. *Journal of Biopharmaceutical Statistics* 20(6): 1171–1177.

Chiu, S. T., Liu, J. P., Chow, S. C. (2014). Applications of the bayesian prior information to evaluation of equivalence of similar biological medicinal products. *Journal of Biopharmaceutical Statistics* 24(6): 1254–1263.

Chow, S. C. (2013). *Biosimilars: Design and Analysis of Follow-on Biologics*. Boca Raton, FL: CRC Press.

Chow, S. C., Lu, Q., Tse, S. K., Chi, E. (2009). Statistical methods for assessment of biosimilarity using biomarker data. *Journal of Biopharmaceutical Statistics* 20(1): 90–105.

Chow, S. C., Yang, L. Y., Starr, A., Chiu, S. T. (2013). Statistical methods for assessing interchangeability of biosimilars. *Statistics in Medicine* 32(3): 442–448.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65(1): 141–151.

Collett, D. (2015). *Modelling Survival Data in Medical Research (3rd edn.)*. Boca Raton, FL: CRC Press.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* 34(2): 187–220.

Di Scala, L., Glimm, E. (2011). Time-to-event analysis with treatment arm selection at interim. *Statistics in Medicine* 30(26): 3067–3081.

DiMasi, J. A., Grabowski, H. G. (2007). Economics of new oncology drug development. *Journal of Clinical Oncology* 25(2): 209–216.

Dmitrienko, A., Wang, M. D. (2006). Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in Medicine* 25(13): 2178–2195.

Eisenhauer, E., Therasse, P., Bogaerts, J., et al. (2009). New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European Journal of Cancer* 45(2): 228–247.

European Medicines Agency (2010). Guideline on the investigation of bioequivalence. *European Medicines Agency*.

European Medicines Agency (2014a). Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: non-clinical and clinical issues. *European Medicines Agency*.

European Medicines Agency (2014b). Guideline on the investigation of subgroups in confirmatory clinical trials. *European Medicines Agency*.

Felson, D. T., Anderson, J. J., Boers, M., et al. (1995). American college of rheumatology preliminary definition of improvement in rheumatoid arthritis. *Arthritis & Rheumatism* 38(6): 727–735.

Fine, J. P., Jiang, H., Chappell, R. (2001). On semi-competing risks data. *Biometrika* 88(4): 907–919.

Fleischer, F., Gaschler-Markefski, B., Bluhmki, E. (2009). A statistical model for the dependence between progression-free survival and overall survival. *Statistics in Medicine* 28(21): 2669–2686.

Fleming, T. R., Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Hoboken, NJ: John Wiley & Sons.

Fleming, T. R., Rothmann, M. D., Lu, H. L. (2009). Issues in using progression-free survival when evaluating oncology products. *Journal of Clinical Oncology* 27(17): 2874–2880.

Food and Drug Administration (2001). Guidance for industry: statistical approaches to establishing bioequivalence. *U.S. Food and Drug Administration*.

Food and Drug Administration (2007). Guidance for industry: clinical trial endpoints for the approval of cancer drugs and biologics. *U.S. Food and Drug Administration*.

Food and Drug Administration (2010). Guidance for industry: adaptive design clinical trials for drugs and biologics, draft. *U.S. Food and Drug Administration*.

Food and Drug Administration (2012). Guidance for industry: enrichment strategies for clinical trials to support approval of human drugs and biological products. *U.S. Food and Drug Administration*.

Food and Drug Administration (2015). Guidance for industry: scientific considerations in demonstrating biosimilarity to a reference product. *U.S. Food and Drug Administration*.

Food and Drug Administration (2016a). Guidance for industry and food and drug administration staff: adaptive designs for medical device clinical studies. *U.S. Food and Drug Administration*.

Food and Drug Administration (2016b). Guidance for industry: non-inferiority clinical trials to establish effectiveness. *U.S. Food and Drug Administration*.

Friede, T., Parsons, N., Stallard, N. (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine* 31(30): 4309–4320.

Fujitani, K., Yang, H, K., Mizusawa, J., et al. (2016). Gastrectomy plus chemotherapy versus chemotherapy alone for advanced gastric cancer with a single non-curable factor (REGATTA): a phase 3, randomised controlled trial. *Lancet Oncology* 17(3): 309–318.

Gallo, P., Mao, L., Shih, V. H. (2014). Alternative views on setting clinical trial futility criteria. *Journal of Biopharmaceutical Statistics* 24(5): 976–993.

Gao, P., Ware, J. H., Mehta, C. (2008). Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics* 18(6): 1184–1196.

Goteti, S., Hirawat, S., Massacesi, C., Fretault, N., Bretz, F., Dharan, B. (2014). Some practical considerations for phase III studies with biomarker evaluations. *Journal of Clinical Oncology* 32(8): 854–855.

Götte, H., Donica, M., Mordenti, G. (2015). Improving probabilities of correct interim decision in population enrichment designs. *Journal of Biopharmaceutical Statistics* 25(5): 1020–1038.

Graf, A. C., Posch, M., Koenig, F. (2015). Adaptive designs for subpopulation analysis optimizing utility functions. *Biometrical Journal* 57(1): 76–89.

Gregorc, V., Novello, S., Lazzari, C., et al. (2014). Predictive value of a proteomic signature in patients with non-small-cell lung cancer treated with second-line erlotinib or chemotherapy (PROSE): a biomarker-stratified, randomised phase 3 trial. *Lancet Oncology* 15(7): 713–721.

Harrison, R. K. (2016). Phase II and phase III failures: 2013-2015. *Nature Reviews Drug Discovery* 15(12): 817–818.

Hasegawa, T. (2014). Sample size determination for the weighted log-rank test with the Fleming–Harrington class of weights in cancer vaccine studies. *Pharmaceutical Statistics* 13(2): 128–135.

Hasegawa, T. (2016). Group sequential monitoring based on the weighted log-rank test statistic with the Fleming–Harrington class of weights in cancer vaccine studies. *Pharmaceutical Statistics* 15(5): 412–419.

Hayes, D. F., Thor, A. D., Dressler, L. G., et al. (2007). HER2 and response to paclitaxel in node-positive breast cancer. *New England Journal of Medicine* 357(15): 1496–1506.

Hsieh, T. C., Chow, S. C., Liu, J. P., Hsiao, C. F., Chi, E. (2009). Statistical test for evaluation of biosimilarity in variability of follow-on biologics. *Journal of Biopharmaceutical Statistics* 20(1): 75–89.

Hsieh, T. C., Chow, S. C., Yang, L. Y., Chi, E. (2013). The evaluation of biosimilarity index based on reproducibility probability for assessing follow-on biologics. *Statistics in Medicine* 32(3): 406–414.

Hua, S. Y., Xu, S., D'Agostino, R. B. (2015). Multiplicity adjustments in testing for bioequivalence. *Statistics in Medicine* 34(2): 215–231.

Irwin, J. (1949). The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Journal of Hygiene* 47(2): 188–189.

Jenkins, M., Stone, A., Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* 10(4): 347–356.

Jennison, C., Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC.

Jennison, C., Turnbull, B. W. (2015). Adaptive sample size modification in clinical trials: start small then ask for more? *Statistics in Medicine* 29(34): 3793–3810.

Kang, S. H., Chow, S. C. (2013). Statistical assessment of biosimilarity based on relative distance between follow-on biologics. *Statistics in Medicine* 32(3): 382–392.

Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282): 457–481.

Karapetis, C. S., Khambata Ford, S., Jonker, D. J., et al. (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine* 359(17): 1757–1765.

Karrison, T. (1987). Restricted mean life with adjustment for covariates. *Journal of the American Statistical Association* 82(400): 1169–1176.

Kay, R., Wu, J., Wittes, J. (2011). On assessing the presence of evaluation-time bias in progression-free survival in randomized trials. *Pharmaceutical Statistics* 10(3): 213–217.

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G., Scheike, T. H. (2013). *Handbook of Survival Analysis*. Boca Raton, FL: CRC Press.

Krisam, J., Kieser, M. (2014). Decision rules for subgroup selection based on a predictive biomarker. *Journal of Biopharmaceutical Statistics* 24(1): 188–202.

Lehmacher, W., Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* 55(4): 1286–1290.

Li, Y., Liu, Q., Wood, P., Johri, A. (2013). Statistical considerations in biosimilar clinical efficacy trials with asymmetrical margins. *Statistics in Medicine* 32(3): 393–405.

Liao, J. J., Darken, P. F. (2013). Comparability of critical quality attributes for establishing biosimilarity. *Statistics in Medicine* 32(3): 462–469.

Lu, Y., Zhang, Z. Z., Chow, S. C. (2014). Frequency estimator for assessing of follow-on biologics. *Journal of Biopharmaceutical Statistics* 24(6): 1280–1297.

Maini, R., St Clair, E. W., Breedveld, F., et al. (1999). Infliximab (chimeric anti-tumour necrosis factor $\alpha$ monoclonal antibody) versus placebo in rheumatoid arthritis patients receiving concomitant methotrexate: a randomised phase III trial. *Lancet* 354(9194): 1932–1939.

Marcus, R., Peritz, E., Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 63(3): 655–660.

Martin, L., Hutchens, M., Hawkins, C., Radnov, A. (2017). How much do clinical trials cost? *Nature Reviews Drug Discovery* 16(6): 381–382.

Mehta, C. R., Pocock, S. J. (2011). Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine* 30(28): 3267–3284.

Mok, T. S., Wu, Y. L., Thongprasert, S., et al. (2009). Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. *New England Journal of Medicine* 361(10): 947–957.

Montague, T. H., Potvin, D., DiLiberti, C. E., Hauck, W. W., Parr, A. F., Schuirmann, D. J. (2012). Additional results for 'Sequential design approaches for bioequivalence studies with crossover designs'. *Pharmaceutical Statistics* 11(1): 8–13.

Müller, H. H., Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 57(3): 886–891.

Nagasaki, M., Ando, Y. (2014). Clinical development and trial design of biosimilar products: a Japanese perspective. *Journal of Biopharmaceutical Statistics* 24(6): 1165–1172.

Ohashi, Y., Hamada, C., Uozumi, R. (2016). *Advanced Survival Analysis: Biostatistics Using SAS (in Japanese)*. Tokyo: University of Tokyo Press.

Pan, H., Yuan, Y., Xia, J. (2017). A calibrated power prior approach to borrow information from historical data with application to biosimilar clinical trials. *Journal of the Royal Statistical Society: Series C* 66(5): 979–996.

Park, J. W., Liu, M. C., Yee, D., et al. (2016). Adaptive randomization of neratinib in early breast cancer. *New England Journal of Medicine* 375(1): 11–22.

Park, W., Hrycaj, P., Jeka, S., et al. (2013). A randomised, double-blind, multicentre, parallel-group, prospective study comparing the pharmacokinetics, safety, and efficacy of CT-P13 and innovator infliximab in patients with ankylosing spondylitis: the PLANETAS study. *Annals of the Rheumatic Diseases* 72(10): 1605–1612.

Potvin, D., DiLiberti, C. E., Hauck, W. W., Parr, A. F., Schuirmann, D. J., Smith, R. A. (2008). Sequential design approaches for bioequivalence studies with crossover designs. *Pharmaceutical Statistics* 7(4): 245–62.

Proschan, M. A., Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* 51(6): 1315–1324.

Rémuzat, C., Dorey, J., Cristeau, O., Ionescu, D., Radière, G., Toumi, M. (2017). Key drivers for market penetration of biosimilars in europe. *Journal of Market Access & Health Policy* 5(1): 1–15.

Rosenstock, J., Hollander, P., Bhargava, A., et al. (2015). Similar efficacy and safety of LY2963016 insulin glargine and insulin glargine (lantus®) in patients with type 2 diabetes who were insulin-naïve or previously treated with insulin glargine: a randomized, double-blind controlled trial (the ELEMENT 2 study). *Diabetes, Obesity and Metabolism* 17(8): 734–741.

Rothmann, M. D., Wiens, B. L., Chan, I. S. (2011). *Design and Analysis of Non-Inferiority Trials*. Boca Raton, FL: CRC Press.

Royston, P., Parmar, M. K. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine* 30(19): 2409–2421.

Rugo, H. S., Olopade, O. I., DeMichele, A., et al. (2016). Adaptive randomization of veliparib–carboplatin treatment in breast cancer. *New England Journal of Medicine* 375(1): 23–34.

Saad, E., Katz, A., Hoff, P., Buyse, M. (2010). Progression-free survival as surrogate and as true end point: insights from the breast and colorectal cancer literature. *Annals of Oncology* 21(1): 7–12.

Saad, E. D., Buyse, M. (2012). Overall survival: patient outcome, therapeutic objective, clinical trial end point, or public health measure? *Journal of Clinical Oncology* 30(15): 1750–1754.

Samuel-Cahn, E. (1996). Is the Simes improved Bonferroni procedure conservative? *Biometrika* 83(4): 928–933.

Sarkar, S. K., Chang, C. K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 92(440): 1601–1608.

Satoh, T., Xu, R. H., Chung, H. C., et al. (2014). Lapatinib plus paclitaxel versus paclitaxel alone in the second-line treatment of HER2-amplified advanced gastric cancer in Asian populations: TyTAN–a randomized, phase III study. *Journal of Clinical Oncology* 32(19): 2039–2049.

Schellekens, H., Lietzan, E., Faccin, F., Venema, J. (2015). Biosimilar monoclonal antibodies: the scientific basis for extrapolation. *Expert Opinion on Biological Therapy* 15(11): 1–14.

Schemper, M. (1992). Cox analysis of survival data with non-proportional hazard functions. *The Statistician* 41(4): 455–465.

Schmidli, H., Bretz, F., Racine, A., Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal* 48(4): 635–643.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* 15(6): 657–680.

Seymour, L., Ivy, S. P., Sargent, D., et al. (2010). The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the National Cancer Institute Investigational Drug Steering Committee. *Clinical Cancer Research* 16(6): 1764–1769.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73(3): 751–754.

Simon, R. (2013). *Genomic Clinical Trials and Predictive Medicine*. Cambridge University Press.

Solomon, B. J., Mok, T., Kim, D. W., et al. (2014). First-line crizotinib versus chemotherapy in ALK-positive lung cancer. *New England Journal of Medicine* 371(23): 2167–2177.

Sozu, T., Sugimoto, T., Hamasaki, T. (2012). Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. *Biometrical Journal* 54(5): 716–729.

Spiegelhalter, D. J., Freedman, L. S., Blackburn, P. R. (1986). Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials* 7(1): 8–17.

Spiessens, B., Debois, M. (2010). Adjusted significance levels for subgroup analyses in clinical trials. *Contemporary Clinical Trials* 31(6): 647–656.

Stallard, N., Hamborg, T., Parsons, N., Friede, T. (2014). Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of Biopharmaceutical Statistics* 24(1): 168–187.

Stebbing, J., Baranau, Y., Baryash, V., et al. (2017). CT-P6 compared with reference trastuzumab for HER2-positive breast cancer: a randomised, double-blind, active-controlled, phase 3 equivalence trial. *Lancet Oncology* 18(7): 917–928.

Takeuchi, T., Yamanaka, H., Tanaka, Y., et al. (2015). Evaluation of the pharmacokinetic equivalence and 54-week efficacy and safety of CT-P13 and innovator infliximab in Japanese patients with rheumatoid arthritis. *Modern Rheumatology* 25(6): 817–824.

Trinquart, L., Jacot, J., Conner, S. C., Porcher, R. (2016). Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology* 34(15): 1813–1819.

Uno, H., Claggett, B., Tian, L., et al. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology* 32(22): 2380–2385.

Uozumi, R., Hamada, C. (2017a). Adaptive seamless design for establishing pharmacokinetic and efficacy equivalence in developing biosimilars. *Therapeutic Innovation & Regulatory Science* 51(6): 761–769.

Uozumi, R., Hamada, C. (2017b). Interim decision-making strategies in adaptive designs for population selection using time-to-event endpoints. *Journal of Biopharmaceutical Statistics* 27(1): 84–100.

Uozumi, R., Hamada, C. Utility-based interim decision rule planning in adaptive population selection design with survival endpoints. (submitted).

Wang, S. J., James Hung, H., O'Neill, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal* 51(2): 358–374.

Wang, Y. M. C., Chow, A. T. (2009). Development of biosimilars–pharmacokinetic and pharmacodynamic considerations. *Journal of Biopharmaceutical Statistics* 20(1): 46–61.

Wiens, B. L. (2003). A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics* 2(3): 211–215.

Wittes, J. (2010). Comments on the FDA draft guidance on adaptive designs. *Journal of Biopharmaceutical Statistics* 20(6): 1166–1170.

Wynne, C., Altendorfer, M., Sonderegger, I., et al. (2016). Bioequivalence, safety and immunogenicity of BI 695501, an adalimumab biosimilar candidate, compared with the reference biologic in a randomized, double-blind, active comparator phase I clinical study (VOLTAIRE®-PK) in healthy subjects. *Expert opinion on investigational drugs* 25(12): 1361–1370.

Xu, J., Audet, C., DiLiberti, C. E., et al. (2016). Optimal adaptive sequential designs for crossover bioequivalence studies. *Pharmaceutical statistics* 15(1): 15–27.

Xu, J., Kalbfleisch, J. D., Tai, B. (2010). Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics* 66(3): 716–725.

Yang, J., Zhang, N., Chow, S. C., Chi, E. (2013). An adapted f-test for homogeneity of variability in follow-on biological products. *Statistics in Medicine* 32(3): 415–423.

Yang, L. Y., Lai, C. H. (2014). Estimation and approximation approaches for biosimilar index based on reproducibility probability. *Journal of Biopharmaceutical Statistics* 24(6): 1298–1311.

Yoo, D. H., Hrycaj, P., Miranda, P., et al. (2013). A randomised, double-blind, parallel-group study to demonstrate equivalence in efficacy and safety of CT-P13 compared with innovator infliximab when coadministered with methotrexate in patients with active rheumatoid arthritis: the PLANETRA study. *Annals of the Rheumatic Diseases* 72(10): 1613–1620.

Zhang, L., Ko, C. W., Tang, S., Sridhara, R. (2013). Relationship between progression-free survival and overall survival benefit a simulation study. *Therapeutic Innovation & Regulatory Science* 47(1): 95–100.

Zhang, N., Yang, J., Chow, S. C., Chi, E. (2014). Nonparametric tests for evaluation of biosimilarity in variability of follow-on biologics. *Journal of biopharmaceutical statistics* 24(6): 1239–1253.

Zhang, N., Yang, J., Chow, S. C., Endrenyi, L., Chi, E. (2013). Impact of variability on the choice of biosimilarity limits in assessing follow-on biologics. *Statistics in Medicine* 32(3): 424–433.

Zheng, C., Zhao, L., Wang, J. (2015). Modifications of sequential designs in bioequivalence trials. *Pharmaceutical Statistics* 14(3): 180–188.

Zucker, D. M. (1998). Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of the American Statistical Association* 93(442): 702–709.

# Appendices

# Appendix A

# Derivation of the correlation model between OS and PFS by Fleischer et al. (2009) in Chapter 3

## A.1 Correlation between OS and PFS with maximal independence

In Chapter 3, we considered the correlation between OS and PFS. One approach is to assume exponentially distributed TTP and OS considering a model that offers maximal independence with respect to OS and PFS. Suppose that each survival endpoint $b \in \{\mathrm{TTP}, \mathrm{OS}\}$ is exponentially distributed with parameter $\lambda_b$ where $\lambda_b > 0$. Because PFS is given by the minimum of TTP and OS, it holds that

$$\begin{aligned} P(\min(\mathrm{TTP}, \mathrm{OS}) < t) &= 1 - P(\min(\mathrm{TTP}, \mathrm{OS}) \geq t) \\ &= 1 - P(\mathrm{TTP} \geq t) \cdot P(\mathrm{OS} \geq t) \\ &= 1 - \exp\{-(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{OS}})t\} \end{aligned}$$

based on the assumption that TTP and OS are completely independent. Thus, PFS is exponentially distributed with parameter $\lambda_{\mathrm{OS}} + \lambda_{\mathrm{TTP}}$.

Let $q_1 = P(\mathrm{PFS} < \mathrm{OS}) = \lambda_{\mathrm{TTP}}/\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{OS}}$ denote the probability that a progression occurs before death (Case 1 in Figure 3.1). Then, the expectation denoted by $E[\mathrm{OS} \cdot \mathrm{PFS}]$ is given by

$$E[\mathrm{OS} \cdot \mathrm{PFS}] = E[\mathrm{OS} \cdot \mathrm{PFS} \mid \mathrm{PFS} < \mathrm{OS}]q_1 + E[\mathrm{OS} \cdot \mathrm{PFS} \mid \mathrm{PFS} = \mathrm{OS}](1 - q_1) \tag{A.1}$$

with OS and PFS being independent. Hence, each expectation can be written as

$$\begin{aligned} E[\mathrm{OS} \cdot \mathrm{PFS} \mid \mathrm{PFS} < \mathrm{OS}] &= E[\mathrm{OS} \cdot \mathrm{PFS} + \mathrm{PFS}^2] \\ &= E[\mathrm{OS}]E[\mathrm{PFS}] + E[\mathrm{PFS}^2] \\ &= \frac{1}{\lambda_{\mathrm{OS}}(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{OS}})} + \frac{2}{(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{OS}})^2} \end{aligned}$$

and

$$E[\text{OS} \cdot \text{PFS} \mid \text{PFS} = \text{OS}] \;=\; E[\text{OS}^2] = \frac{2}{(\lambda_{\text{TTP}} + \lambda_{\text{OS}})^2}.$$

Therefore, Equation (A.1) is rewritten as

$$\begin{aligned}
E[\text{OS} \cdot \text{PFS}] &= E[\text{OS} \cdot \text{PFS} \mid \text{PFS} < \text{OS}]q_1 + E[\text{OS} \cdot \text{PFS} \mid \text{PFS} = \text{OS}](1 - q_1) \\
&= \left(\frac{1}{\lambda_{\text{OS}}(\lambda_{\text{TTP}} + \lambda_{\text{OS}})} + \frac{2}{(\lambda_{\text{TTP}} + \lambda_{\text{OS}})^2}\right)\frac{\lambda_{\text{TTP}}}{\lambda_{\text{TTP}} + \lambda_{\text{OS}}} \\
&\quad + \left(\frac{2}{(\lambda_{\text{TTP}} + \lambda_{\text{OS}})^2}\right)\frac{\lambda_{\text{OS}}}{\lambda_{\text{TTP}} + \lambda_{\text{OS}}} \\
&= \frac{1}{\lambda_{\text{TTP}} + \lambda_{\text{OS}}}\left(\frac{\lambda_{\text{TTP}}}{\lambda_{\text{OS}}(\lambda_{\text{TTP}} + \lambda_{\text{OS}})} + \frac{2}{\lambda_{\text{TTP}} + \lambda_{\text{OS}}}\right) \\
&= \frac{2\lambda_{\text{OS}} + \lambda_{\text{TTP}}}{\lambda_{\text{OS}}(\lambda_{\text{TTP}} + \lambda_{\text{OS}})^2}. \qquad \text{(A.2)}
\end{aligned}$$

Using Equation (A.2), the covariate of OS and PFS can be expressed by

$$\begin{aligned}
\text{Cov}[\text{OS}, \text{PFS}] &= E[\text{OS} \cdot \text{PFS}] - E[\text{OS}]E[\text{PFS}] \\
&= \frac{2\lambda_{\text{OS}} + \lambda_{\text{TTP}}}{\lambda_{\text{OS}}(\lambda_{\text{TTP}} + \lambda_{\text{OS}})^2} - \frac{1}{\lambda_{\text{OS}}(\lambda_{\text{TTP}} + \lambda_{\text{OS}})} \\
&= \frac{1}{(\lambda_{\text{TTP}} + \lambda_{\text{OS}})^2}. \qquad \text{(A.3)}
\end{aligned}$$

Therefore, the correlation coefficient between OS and PFS can be represented as follows:

$$\begin{aligned}
\rho = \text{Corr}[\text{OS}, \text{PFS}] &= \frac{\text{Cov}[\text{OS}, \text{PFS}]}{\sqrt{V[\text{OS}]V[\text{PFS}]}} \\
&= \frac{1}{(\lambda_{\text{TTP}} + \lambda_{\text{OS}})^2}\lambda_{\text{OS}}(\lambda_{\text{TTP}} + \lambda_{\text{OS}}) \\
&= \frac{\lambda_{\text{OS}}}{\lambda_{\text{TTP}} + \lambda_{\text{OS}}}. \qquad \text{(A.4)}
\end{aligned}$$

## A.2 A more general model considering dependencies between OS and TTP

In Chapter 3, we considered the length of PPS when using the correlation between OS and PFS. Thus, we did not apply the model in Equation (A.4). Instead, another approach is to consider dependency between OS and TTP in a more general model. That is, PPS effect is considered. Assume the survival endpoint D, namely, the time to death without tumor progression. Then, OS is calculated as follows:

$$\text{OS} = \begin{cases} \text{PFS} & if \quad \text{PFS} \neq \text{TTP} \\ \text{TTP} + \text{PPS} & otherwise \end{cases}.$$

Suppose that each survival endpoint $a \in \{\mathrm{TTP}, \mathrm{PPS}, \mathrm{D}\}$ is exponentially distributed with parameter $\lambda_a$ where $\lambda_a > 0$. Then, the expectation $E[\mathrm{OS}]$ is given by

$$\begin{aligned}
E[\mathrm{OS}] &= E[\mathrm{OS} \mid \mathrm{PFS} < \mathrm{OS}]q_1 + E[\mathrm{OS} \mid \mathrm{PFS} = \mathrm{OS}](1 - q_1) \\
&= \left(\frac{1}{\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}} + \frac{1}{\lambda_{\mathrm{PPS}}}\right)\frac{\lambda_{\mathrm{TTP}}}{\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}} + \frac{1}{\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}}\frac{\lambda_{\mathrm{D}}}{\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}} \\
&= \frac{\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{PPS}}}{\left(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}\right)\lambda_{\mathrm{PPS}}}
\end{aligned}$$

and the expectation $E[\mathrm{OS}^2]$ is given by

$$\begin{aligned}
E[\mathrm{OS}^2] &= E[\mathrm{OS}^2 \mid \mathrm{PFS} < \mathrm{OS}]q_1 + E[\mathrm{OS}^2 \mid \mathrm{PFS} = \mathrm{OS}](1 - q_1) \\
&= E[(\mathrm{PFS} + \mathrm{PPS})^2 \mid \mathrm{PFS} < \mathrm{OS}]q_1 + E[\mathrm{PFS}^2 \mid \mathrm{PFS} = \mathrm{OS}](1 - q_1) \\
&= \Big(E[\mathrm{PFS}^2 \mid \mathrm{PFS} < \mathrm{OS}] + 2E[\mathrm{PFS} \cdot \mathrm{PPS} \mid \mathrm{PFS} < \mathrm{OS}] + E[\mathrm{PPS}^2 \mid \mathrm{PFS} < \mathrm{OS}]\Big)q_1 \\
&\quad + E[\mathrm{PFS}^2 \mid \mathrm{PFS} = \mathrm{OS}](1 - q_1) \\
&= \left(\frac{2}{(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}})^2} + \frac{2}{\lambda_{\mathrm{PPS}}^2} + \frac{2}{\left(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}\right)\lambda_{\mathrm{PPS}}}\right)\frac{\lambda_{\mathrm{TTP}}}{\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}} + \frac{2}{(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}})^2}\frac{\lambda_{\mathrm{D}}}{\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}} \\
&= \frac{2}{(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}})^2} + \frac{2\lambda_{\mathrm{TTP}}}{\left(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}\right)\lambda_{\mathrm{PPS}}^2} + \frac{2\lambda_{\mathrm{TTP}}}{\left(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}\right)^2\lambda_{\mathrm{PPS}}} \\
&= \frac{2\lambda_{\mathrm{TTP}}^2 + 2\lambda_{\mathrm{TTP}}\lambda_{\mathrm{D}} + 2\lambda_{\mathrm{TTP}}\lambda_{\mathrm{PPS}} + 2\lambda_{\mathrm{PPS}}^2}{\left(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}\right)^2\lambda_{\mathrm{PPS}}^2}.
\end{aligned}$$

Thus,

$$\begin{aligned}
V[\mathrm{OS}] &= E[\mathrm{OS}^2] - (E[\mathrm{OS}])^2 \\
&= \frac{2\lambda_{\mathrm{TTP}}^2 + 2\lambda_{\mathrm{TTP}}\lambda_{\mathrm{D}} + 2\lambda_{\mathrm{TTP}}\lambda_{\mathrm{PPS}} + 2\lambda_{\mathrm{PPS}}^2}{\left(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}\right)^2\lambda_{\mathrm{PPS}}^2} - \left(\frac{\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{PPS}}}{\left(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}\right)\lambda_{\mathrm{PPS}}}\right)^2 \\
&= \frac{\lambda_{\mathrm{TTP}}^2 + 2\lambda_{\mathrm{TTP}}\lambda_{\mathrm{D}} + \lambda_{\mathrm{PPS}}^2}{\left(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}\right)^2\lambda_{\mathrm{PPS}}^2}.
\end{aligned}$$

Because the expectation $E[\mathrm{OS} \cdot \mathrm{PFS}]$ can be rewitten as

$$\begin{aligned}
E[\mathrm{OS} \cdot \mathrm{PFS} \mid \mathrm{PFS} < \mathrm{OS}] &= E[\mathrm{PFS}]E[\mathrm{PPS}] + E[\mathrm{PFS}^2] \\
&= \frac{1}{(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}})\lambda_{\mathrm{PPS}}} + \frac{2}{(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}})^2},
\end{aligned}$$

it holds that

$$\begin{aligned}
E[\mathrm{OS} \cdot \mathrm{PFS}] &= E[\mathrm{OS} \cdot \mathrm{PFS} \mid \mathrm{PFS} < \mathrm{OS}]q_1 + E[\mathrm{OS} \cdot \mathrm{PFS} \mid \mathrm{PFS} = \mathrm{OS}](1 - q_1) \\
&= \left(\frac{1}{(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}})\lambda_{\mathrm{PPS}}} + \frac{2}{(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}})^2}\right)\frac{\lambda_{\mathrm{TTP}}}{\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}} \\
&\quad + \frac{2}{(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}})^2}\frac{\lambda_{\mathrm{D}}}{\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}} \\
&= \frac{2\lambda_{\mathrm{PPS}} + \lambda_{\mathrm{TTP}}}{\lambda_{\mathrm{PPS}}\left(\lambda_{\mathrm{TTP}} + \lambda_{\mathrm{D}}\right)^2},
\end{aligned}$$

and then

$$\begin{aligned}\mathrm{Cov}[\mathrm{OS},\mathrm{PFS}] &= E[\mathrm{OS}\cdot\mathrm{PFS}] - E[\mathrm{OS}]E[\mathrm{PFS}] \\ &= \frac{2\lambda_{\mathrm{PPS}}+\lambda_{\mathrm{TTP}}}{\lambda_{\mathrm{PPS}}\left(\lambda_{\mathrm{TTP}}+\lambda_{\mathrm{D}}\right)^2} - \frac{1}{\lambda_{\mathrm{TTP}}+\lambda_{\mathrm{D}}}\frac{\lambda_{\mathrm{TTP}}+\lambda_{\mathrm{PPS}}}{\left(\lambda_{\mathrm{TTP}}+\lambda_{\mathrm{D}}\right)\lambda_{\mathrm{PPS}}} \\ &= \frac{1}{\left(\lambda_{\mathrm{TTP}}+\lambda_{\mathrm{D}}\right)^2}. \end{aligned} \tag{A.5}$$

Note that Equation (A.5) is equivalent to Equation (A.3). Therefore, the correlation coefficient between OS and PFS can be represented as follows:

$$\begin{aligned}\rho = \mathrm{Corr}[\mathrm{OS},\mathrm{PFS}] &= \frac{\mathrm{Cov}[\mathrm{OS},\mathrm{PFS}]}{\sqrt{V[\mathrm{OS}]V[\mathrm{PFS}]}} \\ &= \frac{1}{\left(\lambda_{\mathrm{TTP}}+\lambda_{\mathrm{D}}\right)^2}\frac{\left(\lambda_{\mathrm{TTP}}+\lambda_{\mathrm{D}}\right)\lambda_{\mathrm{PPS}}}{\sqrt{\lambda_{\mathrm{TTP}}^2+2\lambda_{\mathrm{TTP}}\lambda_{\mathrm{D}}+\lambda_{\mathrm{PPS}}^2}}\left(\lambda_{\mathrm{TTP}}+\lambda_{\mathrm{D}}\right) \\ &= \frac{\lambda_{\mathrm{PPS}}}{\sqrt{\lambda_{\mathrm{TTP}}^2+2\lambda_{\mathrm{TTP}}\lambda_{\mathrm{D}}+\lambda_{\mathrm{PPS}}^2}}. \end{aligned} \tag{A.6}$$

Moreover, it is obtained that

$$\begin{aligned}\sqrt{\frac{V[\mathrm{PFS}]}{V[\mathrm{OS}]}} &= \frac{1}{\lambda_{\mathrm{TTP}}+\lambda_{\mathrm{D}}}\frac{\left(\lambda_{\mathrm{TTP}}+\lambda_{\mathrm{D}}\right)\lambda_{\mathrm{PPS}}}{\sqrt{\lambda_{\mathrm{TTP}}^2+2\lambda_{\mathrm{TTP}}\lambda_{\mathrm{D}}+\lambda_{\mathrm{PPS}}^2}} \\ &= \frac{\lambda_{\mathrm{PPS}}}{\sqrt{\lambda_{\mathrm{TTP}}^2+2\lambda_{\mathrm{TTP}}\lambda_{\mathrm{D}}+\lambda_{\mathrm{PPS}}^2}} \\ &= \mathrm{Corr}[\mathrm{OS},\mathrm{PFS}] = \rho. \end{aligned}$$

Furthermore, Equation (A.6) is represented as follows:

$$\begin{aligned}\rho = \mathrm{Corr}[\mathrm{OS},\mathrm{PFS}] &= \frac{\lambda_{\mathrm{PPS}}}{\sqrt{\lambda_{\mathrm{TTP}}^2+2\lambda_{\mathrm{TTP}}\lambda_{\mathrm{D}}+\lambda_{\mathrm{PPS}}^2}} \\ &\to \frac{\lambda_{\mathrm{OS}}}{\sqrt{\lambda_{\mathrm{TTP}}^2+2\lambda_{\mathrm{TTP}}\lambda_{\mathrm{OS}}+\lambda_{\mathrm{OS}}^2}} \\ &\to \frac{\lambda_{\mathrm{OS}}}{\lambda_{\mathrm{TTP}}+\lambda_{\mathrm{OS}}} \end{aligned}$$

when $\lambda_{\mathrm{D}} \to \lambda_{\mathrm{OS}}, \lambda_{\mathrm{PPS}} \to \lambda_{\mathrm{OS}}$. This equivalence of this model formulation can be seen by using the lack of memory property of the exponential distribution. In other words, the hazard for OS is constant, i.e., $\lambda_{\mathrm{OS}}$, before and after progression happens.

# Appendix B

# Details of simulation results in Chapter 3

Table B.1: Probabilities of selecting $F$ at the interim analysis using PFS only, OS only, or OS and PFS under the assumption that the effect of PPS is small and large based on 10,000 simulation replications per scenario.

| Model | Scenario | $HR_N^{\{a\}}$ | $\rho$ / $\tau$ | PPS effect is small | | | PPS effect is large | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PFS only | OS only | OS and PFS | PFS only | OS only | OS and PFS |
| (a) | 1 | 0.90 | 0.90 | 0.6060 | 0.6256 | 0.5683 | 0.6060 | 0.6256 | 0.5683 |
| | | | 0.70 | 0.6000 | 0.5920 | 0.5506 | 0.6163 | 0.6821 | **0.5561** |
| | | | 0.50 | 0.6011 | 0.5881 | 0.5735 | 0.6184 | 0.6529 | **0.5228** |
| | 2 | 1.00 | 0.90 | 0.4103 | 0.4095 | 0.3647 | 0.4103 | 0.4095 | 0.3647 |
| | | | 0.70 | 0.4161 | 0.4151 | 0.3655 | 0.4117 | 0.4088 | **0.3176** |
| | | | 0.50 | 0.4139 | 0.4161 | 0.3847 | 0.4061 | 0.4096 | **0.2789** |
| | 3 | 1.11 | 0.90 | 0.2378 | 0.2186 | 0.1930 | 0.2378 | 0.2186 | 0.1930 |
| | | | 0.70 | 0.2407 | 0.2540 | 0.2022 | 0.2337 | 0.1778 | **0.1377** |
| | | | 0.50 | 0.2409 | 0.2518 | 0.2191 | 0.2316 | 0.2016 | **0.1181** |
| | 4 | 1.43 | 0.90 | 0.0292 | 0.0176 | 0.0165 | 0.0292 | 0.0176 | 0.0165 |
| | | | 0.70 | 0.0269 | 0.0358 | 0.0193 | 0.0268 | 0.0045 | 0.0040 |
| | | | 0.50 | 0.0285 | 0.0366 | 0.0244 | 0.0252 | 0.0103 | 0.0037 |
| (b) | 1 | 0.90 | 0.90 | 0.6140 | 0.6095 | 0.5999 | 0.6131 | 0.6034 | 0.5865 |
| | | | 0.70 | 0.6258 | 0.6158 | 0.6048 | 0.6214 | 0.5886 | **0.5428** |
| | | | 0.50 | 0.6274 | 0.6303 | 0.6053 | 0.6280 | 0.5687 | **0.5050** |
| | 2 | 1.00 | 0.90 | 0.4107 | 0.4104 | 0.3957 | 0.4095 | 0.4082 | 0.3856 |
| | | | 0.70 | 0.4130 | 0.4131 | 0.3926 | 0.4181 | 0.4081 | **0.3452** |
| | | | 0.50 | 0.4222 | 0.4192 | 0.3967 | 0.4236 | 0.4117 | **0.3129** |
| | 3 | 1.11 | 0.90 | 0.2299 | 0.2338 | 0.2185 | 0.2287 | 0.2356 | 0.2103 |
| | | | 0.70 | 0.2257 | 0.2301 | 0.2100 | 0.2272 | 0.2438 | **0.1764** |
| | | | 0.50 | 0.2355 | 0.2298 | 0.2112 | 0.2391 | 0.2619 | **0.1601** |
| | 4 | 1.43 | 0.90 | 0.0228 | 0.0265 | 0.0217 | 0.0223 | 0.0273 | 0.0195 |
| | | | 0.70 | 0.0184 | 0.0202 | 0.0155 | 0.0236 | 0.0283 | 0.0146 |
| | | | 0.50 | 0.0191 | 0.0187 | 0.0151 | 0.0247 | 0.0414 | 0.0099 |

The bold numbers show the situations for which a greater probability of selecting each population is expected when using both OS and PFS.

Table B.1: continued.

| Model | Scenario | $HR_N^{\{a\}}$ | $\rho$ / $\tau$ | PPS effect is small | | | PPS effect is large | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PFS only | OS only | OS and PFS | PFS only | OS only | OS and PFS |
| (c) | 1 | 0.90 | 0.90 | 0.6122 | 0.6072 | 0.6034 | 0.6147 | 0.6029 | 0.5935 |
| | | | 0.70 | 0.6251 | 0.6118 | 0.6053 | 0.6309 | 0.5771 | **0.5348** |
| | | | 0.50 | 0.6264 | 0.6246 | 0.6019 | 0.6374 | 0.5640 | **0.5096** |
| | 2 | 1.00 | 0.90 | 0.4072 | 0.4083 | 0.3995 | 0.4110 | 0.4093 | 0.3927 |
| | | | 0.70 | 0.4112 | 0.4174 | 0.3957 | 0.4161 | 0.4110 | **0.3379** |
| | | | 0.50 | 0.4197 | 0.4172 | 0.3895 | 0.4232 | 0.4069 | **0.3170** |
| | 3 | 1.11 | 0.90 | 0.2271 | 0.2341 | 0.2224 | 0.2280 | 0.2344 | 0.2169 |
| | | | 0.70 | 0.2279 | 0.2295 | 0.2153 | 0.2239 | 0.2573 | **0.1737** |
| | | | 0.50 | 0.2315 | 0.2312 | 0.2055 | 0.2322 | 0.2614 | **0.1576** |
| | 4 | 1.43 | 0.90 | 0.0228 | 0.0260 | 0.0223 | 0.0220 | 0.0277 | 0.0210 |
| | | | 0.70 | 0.0178 | 0.0208 | 0.0162 | 0.0175 | 0.0388 | 0.0126 |
| | | | 0.50 | 0.0193 | 0.0190 | 0.0150 | 0.0197 | 0.0434 | 0.0094 |
| (d) | 1 | 0.90 | 0.90 | 0.6156 | 0.6056 | 0.5880 | 0.6181 | 0.5995 | 0.5815 |
| | | | 0.70 | 0.6292 | 0.6129 | 0.5915 | 0.6181 | 0.6036 | **0.5500** |
| | | | 0.50 | 0.6363 | 0.6272 | 0.5973 | 0.6230 | 0.5885 | **0.5119** |
| | 2 | 1.00 | 0.90 | 0.4119 | 0.4099 | 0.3876 | 0.4118 | 0.4077 | 0.3772 |
| | | | 0.70 | 0.4143 | 0.4152 | 0.3804 | 0.4156 | 0.4136 | **0.3510** |
| | | | 0.50 | 0.4226 | 0.4214 | 0.3842 | 0.4224 | 0.4173 | **0.3158** |
| | 3 | 1.11 | 0.90 | 0.2305 | 0.2355 | 0.2127 | 0.2289 | 0.2412 | 0.2051 |
| | | | 0.70 | 0.2256 | 0.2305 | 0.1988 | 0.2324 | 0.2371 | **0.1780** |
| | | | 0.50 | 0.2350 | 0.2313 | 0.2011 | 0.2407 | 0.2539 | **0.1549** |
| | 4 | 1.43 | 0.90 | 0.0224 | 0.0263 | 0.0197 | 0.0240 | 0.0305 | 0.0203 |
| | | | 0.70 | 0.0168 | 0.0199 | 0.0135 | 0.0248 | 0.0247 | 0.0148 |
| | | | 0.50 | 0.0167 | 0.0194 | 0.0123 | 0.0262 | 0.0307 | 0.0085 |

The bold numbers show the situations for which a greater probability of selecting each population is expected when using both OS and PFS.

Table B.2: Probabilities of selecting $P$ at the interim analysis using PFS only, OS only, or OS and PFS under the assumption that the effect of PPS is small and large based on 10,000 simulation replications per scenario.

| Model | Scenario | $HR_N^{(a)}$ | $\rho$ / $\tau$ | PPS effect is small | | | PPS effect is large | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PFS only | OS only | OS and PFS | PFS only | OS only | OS and PFS |
| (a) | 1 | 0.90 | 0.90 | 0.3874 | 0.3709 | 0.4245 | 0.3874 | 0.3709 | 0.4245 |
| | | | 0.70 | 0.3929 | 0.3942 | 0.4382 | 0.3814 | 0.3177 | **0.4413** |
| | | | 0.50 | 0.3891 | 0.3980 | 0.4149 | 0.3799 | 0.3461 | **0.4747** |
| | 2 | 1.00 | 0.90 | 0.5811 | 0.5846 | 0.6258 | 0.5811 | 0.5846 | 0.6258 |
| | | | 0.70 | 0.5734 | 0.5668 | 0.6190 | 0.5851 | 0.5911 | **0.6788** |
| | | | 0.50 | 0.5738 | 0.5649 | 0.6015 | 0.5913 | 0.5890 | **0.7169** |
| | 3 | 1.11 | 0.90 | 0.7522 | 0.7743 | 0.7958 | 0.7522 | 0.7743 | 0.7958 |
| | | | 0.70 | 0.7462 | 0.7234 | 0.7803 | 0.7615 | 0.8221 | **0.8569** |
| | | | 0.50 | 0.7431 | 0.7253 | 0.7628 | 0.7650 | 0.7966 | **0.8773** |
| | 4 | 1.43 | 0.90 | 0.9581 | 0.9731 | 0.9699 | 0.9581 | 0.9731 | 0.9699 |
| | | | 0.70 | 0.9566 | 0.9349 | 0.9587 | 0.9672 | 0.9953 | 0.9896 |
| | | | 0.50 | 0.9503 | 0.9351 | 0.9523 | 0.9700 | 0.9875 | 0.9906 |
| (b) | 1 | 0.90 | 0.90 | 0.3820 | 0.3845 | 0.3956 | 0.3842 | 0.3906 | 0.4097 |
| | | | 0.70 | 0.3708 | 0.3787 | 0.3909 | 0.3766 | 0.3972 | **0.4497** |
| | | | 0.50 | 0.3691 | 0.3642 | 0.3902 | 0.3701 | 0.4002 | **0.4751** |
| | 2 | 1.00 | 0.90 | 0.5835 | 0.5820 | 0.5975 | 0.5862 | 0.5826 | 0.6085 |
| | | | 0.70 | 0.5825 | 0.5787 | 0.6016 | 0.5792 | 0.5749 | **0.6451** |
| | | | 0.50 | 0.5724 | 0.5732 | 0.5968 | 0.5736 | 0.5496 | **0.6630** |
| | 3 | 1.11 | 0.90 | 0.7620 | 0.7559 | 0.7723 | 0.7657 | 0.7528 | 0.7818 |
| | | | 0.70 | 0.7680 | 0.7591 | 0.7819 | 0.7694 | 0.7368 | **0.8125** |
| | | | 0.50 | 0.7571 | 0.7603 | 0.7803 | 0.7574 | 0.6928 | **0.8123** |
| | 4 | 1.43 | 0.90 | 0.9656 | 0.9581 | 0.9656 | 0.9692 | 0.9556 | 0.9693 |
| | | | 0.70 | 0.9726 | 0.9641 | 0.9735 | 0.9726 | 0.9484 | 0.9726 |
| | | | 0.50 | 0.9714 | 0.9679 | 0.9737 | 0.9710 | 0.9037 | 0.9593 |

The bold numbers show the situations for which a greater probability of selecting each population is expected when using both OS and PFS.

Table B.2: continued.

| Model | Scenario | $HR_N^{\{a\}}$ | $\rho$ / $\tau$ | PPS effect is small | | | PPS effect is large | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PFS only | OS only | OS and PFS | PFS only | OS only | OS and PFS |
| (c) | 1 | 0.90 | 0.90 | 0.3826 | 0.3872 | 0.3912 | 0.3812 | 0.3908 | 0.4019 |
| | | | 0.70 | 0.3717 | 0.3818 | 0.3904 | 0.3671 | 0.3994 | **0.4513** |
| | | | 0.50 | 0.3693 | 0.3704 | 0.3927 | 0.3613 | 0.3981 | **0.4677** |
| | 2 | 1.00 | 0.90 | 0.5861 | 0.5841 | 0.5936 | 0.5833 | 0.5820 | 0.6004 |
| | | | 0.70 | 0.5841 | 0.5731 | 0.5985 | 0.5815 | 0.5606 | **0.6455** |
| | | | 0.50 | 0.5743 | 0.5755 | 0.6029 | 0.5743 | 0.5481 | **0.6557** |
| | 3 | 1.11 | 0.90 | 0.7638 | 0.7549 | 0.7682 | 0.7640 | 0.7540 | 0.7739 |
| | | | 0.70 | 0.7652 | 0.7585 | 0.7761 | 0.7726 | 0.7099 | **0.8065** |
| | | | 0.50 | 0.7602 | 0.7587 | 0.7851 | 0.7640 | 0.6873 | **0.8101** |
| | 4 | 1.43 | 0.90 | 0.9654 | 0.9589 | 0.9653 | 0.9669 | 0.9556 | 0.9665 |
| | | | 0.70 | 0.9715 | 0.9627 | 0.9712 | 0.9783 | 0.9219 | 0.9650 |
| | | | 0.50 | 0.9692 | 0.9670 | 0.9722 | 0.9755 | 0.8957 | 0.9552 |
| (d) | 1 | 0.90 | 0.90 | 0.3814 | 0.3877 | 0.4077 | 0.3799 | 0.3930 | 0.4144 |
| | | | 0.70 | 0.3684 | 0.3812 | 0.4044 | 0.3792 | 0.3878 | **0.4441** |
| | | | 0.50 | 0.3595 | 0.3667 | 0.3975 | 0.3746 | 0.3925 | **0.4758** |
| | 2 | 1.00 | 0.90 | 0.5840 | 0.5815 | 0.6067 | 0.5853 | 0.5817 | 0.6169 |
| | | | 0.70 | 0.5825 | 0.5764 | 0.6137 | 0.5811 | 0.5740 | **0.6410** |
| | | | 0.50 | 0.5724 | 0.5702 | 0.6089 | 0.5746 | 0.5590 | **0.6698** |
| | 3 | 1.11 | 0.90 | 0.7643 | 0.7524 | 0.7795 | 0.7672 | 0.7452 | 0.7871 |
| | | | 0.70 | 0.7700 | 0.7589 | 0.7934 | 0.7634 | 0.7476 | **0.8121** |
| | | | 0.50 | 0.7589 | 0.7583 | 0.7905 | 0.7558 | 0.7185 | **0.8283** |
| | 4 | 1.43 | 0.90 | 0.9696 | 0.9566 | 0.9696 | 0.9707 | 0.9515 | 0.9692 |
| | | | 0.70 | 0.9766 | 0.9640 | 0.9757 | 0.9706 | 0.9536 | 0.9729 |
| | | | 0.50 | 0.9755 | 0.9654 | 0.9764 | 0.9695 | 0.9349 | 0.9730 |

The bold numbers show the situations for which a greater probability of selecting each population is expected when using both OS and PFS.

Table B.3: Probabilities of discontinuing the trial for futility at the interim analysis using PFS only, OS only, or OS and PFS under the assumption that the effect of PPS is small and large based on 10,000 simulation replications per scenario.

| Model | Scenario | $HR_N^{\{a\}}$ | $\rho / \tau$ | PPS effect is small | | | PPS effect is large | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PFS only | OS only | OS and PFS | PFS only | OS only | OS and PFS |
| (a) | 1 | 0.90 | 0.90 | 0.0066 | 0.0035 | 0.0072 | 0.0066 | 0.0035 | 0.0072 |
| | | | 0.70 | 0.0071 | 0.0138 | 0.0112 | 0.0023 | 0.0002 | 0.0026 |
| | | | 0.50 | 0.0098 | 0.0139 | 0.0116 | 0.0017 | 0.0010 | 0.0025 |
| | 2 | 1.00 | 0.90 | 0.0086 | 0.0059 | 0.0095 | 0.0086 | 0.0059 | 0.0095 |
| | | | 0.70 | 0.0105 | 0.0181 | 0.0155 | 0.0032 | 0.0001 | 0.0036 |
| | | | 0.50 | 0.0123 | 0.0190 | 0.0138 | 0.0026 | 0.0014 | 0.0042 |
| | 3 | 1.11 | 0.90 | 0.0100 | 0.0071 | 0.0112 | 0.0100 | 0.0071 | 0.0112 |
| | | | 0.70 | 0.0131 | 0.0226 | 0.0175 | 0.0048 | 0.0001 | 0.0054 |
| | | | 0.50 | 0.0160 | 0.0229 | 0.0181 | 0.0034 | 0.0018 | 0.0046 |
| | 4 | 1.43 | 0.90 | 0.0127 | 0.0093 | 0.0136 | 0.0127 | 0.0093 | 0.0136 |
| | | | 0.70 | 0.0165 | 0.0293 | 0.0220 | 0.0060 | 0.0002 | 0.0064 |
| | | | 0.50 | 0.0212 | 0.0283 | 0.0233 | 0.0048 | 0.0022 | 0.0057 |
| (b) | 1 | 0.90 | 0.90 | 0.0040 | 0.0060 | 0.0045 | 0.0027 | 0.0060 | 0.0038 |
| | | | 0.70 | 0.0034 | 0.0055 | 0.0043 | 0.0020 | 0.0142 | 0.0075 |
| | | | 0.50 | 0.0035 | 0.0055 | 0.0045 | 0.0019 | 0.0311 | 0.0199 |
| | 2 | 1.00 | 0.90 | 0.0058 | 0.0076 | 0.0068 | 0.0043 | 0.0092 | 0.0059 |
| | | | 0.70 | 0.0045 | 0.0082 | 0.0058 | 0.0027 | 0.0170 | 0.0097 |
| | | | 0.50 | 0.0054 | 0.0076 | 0.0065 | 0.0028 | 0.0387 | 0.0241 |
| | 3 | 1.11 | 0.90 | 0.0081 | 0.0103 | 0.0092 | 0.0056 | 0.0116 | 0.0079 |
| | | | 0.70 | 0.0063 | 0.0108 | 0.0081 | 0.0034 | 0.0194 | 0.0111 |
| | | | 0.50 | 0.0074 | 0.0099 | 0.0085 | 0.0035 | 0.0453 | 0.0276 |
| | 4 | 1.43 | 0.90 | 0.0116 | 0.0154 | 0.0127 | 0.0085 | 0.0171 | 0.0112 |
| | | | 0.70 | 0.0090 | 0.0157 | 0.0110 | 0.0038 | 0.0233 | 0.0128 |
| | | | 0.50 | 0.0095 | 0.0134 | 0.0112 | 0.0043 | 0.0549 | 0.0308 |

Table B.3: continued.

| Model | Scenario | $HR_N^{\{a\}}$ | $\rho$ / $\tau$ | PPS effect is small | | | PPS effect is large | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PFS only | OS only | OS and PFS | PFS only | OS only | OS and PFS |
| (c) | 1 | 0.90 | 0.90 | 0.0052 | 0.0056 | 0.0054 | 0.0041 | 0.0063 | 0.0046 |
| | | | 0.70 | 0.0032 | 0.0064 | 0.0043 | 0.0020 | 0.0235 | 0.0139 |
| | | | 0.50 | 0.0043 | 0.0050 | 0.0054 | 0.0013 | 0.0379 | 0.0227 |
| | 2 | 1.00 | 0.90 | 0.0067 | 0.0076 | 0.0069 | 0.0057 | 0.0087 | 0.0069 |
| | | | 0.70 | 0.0047 | 0.0095 | 0.0058 | 0.0024 | 0.0284 | 0.0166 |
| | | | 0.50 | 0.0060 | 0.0073 | 0.0076 | 0.0025 | 0.0450 | 0.0273 |
| | 3 | 1.11 | 0.90 | 0.0091 | 0.0110 | 0.0094 | 0.0080 | 0.0116 | 0.0092 |
| | | | 0.70 | 0.0069 | 0.0120 | 0.0086 | 0.0035 | 0.0328 | 0.0198 |
| | | | 0.50 | 0.0083 | 0.0101 | 0.0094 | 0.0038 | 0.0513 | 0.0323 |
| | 4 | 1.43 | 0.90 | 0.0118 | 0.0151 | 0.0124 | 0.0111 | 0.0167 | 0.0125 |
| | | | 0.70 | 0.0107 | 0.0165 | 0.0126 | 0.0042 | 0.0393 | 0.0224 |
| | | | 0.50 | 0.0115 | 0.0140 | 0.0128 | 0.0048 | 0.0609 | 0.0354 |
| (d) | 1 | 0.90 | 0.90 | 0.0030 | 0.0067 | 0.0043 | 0.0020 | 0.0075 | 0.0041 |
| | | | 0.70 | 0.0024 | 0.0059 | 0.0041 | 0.0027 | 0.0086 | 0.0059 |
| | | | 0.50 | 0.0042 | 0.0061 | 0.0052 | 0.0024 | 0.0190 | 0.0123 |
| | 2 | 1.00 | 0.90 | 0.0041 | 0.0086 | 0.0057 | 0.0029 | 0.0106 | 0.0059 |
| | | | 0.70 | 0.0032 | 0.0084 | 0.0059 | 0.0033 | 0.0124 | 0.0080 |
| | | | 0.50 | 0.0050 | 0.0084 | 0.0069 | 0.0030 | 0.0237 | 0.0144 |
| | 3 | 1.11 | 0.90 | 0.0052 | 0.0121 | 0.0078 | 0.0039 | 0.0136 | 0.0078 |
| | | | 0.70 | 0.0044 | 0.0106 | 0.0078 | 0.0042 | 0.0153 | 0.0099 |
| | | | 0.50 | 0.0061 | 0.0104 | 0.0084 | 0.0035 | 0.0276 | 0.0168 |
| | 4 | 1.43 | 0.90 | 0.0080 | 0.0171 | 0.0107 | 0.0053 | 0.0180 | 0.0105 |
| | | | 0.70 | 0.0066 | 0.0161 | 0.0108 | 0.0046 | 0.0217 | 0.0123 |
| | | | 0.50 | 0.0078 | 0.0152 | 0.0113 | 0.0043 | 0.0344 | 0.0185 |

Table B.4: Probabilities of rejecting $H_F$ at the final analysis using PFS only, OS only, or OS and PFS at the interim analysis under the assumption that the effect of PPS is small and large based on 10,000 simulation replications per scenario.

| Model | Scenario | $HR_N^{(a)}$ | $\rho$ / $\tau$ | PPS effect is small | | | PPS effect is large | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PFS only | OS only | OS and PFS | PFS only | OS only | OS and PFS |
| (a) | 1 | 0.90 | 0.90 | 0.5957 | 0.6158 | 0.5602 | 0.5957 | 0.6158 | 0.5602 |
| | | | 0.70 | 0.5624 | 0.5615 | 0.5220 | 0.6150 | 0.6816 | **0.5556** |
| | | | 0.50 | 0.5682 | 0.5584 | 0.5449 | 0.6126 | 0.6504 | **0.5206** |
| | 2 | 1.00 | 0.90 | 0.3893 | 0.3935 | 0.3505 | 0.3893 | 0.3935 | 0.3505 |
| | | | 0.70 | 0.3661 | 0.3726 | 0.3292 | 0.4075 | 0.4077 | **0.3167** |
| | | | 0.50 | 0.3702 | 0.3794 | 0.3496 | 0.3899 | 0.4042 | **0.2752** |
| | 3 | 1.11 | 0.90 | 0.2093 | 0.2002 | 0.1753 | 0.2093 | 0.2002 | 0.1753 |
| | | | 0.70 | 0.1893 | 0.2121 | 0.1651 | 0.2214 | 0.1753 | **0.1355** |
| | | | 0.50 | 0.1961 | 0.2087 | 0.1825 | 0.2040 | 0.1945 | **0.1133** |
| | 4 | 1.43 | 0.90 | 0.0160 | 0.0105 | 0.0099 | 0.0160 | 0.0105 | 0.0099 |
| | | | 0.70 | 0.0133 | 0.0199 | 0.0107 | 0.0117 | 0.0038 | 0.0031 |
| | | | 0.50 | 0.0139 | 0.0200 | 0.0129 | 0.0087 | 0.0077 | 0.0024 |
| (b) | 1 | 0.90 | 0.90 | 0.5935 | 0.5904 | 0.5809 | 0.5869 | 0.5810 | 0.5638 |
| | | | 0.70 | 0.6093 | 0.6004 | 0.5901 | 0.5744 | 0.5592 | **0.5129** |
| | | | 0.50 | 0.6135 | 0.6192 | 0.5930 | 0.5073 | 0.4940 | **0.4351** |
| | 2 | 1.00 | 0.90 | 0.3794 | 0.3806 | 0.3671 | 0.3731 | 0.3762 | 0.3545 |
| | | | 0.70 | 0.3863 | 0.3890 | 0.3691 | 0.3608 | 0.3686 | **0.3111** |
| | | | 0.50 | 0.3987 | 0.4007 | 0.3777 | 0.2977 | 0.3239 | **0.2456** |
| | 3 | 1.11 | 0.90 | 0.1942 | 0.2023 | 0.1862 | 0.1929 | 0.2005 | 0.1804 |
| | | | 0.70 | 0.1928 | 0.2002 | 0.1822 | 0.1740 | 0.2027 | **0.1451** |
| | | | 0.50 | 0.2037 | 0.2021 | 0.1863 | 0.1395 | 0.1814 | **0.1109** |
| | 4 | 1.43 | 0.90 | 0.0124 | 0.0145 | 0.0119 | 0.0123 | 0.0145 | 0.0109 |
| | | | 0.70 | 0.0099 | 0.0118 | 0.0090 | 0.0102 | 0.0142 | 0.0080 |
| | | | 0.50 | 0.0095 | 0.0100 | 0.0078 | 0.0058 | 0.0145 | 0.0036 |

The bold numbers show that incorporation of information for both OS and PFS results in good performance.

Table B.4: continued.

| Model | Scenario | $HR_N^{\{a\}}$ | $\rho$ / $\tau$ | PPS effect is small | | | PPS effect is large | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PFS only | OS only | OS and PFS | PFS only | OS only | OS and PFS |
| (c) | 1 | 0.90 | 0.90 | 0.5907 | 0.5864 | 0.5825 | 0.5888 | 0.5799 | 0.5701 |
| | | | 0.70 | 0.6076 | 0.5958 | 0.5895 | 0.5427 | 0.5199 | **0.4771** |
| | | | 0.50 | 0.6129 | 0.6137 | 0.5905 | 0.4991 | 0.4802 | **0.4291** |
| | 2 | 1.00 | 0.90 | 0.3760 | 0.3791 | 0.3695 | 0.3734 | 0.3757 | 0.3593 |
| | | | 0.70 | 0.3829 | 0.3911 | 0.3699 | 0.3260 | 0.3439 | **0.2801** |
| | | | 0.50 | 0.3963 | 0.3985 | 0.3710 | 0.2886 | 0.3152 | **0.2397** |
| | 3 | 1.11 | 0.90 | 0.1947 | 0.2021 | 0.1913 | 0.1895 | 0.1997 | 0.1820 |
| | | | 0.70 | 0.1938 | 0.1982 | 0.1849 | 0.1522 | 0.1929 | **0.1303** |
| | | | 0.50 | 0.2002 | 0.2027 | 0.1819 | 0.1340 | 0.1765 | **0.1058** |
| | 4 | 1.43 | 0.90 | 0.0120 | 0.0133 | 0.0117 | 0.0124 | 0.0152 | 0.0119 |
| | | | 0.70 | 0.0093 | 0.0116 | 0.0090 | 0.0071 | 0.0170 | 0.0060 |
| | | | 0.50 | 0.0100 | 0.0108 | 0.0075 | 0.0047 | 0.0178 | 0.0035 |
| (d) | 1 | 0.90 | 0.90 | 0.5899 | 0.5844 | 0.5664 | 0.5857 | 0.5748 | 0.5561 |
| | | | 0.70 | 0.6118 | 0.5992 | 0.5786 | 0.5880 | 0.5832 | **0.5303** |
| | | | 0.50 | 0.6193 | 0.6139 | 0.5849 | 0.5522 | 0.5451 | **0.4711** |
| | 2 | 1.00 | 0.90 | 0.3769 | 0.3787 | 0.3574 | 0.3721 | 0.3739 | 0.3452 |
| | | | 0.70 | 0.3847 | 0.3881 | 0.3567 | 0.3705 | 0.3813 | **0.3228** |
| | | | 0.50 | 0.3953 | 0.4022 | 0.3638 | 0.3363 | 0.3615 | **0.2729** |
| | 3 | 1.11 | 0.90 | 0.1936 | 0.2006 | 0.1820 | 0.1909 | 0.2052 | 0.1753 |
| | | | 0.70 | 0.1912 | 0.2016 | 0.1726 | 0.1857 | 0.2023 | **0.1516** |
| | | | 0.50 | 0.2040 | 0.2034 | 0.1793 | 0.1606 | 0.1966 | **0.1187** |
| | 4 | 1.43 | 0.90 | 0.0121 | 0.0134 | 0.0110 | 0.0124 | 0.0157 | 0.0112 |
| | | | 0.70 | 0.0092 | 0.0110 | 0.0081 | 0.0110 | 0.0135 | 0.0089 |
| | | | 0.50 | 0.0083 | 0.0105 | 0.0063 | 0.0078 | 0.0136 | 0.0040 |

The bold numbers show that incorporation of information for both OS and PFS results in good performance.

Table B.5: Probabilities of rejecting $H_P$ at the final analysis using PFS only, OS only, or OS and PFS at the interim analysis under the assumption that the effect of PPS is small and large based on 10,000 simulation replications per scenario.

| Model | Scenario | $HR_N^{(a)}$ | $\rho$ / $\tau$ | PPS effect is small | | | PPS effect is large | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PFS only | OS only | OS and PFS | PFS only | OS only | OS and PFS |
| (a) | 1 | 0.90 | 0.90 | 0.3854 | 0.3689 | 0.4225 | 0.3854 | 0.3689 | 0.4225 |
| | | | 0.70 | 0.3787 | 0.3814 | 0.4238 | 0.3812 | 0.3175 | **0.4411** |
| | | | 0.50 | 0.3758 | 0.3858 | 0.4011 | 0.3794 | 0.3458 | **0.4741** |
| | 2 | 1.00 | 0.90 | 0.5781 | 0.5811 | 0.6224 | 0.5781 | 0.5811 | 0.6224 |
| | | | 0.70 | 0.5504 | 0.5467 | 0.5966 | 0.5848 | 0.5908 | **0.6785** |
| | | | 0.50 | 0.5520 | 0.5461 | 0.5792 | 0.5902 | 0.5883 | **0.7159** |
| | 3 | 1.11 | 0.90 | 0.7465 | 0.7688 | 0.7903 | 0.7465 | 0.7688 | 0.7903 |
| | | | 0.70 | 0.7142 | 0.6967 | 0.7494 | 0.7613 | 0.8219 | **0.8567** |
| | | | 0.50 | 0.7144 | 0.7008 | 0.7349 | 0.7637 | 0.7954 | **0.8760** |
| | 4 | 1.43 | 0.90 | 0.9490 | 0.9645 | 0.9613 | 0.9490 | 0.9645 | 0.9613 |
| | | | 0.70 | 0.9086 | 0.8958 | 0.9136 | 0.9669 | 0.9950 | 0.9893 |
| | | | 0.50 | 0.9115 | 0.9011 | 0.9147 | 0.9678 | 0.9860 | 0.9889 |
| (b) | 1 | 0.90 | 0.90 | 0.3741 | 0.3775 | 0.3876 | 0.3745 | 0.3818 | 0.3997 |
| | | | 0.70 | 0.3645 | 0.3724 | 0.3845 | 0.3624 | 0.3844 | **0.4344** |
| | | | 0.50 | 0.3650 | 0.3600 | 0.3858 | 0.3263 | 0.3610 | **0.4251** |
| | 2 | 1.00 | 0.90 | 0.5704 | 0.5698 | 0.5848 | 0.5715 | 0.5690 | 0.5933 |
| | | | 0.70 | 0.5718 | 0.5691 | 0.5911 | 0.5558 | 0.5551 | **0.6217** |
| | | | 0.50 | 0.5645 | 0.5662 | 0.5888 | 0.5009 | 0.4961 | **0.5928** |
| | 3 | 1.11 | 0.90 | 0.7452 | 0.7396 | 0.7556 | 0.7442 | 0.7340 | 0.7605 |
| | | | 0.70 | 0.7545 | 0.7465 | 0.7688 | 0.7375 | 0.7122 | **0.7828** |
| | | | 0.50 | 0.7456 | 0.7501 | 0.7687 | 0.6604 | 0.6260 | **0.7236** |
| | 4 | 1.43 | 0.90 | 0.9389 | 0.9332 | 0.9394 | 0.9355 | 0.9272 | 0.9374 |
| | | | 0.70 | 0.9493 | 0.9440 | 0.9518 | 0.9274 | 0.9155 | 0.9333 |
| | | | 0.50 | 0.9523 | 0.9516 | 0.9555 | 0.8408 | 0.8159 | 0.8505 |

The bold numbers show that incorporation of information for both OS and PFS results in good performance.

Table B.5: continued.

| Model | Scenario | $HR_N^{\{a\}}$ | $\rho$ / $\tau$ | PPS effect is small | | | PPS effect is large | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PFS only | OS only | OS and PFS | PFS only | OS only | OS and PFS |
| (c) | 1 | 0.90 | 0.90 | 0.3741 | 0.3790 | 0.3828 | 0.3714 | 0.3813 | 0.3919 |
| | | | 0.70 | 0.3650 | 0.3758 | 0.3837 | 0.3353 | 0.3706 | **0.4174** |
| | | | 0.50 | 0.3643 | 0.3653 | 0.3877 | 0.3114 | 0.3518 | **0.4096** |
| | 2 | 1.00 | 0.90 | 0.5736 | 0.5721 | 0.5812 | 0.5677 | 0.5673 | 0.5849 |
| | | | 0.70 | 0.5736 | 0.5639 | 0.5881 | 0.5294 | 0.5198 | **0.5954** |
| | | | 0.50 | 0.5667 | 0.5678 | 0.5952 | 0.4889 | 0.4815 | **0.5719** |
| | 3 | 1.11 | 0.90 | 0.7477 | 0.7395 | 0.7520 | 0.7437 | 0.7345 | 0.7539 |
| | | | 0.70 | 0.7504 | 0.7450 | 0.7619 | 0.7029 | 0.6581 | **0.7429** |
| | | | 0.50 | 0.7488 | 0.7488 | 0.7740 | 0.6500 | 0.6061 | **0.7062** |
| | 4 | 1.43 | 0.90 | 0.9401 | 0.9358 | 0.9405 | 0.9364 | 0.9286 | 0.9372 |
| | | | 0.70 | 0.9466 | 0.9410 | 0.9475 | 0.8847 | 0.8553 | 0.8865 |
| | | | 0.50 | 0.9509 | 0.9503 | 0.9548 | 0.8218 | 0.7876 | 0.8266 |
| (d) | 1 | 0.90 | 0.90 | 0.3716 | 0.3791 | 0.3978 | 0.3690 | 0.3824 | 0.4031 |
| | | | 0.70 | 0.3631 | 0.3759 | 0.3985 | 0.3696 | 0.3801 | **0.4342** |
| | | | 0.50 | 0.3551 | 0.3622 | 0.3926 | 0.3509 | 0.3738 | **0.4511** |
| | 2 | 1.00 | 0.90 | 0.5672 | 0.5664 | 0.5897 | 0.5673 | 0.5658 | 0.5990 |
| | | | 0.70 | 0.5722 | 0.5675 | 0.6036 | 0.5655 | 0.5606 | **0.6253** |
| | | | 0.50 | 0.5640 | 0.5628 | 0.6006 | 0.5369 | 0.5326 | **0.6331** |
| | 3 | 1.11 | 0.90 | 0.7430 | 0.7345 | 0.7588 | 0.7418 | 0.7248 | 0.7629 |
| | | | 0.70 | 0.7552 | 0.7463 | 0.7799 | 0.7405 | 0.7306 | **0.7902** |
| | | | 0.50 | 0.7459 | 0.7483 | 0.7788 | 0.7054 | 0.6838 | **0.7823** |
| | 4 | 1.43 | 0.90 | 0.9353 | 0.9277 | 0.9371 | 0.9313 | 0.9193 | 0.9333 |
| | | | 0.70 | 0.9531 | 0.9452 | 0.9546 | 0.9323 | 0.9261 | 0.9391 |
| | | | 0.50 | 0.9546 | 0.9489 | 0.9578 | 0.9023 | 0.8892 | 0.9163 |

The bold numbers show that incorporation of information for both OS and PFS results in good performance.