Impact of Informative Censoring on Statistics Used in the Validation of Surrogate Endpoints in Oncology

IMPACT OF INFORMATIVE CENSORING ON STATISTICS USED IN THE VALIDATION OF SURROGATE ENDPOINTS IN ONCOLOGY

By

Yumeng Liu, B. Math.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2014) McMaster University

(Department of Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE: Impact of Informative Censoring on Statistics Used in the Validation of Surrogate Endpoints in Oncology

AUTHOR: Yumeng Liu,
B.Math. (Statistics)

SUPERVISOR: Dr. Gregory R. Pond

NUMBER OF PAGES: ix, 74

Abstract

In the past few years, biomarkers such as progression free survival (PFS) and time to progression (TTP), have been increasingly used as surrogate endpoints for overall survival (OS) in clinical trials in oncology. An issue occurs when clinical trials which demonstrated statistically significant treatment effect for the surrogate marker, shows no significant effect on the true outcome of interest, OS. It is possible that this lack of concordant results was due to informative censoring. Although it is known that informative censoring may bias the observed results, it is not clear what impact informative censoring has on the surrogacy of one marker in relation to a true outcome. In this thesis, we investigated how informative censoring could affect the results of a surrogate endpoint, and how would that affect the surrogacy of that endpoint. A simulation study was conducted to evaluate the impact of informative censoring on the treatment effect on TTP and the outcomes of the surrogate validation methods relative effect (RE), surrogate threshold effect (STE), and the difference between the treatment effect on TTP and on OS (IRE). The results of the simulation showed that having informative censoring for TTP will indeed bias the treatment effect on TTP as well as the results for the validation methods, RE, STE, and IRE. Hence, we conclude that the effect of informative censoring can greatly influence the ability to validate a surrogate marker, and additionally can bias the ability to determine the efficacy of a new therapy from a clinical trial using a surrogate marker as the primary outcome.

Acknowledgements

First, I would like to take this opportunity to express my gratitude to my extraordinary supervisor, Dr. Gregory Pond, who has supported me throughout my thesis with his patience, motivation, and knowledge. Next, I would like to thank the rest of my thesis examination committee members, Dr. Roman Viveros-Aguillera, and Dr. Joseph Beyene for their support, comments, and recommendations. Last but not the least, I would like to thank my parents and grandparents for their love and support throughout my life and my academic pursuits.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The goal of clinical trials in cancer research is to identify new promising therapies for cancer patients. Survival analysis is a branch of statistics that is commonly used in clinical trials for analyzing data where time-to-event endpoints are the outcome variables. A time-to-event endpoint measures the time from randomization to the event of interest. The event can be death, tumor progression, occurrence of a disease, etc. An example of a time-to-event endpoint is overall survival (OS), which measures the time from randomization until death from any cause.

In survival analysis, patients are followed over a specified period of time, and the time at which the event of interest occurs is recorded. If the event of interest (e.g. death) did not occur for some patients at the end of the follow up period during a study, these patients are said to be censored from the study. Censoring also occurs when patients are lost to follow up during the study. For a patient who was censored, investigators know his or her event of interest happens after the time which he or she was censored, but they do not know exactly when the patient experiences that event. The censored observations will also need to be taken into account when analyzing

time-to-event data.

In a cancer clinical trial, investigators evaluate the new therapy by comparing the treatment effect of the new therapy with the standard therapy on the endpoint of choice. The preferred endpoint for advanced cancer trials is OS, but this endpoint requires a very large sample size and prolonged follow up. Therefore, investigators often consider using a biomarker, such as time to progression and progression free survival as a surrogate endpoint for OS in cancer trials. Before a biomarker can be adopted as a surrogate endpoint in clinical practice, it must be statistically validated using standard statistical methods, and will often include reporting statistics such as proportion explained and relative effect. Surrogate validation involves determining if a surrogate marker can be used as a substitute for the true outcome of interest. This involves evaluating if the surrogate predicts the outcome of interest, but also if the treatment effect on the surrogate can accurately predict the treatment effect on the true outcome. The treatment effect on a time-to-event endpoint can be estimated using the hazard ratio calculated from the Cox proportional hazards regression model. One assumption for the Cox regression model is that the censored patients are at the same risk for treatment failure than those who remain in the study. However, if a large percentage of patients drop out the study due to reasons such as adverse effect and declining health, the above assumption will be violated. In these situations, the estimated treatment effect and the results of the validation methods for surrogacy may be biased, which can be problematic. The type of censoring that is mentioned above is called informative censoring.

This thesis will focus on estimating the impact of informative censoring on the surrogate endpoint as well as the impact on the surrogate validation methods. Chapter 2 discusses the commonly used surrogate endpoints for OS and different statistical validation methods of surrogacy. Chapter 3 is a literature review which describes

situations in which informative censoring may occur. Chapter 4 explains the design of the simulation that evaluates the impact of informative censoring. The results of the simulation are presented in Chapter 5. Chapter 6 concludes the study with a discussion on the simulation results.

Chapter 2

Surrogate Endpoints

2.1 Definitions

Clinical Endpoint - defined as a characteristic that measures how a patient feels, functions, or survives (Lesko and Atkinson, 2001).

Biomarker - defined as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Biomarkers Definitions Working Group, 2001). A biomarker can be a laboratory measure, imaging result, clinical endpoint or any other objectively measured value.

Surrogate Endpoint - defined as a biomarker which serves as a substitute of a clinical endpoint that is expected to accurately predict the effect of treatment on the clinical endpoint (Lesko and Atkinson, 2001).

Overall Survival - defined as the time from randomization until death from any cause (Food and Drug Administration, 2007).

Time to Progression - defined as the time from randomization until tumor progression (Food and Drug Administration, 2007).

Progression Free Survival - defined as the time from randomization until tumor progression or death from any cause (Food and Drug Administration, 2007).

Post Progression Survival - defined as the time from tumor progression until death from any cause.

Censoring - occurs when a patient's event of interest is not completely observed but some information on the value is available at the time of analysis (Berger, 2005).

Non-informative Censoring - occurs when a patient who was censored from a study and the reason for censoring is unrelated to the future likelihood of the event of interest.

Informative Censoring - occurs when a patient who was censored from a study and the reason for censoring is related to the future likelihood of the event of interest.

2.2 Endpoints in Randomized Clinical Trials in Oncology

2.2.1 Introduction

The gold standard clinical endpoint for assessing efficacy of new therapies for advanced cancer is OS (Denne *et al.*, 2013) which measures the time from randomization until death from any cause. In the context of clinical trials, OS can be objectively measured and is of substantial clinical importance to patients, investigators and regulatory

authorities. OS events are also somewhat rare, and they require prolonged periods of follow-up to observe patient deaths, particularly in some diseases like breast or prostate cancer. Using this endpoint as the primary endpoint in a randomized cancer trial may require substantial time to observe a beneficial treatment effect of a new therapy, since the new therapy can only be approved after its efficacy is proven in a clinical trial, which may take years. Moreover, the potential treatment effects of early therapies on OS may be confounded by the effects of any subsequent therapies (Burzykowski *et al.*, 2008). Therefore, using surrogate endpoints for survival has become an increasingly popular choice for investigators in phase 3 randomized controlled cancer trials.

2.2.2 Tumor Response Versus Disease Progression

Although numerous biomarkers have been proposed, the two most common types of biomarkers used in cancer research in lieu of overall survival is "tumor response" and "disease progression" (Oxnard *et al.*, 2012). Tumor response is a "time-tested marker of therapeutic efficacy" (Oxnard *et al.*, 2012, P. 1534). Tumor response is primarily used to calculate the response rate, which is the proportion of responders among all participants in a trial to quantify the efficacy of the new drug. It is objectively measured using standard criteria, such as the Response Evaluation Criteria In Solid Tumors (RECIST) (Eisenhauer, 2009). Tumor response can be assessed early during a treatment, and it is not normally used to determine whether the patients should change to a new therapy or not (Oxnard *et al.*, 2012).

Disease progression is a marker of treatment failure, and it is "usually reserved for patients with advanced disease" (Oxnard *et al.*, 2012, p. 1534). Disease progressions are frequently measured as time-to-event endpoints, such as time to progression (TTP) and progression free survival (PFS). Both TTP and PFS measure the length of time

from randomization until tumor progression, but PFS also counts the patients who die from any cause whereas TTP does not include deaths. Between TTP and PFS, PFS is the preferred regulatory endpoint by the U.S. Food and Drug Administration (FDA), because they believe that death from any cause are randomly related with tumor progression (Food and Drug Administration, 2007).

None of tumor response, TTP or PFS is validated as a surrogate marker for OS in all circumstances. PFS has been validated as a surrogate for OS in some, selected situations. For example, PFS has been validated as an acceptable surrogate for survival in advanced colorectal (Buyse *et al.*, 2007 and Tang *et al*, 2007) and advanced ovarian cancers (Bast *et al.*, 2007). It is important to note that these validated surrogates are very context dependent; for instance, where PFS is validated as a surrogate for OS in advanced colorectal cancer is specific to the case where 5-fluorouracil (5-FU) based chemotherapies were used, and relates to 2-year or 3-year PFS as a surrogate for 5-year OS. (Buyse *et al.*, 2007). This means that 2-year PFS may not be a valid surrogate in future studies, which may use different treatments, such as molecularly targeted agents. In addition, over time, a marker may no longer be valid as a surrogate due to changing therapies, such as when increasing numbers of useful salvage therapies become available (Buyse *et al.*, 2010). Although only validated in certain situations, PFS is becoming more frequently used as the primary outcome in randomized clinical trials in many different types of cancer.

2.2.3 Advantages of PFS

There are three major advantages of using PFS as the primary endpoint in phase 3 randomized cancer trials. First, tumor progression can be observed earlier than death. Hence, using PFS can significantly shorten the study duration since less time

is required to observe the same number of events. Second, trials using PFS require a much smaller sample size than trials using OS. This is because tumor progression can be observed much more frequently among patients than death, further decreasing the sample size. For example, in a phase 3 clinical trial for a particular type of advanced cancer, 500 events may be required for the final analysis. One might assume that 2/3rds of patients enrolled to a study will have a PFS event in a defined period of time, compared with only 1/3rd having an OS event. Hence, only 750 patients would be required to observe 500 PFS events, as compared with 1500 patients to observe 500 OS events. Third, the effect of treatment on OS is usually smaller than on PFS, since the potential treatment effect of the new therapy on survival may be confounded or diminished by the effect of subsequent therapies given to patients after progression (Burzykowski *et al.*, 2008). Thus, PFS is better at detecting treatment effect than OS.

Unfortunately, as mentioned previously, PFS is not validated as a surrogate for OS in many circumstances (see for example: Laporte *et al.*, 2013, and Paoletti *et al.*, 2013). In these situations, it is possible that PFS is not a true surrogate for OS, and significant results may be observed using the PFS outcome, which does not translate to significant results for the OS outcome. Alternatively, PFS may be a true surrogate for OS, but it has just not been demonstrated statistically. To separate these two situations, a vast amount of statistical work has occurred in recent years in an attempt to validate, or not validate, the use of PFS as a surrogate for OS across a large variety of contexts (Buyse *et al.*, 2007 , Burzykowski *et al.*, 2008, Laporte *et al.*, 2013, and Paoletti *et al.*, 2013)

2.3 Statistical Definitions of Surrogacy

According to the FDA (2007), for a novel therapy to receive regulatory approval, clinical evidence is required which demonstrates efficacy of the novel therapy, generally observed in a well-run confirmatory randomized trial using a primary outcome which is well-known to be of clinical importance or which is a validated surrogate of a clinically important endpoint. What is clinically important, however, is very subjective, with the only universally agreed clinically important outcome for oncology clinical trials being overall survival (Food and Drug Administration, 2007). Use of tumor response, TTP or PFS endpoints are then used based on the belief that these endpoints are reasonable surrogates of overall survival. While this belief is biologically plausible, and there is usually, if not always, a strong level of correlation between these endpoints and OS, this does not, by itself, confirm the use of tumor response, PFS or TTP as validated surrogates of OS. It is well-known that correlation does not imply surrogacy (Fleming and DeMets, 1996). Thus, for a marker to be considered a statistically validated surrogate, there are statistical criteria that must be satisfied.

2.3.1 Prentice's Criteria

Prentice's criteria were proposed by Prentice in 1989, and it is a set of criteria that are used to define a surrogate endpoint and are often considered the gold standard in terms of validation criteria. To interpret the Prentice criteria, denote the potential surrogate and the true clinical endpoint by random variables S and T respectively, and let Z be an indicator variable for the treatment. Moreover, let parameters α denote the effect of Z on S, β denote the effect of Z on T, γ denote the effect of S on T, β_s denote the effect of Z on T after adjustment for S, and γ_z denote the association

between S and T after adjustment for the effect of Z. (Buyse and Molenberghs, 1998). Prentice's criteria then consist of a set of four criteria, and a surrogate endpoint is validated if it fulfills all four criteria. These four criteria can be written as:

$$f(S|Z) \neq f(S) \tag{2.1}$$

$$f(T|Z) \neq f(T) \tag{2.2}$$

$$f(T|S) \neq f(T) \tag{2.3}$$

$$f(T|S,Z) = f(T|S), \tag{2.4}$$

where $f(X)$ denotes the probability distribution of random variable X and $f(X|Z)$ denotes the probability distribution of X conditional on the value of Z. These criteria can be verified using appropriate hypothesis tests. The interpretation of the four criteria, with corresponding statistical tests of significance, are:

1. Z has a significant effect on S ($H_0 : \alpha = 0$).
2. Z has a significant effect on T ($H_0 : \beta = 0$).
3. S has a significant effect on T ($H_0 : \gamma = 0$)
4. The effect of Z on T after adjustment for S is no longer significant ($H_0 : \beta_s \neq 0$).

One must reject H_0 for all four tests to meet the criteria of a surrogate. An example is provided which illustrates how one might validate a surrogate in the context of Prentice's criteria.

2.3.1.1 Example

Assume the surrogate and the true endpoint are both binary endpoints. Let,

$$
\begin{aligned}
Z_i &= \begin{cases} 0, \text{ if patient } i \text{ received placebo} \\ 1, \text{ if patient } i \text{ received the new drug} \end{cases} \\
S_i &= \begin{cases} 0, \text{ if patient } i \text{ have achieved the surrogate endpoint} \\ 1, \text{ otherwise} \end{cases} \\
T_i &= \begin{cases} 0, \text{ if patient } i \text{ have achieved the true endpoint} \\ 1, \text{ otherwise,} \end{cases}
\end{aligned}
$$

where Z_i, S_i and T_i are indicator variables for treatment, the surrogate endpoint and the true endpoint for patient i in the trial, $i = 1, ..., n$. The logistic regression model is commonly used when both endpoints are binary (Buyse and Molenberghs, 1998). The logistic model for the effect of Z on S can be expressed as:

$$
log\left(\frac{P(S_i = 1|Z_i)}{P(S_i = 0|Z_i)}\right) = \mu_{ZS} + \alpha Z_i, \tag{2.5}
$$

where μ_{ZS} is the intercept of the model and $\alpha = log(OR_{ZS})$, which is the log odds ratio between Z and S. Similarly, the logistic model for the effect of Z on T and the effect of S on T are:

$$
\begin{aligned}
log\left(\frac{P(T_i = 1|Z_i)}{P(T_i = 0|Z_i)}\right) &= \mu_{ZT} + \beta Z_i & (2.6) \\
log\left(\frac{P(T_i = 1|S_i)}{P(T_i = 0|S_i)}\right) &= \mu_{ZT} + \gamma S_i, & (2.7)
\end{aligned}
$$

where $\beta = log(OR_{ZT})$ and $\gamma = log(OR_{ST})$. To obtain the value for β_s, the effect of Z on T after the adjustment for S, and to obtain the value for γ_s, the effect of S on T after the adjustment for Z, the following model can be used:

$$log\left(\frac{P(T_i = 1|Z_i, S_i)}{P(T_i = 0|Z_i, S_i)}\right) = \mu_{ZT|S} + \beta_S Z_i + \gamma_Z Si + \delta Z_i S_i, \tag{2.8}$$

where $\beta_s = log(OR_{ZT|S})$, $\gamma_Z = log(OR_{ST|Z})$, and δ is the coefficient of the three way interaction(Buyse and Molenberghs, 1998). If the interaction is non-significant, it can be removed from model. In that case, the model becomes:

$$log\left(\frac{P(T_i = 1|Z_i, S_i)}{P(T_i = 0|Z_i, S_i)}\right) = \mu_{ZT|S} + \beta_S Z_i + \gamma_Z Si. \tag{2.9}$$

Then, the four null hypothesis corresponding to each criterion can be tested using models (2.5), (2.6), (2.7) and (2.9), respectively. To verify the fourth criterion, the statistical test for the effect of treatment on the true endpoint after adjustment of the surrogate endpoint needs to be non-significant. This is equivalent in saying that the hypothesis test for $H_0 : \beta_s = 0$ with $H_a : \beta_s \neq 0$ in model (2.9) needs to be non-significant. However, one can never prove the $H_0 : \beta_s = 0$ and reject $H_a : \beta_s \neq 0$ based on the fact that the statistical test is not significant. Since it is not possible to reject $\beta_s \neq 0$ in favour of $\beta_s = 0$, it is impossible to prove the last criterion. In other words, it is impossible to prove that the treatment effect on the true endpoint is no longer significant after the adjustment for the surrogate endpoint.

Fleming *et al.* (1994) pointed out that Prentice's criteria is too stringent. It can be useful in rejecting inappropriate surrogates, but it is inadequate in terms of validating a good surrogate. As demonstrated, one can never prove the fourth criterion, and in addition, it requires the treatment to have a significant effect on both of the surrogate

and true endpoint in order to verify the first and second criterion. For example, It is possible to have a valid surrogate, but the treatment has no effect on either the surrogate and true endpoint. This would violate the first and second criteria, simply because the treatment has no effect on outcomes. Moreover, Prentice's criteria is only sufficient to validate binary endpoints (Buyse and Molenberghs, 1998).

2.3.2 Proportion Explained

Freedman *et al.* (1992) addressed the problems with Prentice's criteria and proposed a statistic, called the proportion explained (PE), which can be used in the validation of surrogate markers. This method focuses on the treatment effect on the clinical true endpoint with and without adjustment for the surrogate endpoint. Let PE be the proportion of the treatment effect on the true endpoint that can be explained by the surrogate endpoint, then

$$PE = \frac{\beta - \beta_S}{\beta} = 1 - \frac{\beta_s}{\beta}, \tag{2.10}$$

where β is the effect of Z on T, and β_S is the effect of Z on T with the adjustment for S. The theory behind proportion explained is that if S is a valid surrogate for T, then there would remain very minimal effect of Z on T after adjusting for S, in other words, β_s will be small in comparison to β. Prentice's fourth criteria requires $\beta_s = 0$, in which case, $PE = 1$. For a valid surrogate endpoint, $PE \approx 1$, and the lower confidence limit of PE is larger than a certain value, such as 0.5 and 0.75. This indicates that a large amount of treatment effect on the true endpoint can be explained by the surrogate (Buyse and Molenberghs, 1998).

Unlike Prentice's criteria, the estimate of PE can be defined in most types of endpoints, such as binary, normally distributed endpoints and time-to-event endpoints

(Lin *et al.*, 1997). Buyse and Molenberghs (1998) pointed out that PE is in fact not a proportion, since its value does not necessarily lie between 0 and 1. They also identified that the confidence intervals for PE are generally too wide for small and median sized randomized trials, and they can often extend beyond (0, 1).

2.3.2.1 Example

In the previous example where the surrogate and the true endpoint are both binary, β can be calculated using the log odds ratio between Z and T and β_S can be calculated using the log odds ratio of Z and T given S. Using these values, PE can be expressed as:

$$PE = 1 - \frac{\beta_s}{\beta} = 1 - \frac{log(OR_{ZT|S})}{log(OR_{ZT})}. \tag{2.11}$$

2.3.3 Relative Effect and Adjusted Association

Buyse and Molenberghs (1998) introduced relative effect (RE) as a possible improvement upon proportion explained for validating surrogates. They mentioned that "for a surrogate endpoint to be useful in practice, the investigators must be able to predict the effect of treatment on the true endpoint based on the observed effect of treatment on surrogate" (Buyse and Molenberghs, 1998, p. 1022). Let RE stand for the effect of Z on T relative to the effect of Z on the S, then RE can be calculated using the following equation: $RE = \frac{\beta}{\alpha}$. A good surrogate is where $RE \approx 1$, as the treatment effect on the surrogate can accurately predict the treatment effect on the true endpoint. In most cases, RE is less than 1 since the treatment is likely to have a greater effect on the surrogate than on the true endpoint (Buyse and Molenberghs, 1998).

In combination with RE, another validation method for surrogate endpoints, ad-

justed association, was also proposed by Buyse and Molenberghs (1998). As mentioned previously, parameter γ_z denotes the association between the surrogate and the true endpoint after adjusting for the effect of the treatment. For a good surrogate, γ_z is close to infinity for binary endpoints and 1 for continuous or time-to-event endpoints (Buyse and Molenberghs, 1998).

Relative effect describes the effect of treatment on both S and T at the population level (Buyse *et al.*, 2000). It would be useful in predicting treatment effect on the true endpoint based on the treatment effect on the surrogate endpoint in future trials. However, like proportion explained, the confidence interval for RE will generally be too wide for small and medium sized trials. The adjusted association connects S and T at the individual level, and it can be used to predict the outcome for an individual patient. γ_z has a much narrower confidence interval compare to PE and RE when it is estimated from clinical trials with small and median sample sizes (Buyse and Molenberghs, 1998). Relative effect and adjusted association can be applied on binary, time-to-event, and normally distributed endpoints (Buyse, 2008). One problem with these two methods is that they are designed to validate surrogates in a single trial. Buyse (2008) pointed out the needs for "repeated demonstrations of a strong correlation between the marker and the clinical outcome" in order for a biomarker to be accepted as a surrogate. The next two methods that will be addressed here extended relative effect and adjusted association in a multiple randomized trials setting using the meta-analysis approach.

2.3.3.1 Example

As it was found previously, when both endpoints are binary, $\alpha = log(OR_{ZS})$ and $\beta = log(OR_{ZT})$, RE can be expressed as:

$$RE = \frac{\beta}{\alpha} = \frac{log(OR_{ZT})}{log(OR_{ZS})}. \tag{2.12}$$

The adjusted association, γ_z, can be estimated using model (2.9) or calculated as $log(OR_{ST|Z})$.

For time-to-event surrogate and true endpoints, $\alpha = log(HR_{ZS})$, the log hazard ratio (HR) of the surrogate endpoint and $\beta = log(HR_{ZT})$, the log HR of the true endpoint. RE can then be calculated as:

$$RE = \frac{\beta}{\alpha} = \frac{log(HR_{ZT})}{log(HR_{ZS})}. \tag{2.13}$$

Hazard ratios can be estimated using the Cox proportional hazards regression model. Further details on HR and the Cox regression model are provided in Appendix A.

2.3.4 Individual and Trial Level Association

The individual and trial level associations were proposed by Buyse *et al.* in 2000. The individual level association is like the adjusted association but carried over when data are available on multiple similar trials. Let R^2_{indiv} be the rank correlation coefficient between the potential surrogate and the true endpoint after the adjustment for both the trial effect and the treatment effect. R^2_{indiv} indicates the correlation between the surrogate and the true endpoint at an individual level. A surrogate is said to be valid at the individual level if $R^2_{indiv} \approx 1$, indicating a strong association between the surrogate and the true endpoints.

The trial level association is similar to relative effect, but it can be used on multiple similar trials of the same experiment, or experiments with similar treatment comparisons. The association between the treatment effect on the surrogate and on the true endpoint can be analyzed using a linear regression model. Let R^2_{trial} denote the coefficient of determination of the model, and it can be used to quantify the precision of the prediction of the trial-specific treatment effect on the true endpoint from the

treatment effect on the surrogate. A surrogate is said to be valid at the trial level if R^2_{trial} is close to 1, indicating a strong association between the effect of treatment on the surrogate and on the true endpoint (Buyse *et al.*, 2000).

The confidence limits of the estimates of R^2_{indiv} and R^2_{trial} are generally much narrower than the confidence limits of PE and RE. Furthermore, these two methods do not require the treatment to have a significant effect on the surrogate or the true endpoint, since "even though the treatment may not have any effect on the surrogate endpoint as a whole, the fluctuations around zero in individual trials (or other experimental units) can be very strongly predictive of the effect on the true endpoint" (Buyse *et al.*, 2000, p. 58). A biomarker could be claimed to be a valid surrogate for the clinical endpoint if both R^2_{indiv} and R^2_{trial} are close to one, which implies that the surrogate is strongly related to the true outcome at both the trial-level and individual-level. Just like relative effect and adjusted association, the individual and the trial level association can be applied on various types of surrogate and true endpoints, such as binary, time-to-event, and normally distributed endpoints (Buyse, 2008).

2.3.5 Surrogate Threshold Effect

An issue was raised by Burzykowski and Buyse (2006) with R^2_{trial}. They pointed out that it is difficult to interpret its value. Even though, one would desire $R^2_{trial} = 1$ which indicates a perfect association between the treatment effect on the true endpoint and on the surrogate, such a situation does not exist in practice. Thus, it is not clear what the value of R^2_{trial} should be in order for the treatment effect on the surrogate to accurately predict the treatment effect on the true endpoint. To address this difficulty, Burzykowski and Buyse (2006) proposed the surrogate threshold effect (STE). STE

can be defined as "the minimum value of treatment effect on the surrogate endpoint, for which the predicted effect on the true endpoint would be significantly different from 0". (Burzykowski and Buyse. 2006, p. 183). STE can be computed for any type of surrogate and true endpoints, such as binary, time to event and normally distributed endpoints (Burzykowski and Buyse, 2006).

Realistically, STE is not so much a measure of determining if a surrogate could be adequate or not, but a measure that is useful in distinguishing between treatment effects on the surrogate that are sufficient in predicting a clinical meaningful effect on the true endpoint and effects that are not. For example, suppose PFS has been evaluated as a surrogate endpoint for OS. Since both PFS and OS are time-to-event endpoints, log hazard ratio (LHR) could be used to quantify the treatment effect on PFS and on OS, which can be expressed respectively as LHR_{PFS} and LHR_{OS} (Buyse *et al.*, 2007). The calculation of the STE is based on the upper bound of the 95% prediction interval for LHR_{OS}. LHR_{OS} and LHR_{PFS} are related through a simple linear regression model (Burzykowski and Buyse, 2006). Let parameter LHR_{PFSi} and LHR_{OSi} denote the LHR_{PFS} and LHR_{OS} for subject i, $i = 1, ..., n$, respectively. Then, the regression model can be written as:

$$LHR_{OSi} = \alpha + \beta LHR_{PFSi} + \epsilon_i, \tag{2.14}$$

where α is the intercept, β is the slope, and ϵ_i is a random variable with mean 0 and variance σ^2. Using estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma^2}$, for a given LHR^*_{PFS}, LHR_{OS} can be estimated as:

$$\widehat{LHR_{OS}} = \hat{\alpha} + \hat{\beta} LHR^*_{PFS}. \tag{2.15}$$

For a given LHR^*_{PFS}, the upper bound of the 95% prediction interval for LHR_{OS}, denote by $U(LHR_{OS})$, can be obtained as:

$$U(LHR_{OS}) = \widehat{LHR_{OS}} + z_{0.975}\sqrt{Var(\widehat{LHR_{OS}} - LHR_{OS})} \quad (2.16)$$

$$= \hat{\alpha} + \hat{\beta} LHR^*_{PFS} + z_{0.975}\hat{\sigma}, \quad (2.17)$$

where $z_{0.975}$ is the 0.975 quantile of the standard normal distribution.

Let LHR^*_{PFS} be a value that solves $U(LHR_{OS}) = 0$, then $STE = exp(LHR^*_{PFS})$. The upper prediction bound is used here because the reduction of the hazard ratio, is considered beneficial to patients. The larger the variance and the intercept, the smaller the value of STE (Burzykowski and Buyse, 2006). STE measures what HR for PFS would necessarily translate to a HR for OS that is less than 1. For instance, a STE of 0.75 means that in order to ensure a nonzero treatment benefit in OS in a future trial, or equivalently, $HR_{OS} < 1$, a HR of 0.75 or lower would need to be obtained for HR_{PFS}. In terms of assessing the validity of a surrogate marker, if one observes a STE which is impossibly small, then one would have a non-useful surrogate. For example, if the $STE = 0.3$, then one would need to observe a $HR_{PFS} = 0.3$ to ensure that $HR_{OS} < 1$. Since this is unrealistic to be observed in reality, PFS would not be a good surrogate in this situation.

2.4 Informative Censoring

Censoring can be categorized into non-informative and informative. Non-informative censoring occurs when the censoring time of patients who are lost to follow-up is unrelated to the outcome of the study (e.g. move to a different country, study terminates and the patient has not yet experienced the study outcome of interest).

Conversely, informative censoring occurs when patients are censored, but due to reasons which may be related to the outcome (e.g. patients are in declining health and feels the treatment is not working, even though they have not met the objective criteria for the event or are too sick to come back for a follow-up appointment).

There are outstanding questions related to the effect of informative censored variables on surrogacy. It is well known that informative censoring can bias the results of a study using time-to-event outcomes. PFS is an outcome which could be affected by informative censoring, since cancer patients often come off clinical trials due to toxicity, feeling unwell or other reasons, prior to determination of objective progression. However, it is not known how a biased result on PFS affects the interpretation of the treatment effect on OS. That is, when PFS is used as a surrogate for OS, then measures such as STE and RE, which are used to allow interpretation of the treatment effect on OS based on the PFS results, must be affected if PFS is biased. The question is to what extent?

2.4.1 Theoretical Explanation on the Impact of Informative Censoring Using the Validation Statistics of Surrogacy

Assume there are two treatment arms in a clinical trial: one control arm where patients in this arm received placebo, and one experimental arm where patients received the new drug. Let the log hazard ratio for PFS and the log hazard ratio for OS quantify the treatment effect on PFS and on OS, and they can be expressed respectively as LHR_{PFS} and LHR_{OS}.

2.4.1.1 Surrogate Threshold Effect

As discussed in Section 2.3.5, LHR_{OS} can be expressed by LHR_{PFS} using a simple linear regression model, and the estimate of LHR_{OS} can be written as:

$$\widehat{LHR_{OS}} = \hat{\alpha} + \hat{\beta} LHR_{PFS}, \tag{2.18}$$

where $\hat{\alpha}$ is estimate of the intercept α, and $\hat{\beta}$ is the estimate of the slope β. It is worth noting that $\hat{\beta}$ will be greater than 0 since LHR_{OS} and LHR_{PFS} are generally positively correlated. Let SD_{OS} be the standard deviation for LHR_{OS} and SD_{PFS} be the standard deviation for LHR_{PFS}, and they can be calculated as follows:

$$SD_{OS} = \sqrt{\frac{\sum(LHR_{OSi} - \widehat{LHR_{OS}})^2}{n-1}} \tag{2.19}$$

$$SD_{PFS} = \sqrt{\frac{\sum(LHR_{PFSi} - \overline{LHR_{PFS}})^2}{n-1}}. \tag{2.20}$$

The 95% confidence interval for LHR_{OS}, for a given value LHR^*_{PFS}, can be obtained as:

$$\widehat{LHR_{OS}} \pm t_{(0.025),n-2} SD_{OS} \sqrt{\frac{(LHR^*_{PFS} - \overline{LHR_{PFS}})^2}{(n-1)SD^2_{PFS}}}. \tag{2.21}$$

Where $t_{(0.025),n-2}$ is the upper 0.025 quantile of the t-distribution with $n-2$ degrees of freedom. Investigators usually are only concerned with the cases where the new treatment is superior to the standard treatment (i.e. $HR_{OS} < 1$). Hence, only the upper bound of the 95% confidence interval will be discussed here. Let $U(LHR_{OS})$ represent the upper bound of the 95% confidence interval for LHR_{OS}, then $U(LHR_{OS})$

can be expressed as:

$$\begin{aligned} U(LHR_{OS}) &= \hat{\alpha} + \hat{\beta} LHR^*_{PFS} + t_{(0.025),n-2} SD_{OS} \sqrt{\frac{(LHR^*_{PFS} - \overline{LHR_{PFS}})^2}{(n-1)SD^2_{PFS}}} \\ &= \hat{\alpha} + \hat{\beta} LHR^*_{PFS} + t_{(0.025),n-2} \frac{SD_{OS}}{\sqrt{n-1}SD_{PFS}} \sqrt{(LHR^*_{PFS} - \overline{LHR_{PFS}})^2} \\ &= \hat{\alpha} + \hat{\beta} LHR^*_{PFS} + t_{(0.025),n-2} \frac{SD_{OS}}{\sqrt{n-1}SD_{PFS}} \left| LHR^*_{PFS} - \overline{LHR_{PFS}} \right| \end{aligned}$$

Let the true value for LHR^*_{PFS}, the log hazard ratio for PFS, be LHR_{PFSt}. Then $U(LHR_{OSt})$, the upper bound of the 95% confidence interval for LHR_{OS} given LHR_{PFSt} can be written as:

$$\begin{aligned} U(LHR_{OSt}) &= \hat{\alpha} + \hat{\beta} LHR_{PFSt} + t_{(0.025),n-2} \frac{SD_{OS}}{\sqrt{n-1}SD_{PFS}} \left| LHR_{PFSt} - \overline{LHR_{PFS}} \right| \\ &= \widehat{LHR_{OSt}} + t_{(0.025),n-2} SE_{OSt}, \end{aligned} \tag{2.22}$$

where $\widehat{LHR_{OSt}} = \hat{\alpha} + \hat{\beta} LHR_{PFSt}$ and $SE_{OSt} = \frac{SD_{OS}}{\sqrt{n-1}SD_{PFS}} \left| LHR_{PFSt} - \overline{LHR_{PFS}} \right|$. Assume informative censoring occurred in the experimental arm which caused the treatment effect on PFS to be overestimated. If such scenario occurred, the estimated log hazard ratio for PFS will be smaller than the true log hazard ratio (i.e. a larger treatment effect will be observed). Let b denote the bias of LHR_{PFS} caused by informative censoring, then the estimated value of LHR_{PFS} can be written as: $LHR_{PFSt} - b$, where $b > 0$. $U(LHR_{OSb})$, the upper bound of the 95% confidence

interval given $LHR_{PFSt} - b$, can be expressed as:

$$\begin{aligned} U \quad & (LHR_{OSb}) \\ = \quad & \hat{\alpha} + \hat{\beta}(LHR_{PFSt} - b) + t_{(0.025),n-2}\frac{SD_{OS}}{\sqrt{n-1}SD_{PFS}}\left|LHR_{PFSt} - b - \overline{LHR_{PFS}}\right| \\ = \quad & \hat{\alpha} + \hat{\beta}LHR_{PFSt} - \hat{\beta}b + t_{(0.025),n-2}\frac{SD_{OS}}{\sqrt{n-1}SD_{PFS}}\left|LHR_{PFSt} - b - \overline{LHR_{PFS}}\right| \\ = \quad & \widehat{LHR_{OSb}} + t_{(0.025),n-2}SE_{OSb}, \end{aligned} \tag{2.23}$$

where

$$\widehat{LHR_{OSb}} = \hat{\alpha} + \hat{\beta}LHR_{PFSt} - \hat{\beta}b \tag{2.24}$$

$$SE_{OSb} = \frac{SD_{OS}}{\sqrt{n-1}SD_{PFS}}\left|LHR_{PFSt} - b - \overline{LHR_{PFS}}\right|. \tag{2.25}$$

It is obvious that $\widehat{LHR_{OSb}}$ will always be smaller than $\widehat{LHR_{OSt}}$ by $\hat{\beta}b$, since $\hat{\beta} > 0$ and $b > 0$.

For the comparison of SE_{OSt} and SE_{OSb}, the following cases need to be considered:

1. If $LHR_{PFSt} \leq \overline{LHR_{PFS}}$, then $SE_{OSb} > SE_{OSt}$ by $\frac{SD_{OS}}{\sqrt{n-1}SD_{PFS}}b$.
2. If $LHR_{PFSt} > \overline{LHR_{PFS}}$ and $b > 2LHR_{PFSt} - 2\overline{LHR_{PFS}}$, then $SE_{OSb} > SE_{OSt}$ by $\frac{SD_{OS}}{\sqrt{n-1}SD_{PFS}}[b - 2LHR_{PFSt} + 2\overline{LHR_{PFS}}]$.
3. If $LHR_{PFSt} > \overline{LHR_{PFS}}$ and $b = 2LHR_{PFSt} - 2\overline{LHR_{PFS}}$, then $SE_{OSb} = SE_{OSt}$.
4. If $LHR_{PFSt} > \overline{LHR_{PFS}}$ and $b < 2LHR_{PFSt} - 2\overline{LHR_{PFS}}$, then $SE_{OSb} < SE_{OSt}$.

In the last two cases, where $SE_{OSb} = SE_{OSt}$ and $SE_{OSb} < SE_{OSt}$, $U(LHR_{OSb}) < U(LHR_{OSt})$ always. For the first two cases, where $SE_{OSb} > SE_{OSt}$, $U(LHR_{OSb})$ would also be smaller than or equal to $U(LHR_{OSt})$ except when

1. $\hat{\beta}b < t_{(0.025),n-2}\frac{SD_{OS}}{\sqrt{n-1}SD_{PFS}}b$, if $LHR_{PFSt} \leq \overline{LHR_{PFS}}$,

2. $\hat{\beta}b < t_{(0.025),n-2}\frac{SD_{OS}}{\sqrt{n-1}SD_{PFS}}[b-2LHR_{PFSt}+2\overline{LHR_{PFS}}]$, if $LHR_{PFSt} > \overline{LHR_{PFS}}$ and $b > 2LHR_{PFSt} - 2\overline{LHR_{PFS}}$.

The above results theoretically show that $U(LHR_{OSb})$ will be smaller than $U(LHR_{OSt})$ most of the time except for the two cases above. This imply that the upper bound of the 95% confidence interval for LHR_{OS} given $LHR_{PFSt} - b$ will more likely be smaller than what it should to be. That is, when there is a biased estimate of the treatment effect on PFS due to informative censoring, one is more likely to make a mistake and conclude that a real treatment effect on OS exists, when in truth, there is no evidence of a treatment effect. Furthermore, the results above also showed that if LHR_{PFS} is biased, the upper bound of the confidence interval for LHR_{OS} will be biased as well. This implies that the estimated STE will also be biased and give a greater value (i.e. a value closer to 1), since STE is calculated based on the upper bound of the confidence interval. From a clinical trial perspective, if one is using PFS as a primary outcome and a surrogate of OS, and uses STE to determine if the observed trial HR_{PFS} indicates superiority of an experimental treatment in terms of OS, then investigators will deem an experimental treatment as superior more likely than they actually should.

2.4.1.2 Relative Effect

Recall that RE represents the effect of the treatment on the true endpoint relative to the effect on the surrogate endpoint. In this example, $RE = \frac{LHR_{OS}}{LHR_{PFS}}$. Since we are only considering the case where the hazard ratios for OS and PFS are smaller than 1, the values of LHR_{OS} and LHR_{PFS} will be between negative infinity to 0. Let the true value for LHR_{PFS} be LHR_{PFSt}, then the true value of RE, RE_t, can be expressed as:

$$RE_t = \frac{LHR_{OS}}{LHR_{PFSt}}. \tag{2.26}$$

Let informative censoring occur in the experimental arm for PFS, and let LHR_{PFSb} denote the estimated treatment effect for PFS that is biased by informative censoring. Then RE_b, the RE value given LHR_{PFSb}, can be calculated as:

$$RE_b = \frac{LHR_{OS}}{LHR_{PFSb}}. \tag{2.27}$$

Because of informative censoring, the estimated hazard ratio for PFS will be smaller than the actual hazard ratio, which makes LHR_{PFSb} closer to negative infinity than LHR_{PFSt}. This change in the denominator will cause the value of RE to be decreased. Thus, RE_b will be smaller than RE_t at all times. In other words, when there is a biased estimate of the treatment effect on PFS due to informative censoring, it is more difficult to validate PFS as a valid surrogate for OS using RE as the validation statistics of surrogacy.

Chapter 3

Present State of Clinical Trials in Oncology

3.1 Statistical Validation in the Literature

3.1.1 Validation of Progression Free Survival as a Surrogate for Overall Survival

In the past few years, the individual level association, the trial level association, and the surrogate threshold effect have been frequently used to attempt to validate progression free survival as a valid surrogate for overall survival. Here are a few examples:

Buyse *et al.* (2007) investigated whether PFS is a surrogate for OS in advanced colorectal cancer. A meta-analysis was conducted using individual patient data from 3089 patients enrolled in 10 historical trials comparing medications fluorouracil (FU) plus leucovorin with either FU alone or with raltitrexed. The results showed that R^2_{indiv} is 0.82 (95% confidence interval (CI), 0.82 to 0.83) and R^2_{trial} is 0.99 (95% CI,

0.94 to 1.04). The surrogate threshold effect was 0.86, which implied that a hazard ratio of at most 0.86 is needed in PFS in order to predict a significant treatment benefit in OS in future trials. With the exclusion of one highly influential trial, R^2_{trial} and STE dropped to 0.75 (95% CI, 0.44 to 1.04) and 0.77, respectively (Buyse *et al.*, 2007). As a result, Buyse *et al.* (2007) concluded that PFS was a valid surrogate endpoint for OS for patients in advanced colorectal cancer treated with FU plus leucovorin.

Paoletti *et al.* (2013) conducted a meta-analysis consisting of 4069 patients in 20 randomized trials which investigated 5-FU, mitomycin C, anthracyclines, platinum agents, irinotecan and taxanes (GASTRIC, 2013). This study evaluated PFS as a surrogate for OS in trials dealing with advanced/recurrent gastric cancer. In this analysis, R^2_{indiv} was 0.85 (95% CI, 0.85 to 0.86), R^2_{trial} was 0.64 (95% CI, 0.04 to 1.00), and STE was 0.56. The value of R^2_{indiv} showed that PFS is highly correlated with OS, but the value of R^2_{trial} only showed a moderate correlation between the treatment effects on PFS and on OS (Paoletti *et al.*, 2013). Based on the STE of 0.56, a novel treatment must have a considerable impact on PFS (risk reduction of at least 44%) in order to observe a non-zero benefit in OS. Thus, Paoletti *et al.* (2013) could not conclude PFS is a valid surrogate for OS in advanced/recurrent gastric cancer.

PFS has also been evaluated as a surrogate endpoint for survival in advanced non-small-cell lung cancer (NSCLC) by Laporte *et al.* (2013). A meta-analysis of 2334 patients from 5 randomized trials comparing docetaxel-based chemotherapy with vinorelbine-based chemotherapy as the first-line treatment for advanced NSCLC was conducted. The results showed a moderate correlation between PFS and OS at the individual level (R^2_{indiv}=0.59; 95% CI, 0.58 to 0.61). The correlation between the treatment effect on PFS and on OS at the trial level were estimated within units of analysis consisting of 135 centres or 64 prognostic strata. R^2_{trial} using centres was 0.62 (95% CI, 0.52 to 0.72), and R^2_{trial} using strata was 0.72 (95% CI, 0.60 to 0.84).

The estimated STE was 0.49 within centres and 0.53 within strata (Laporte *et al.*, 2013). The values of R^2_{indiv} and R^2_{trial} showed that PFS and OS are only moderately correlated at the individual and the trial level. Thus, Laporte *et al.* (2013) could not verify PFS as an adequate surrogate for OS in advanced NSCLC, and they concluded that "only treatment that have a major impact on PFS (risk reduction of at least 50%) would be expected to also have a significant effect on OS" (Laporte *et al.*, 2013, p. 1).

3.1.2 Validation of Other Biomarkers as Surrogates for Overall Survival

Many additional studies have been conducted attempting to statistically validate other biomarkers of surrogacy, such as disease free survival (DFS) for OS in patients with adjuvant colon cancer (Sargent *et al.*, 2005) and gastric cancer (Oba *et al.*, 2013), leukemia-free survival (LFS) for OS in patients with acute myeloid leukemia (Buyse *et al.*, 2011), tumor response for OS in advanced colorectal cancer (Buyse *et al.*, 2000) and metastatic breast cancer (Burzykowski *et al.*, 2008), event-free survival (EFS) for OS in advanced head and neck cancer (Michiels *et al.*, 2009), and prostate-specific antigen as a surrogate for OS in hormonally treated patients with metastatic prostate cancer (Collette *et al.*, 2005). Among the studies mentioned above, only DFS in adjuvant colon cancer and gastric cancer, LFS in acute myeloid leukemia, and EFS in advanced head and neck cancer were considered statistically valid as surrogate endpoints for OS (Sargent *et al.*, 2005, Buyse *et al.*, 2011, Oba *et al.*, 2013, and Michiels *et al.*, 2009).

3.2 Clinical Question

Approval of a new drug by regulatory bodies occurs as a result of proven efficacy, which is generally shown in phase 3 clinical trials. Despite only moderate success in validating surrogates, the use of PFS as a surrogate for OS in phase 3 trials is increasing. A problem with determining if a drug has efficacy arises when trials that show a new drug has significant improvement in PFS, but the OS results are not yet available. Some trials have not yet shown a clinically meaningful benefit in OS, but still have led to approval of new drugs due to the drugs' superior performance in PFS (Booth and Eisenhauer, 2012). In instances such as this, there is a question as to whether the improvement in PFS will translate into clinical benefit for patients, in terms of an improvement in OS. The following trials are provided as a few examples where the results of OS did not agree with the results of PFS.

3.2.1 CALGB 90401

CALGB 90401 is a phase 3 randomized, placebo-controlled clinical trial. In this trial, Kelly *et al.* (2012) evaluated the addition of bevacizumab (BEV) to standard therapy docetaxel and prednisone (DP) in male patients with metastatic castration-resistant prostate cancer. 1050 patients were randomly assigned with 524 patients in the BEV plus DP arm and 526 patients in the DP alone arm. OS was the primary endpoint in this trial, and PFS was one of the secondary endpoints. In the final analysis, the median PFS for patients who received BEV plus DP was 9.9 months compared at 7.4 months for patients who received DP alone, and the HR for PFS was 0.8 ($P < 0.001$). Even though the results showed that PFS was significantly improved in the BEV plus DP group, OS did not show a corresponding increase in this combination-therapy

group. The median OS was 22.6 months and 21.5 months for patients in the BEV plus DP group and the DP alone group respectively, with estimated HR for OS to be 0.91 ($P = 0.181$) (Kelly *et al.*, 2012). One thing worth noting is that serious toxicity (75.4% vs. 56.2%) among patients who were given BEV plus DP happened more frequently than patients who were given DP alone, as well as treatment-related death (4% vs. 1.2%) (Kelly *et al.*, 2012).

The addition of BEV to the standard therapy DP significantly prolonged PFS, but it also added severe toxicity to patients in the experimental arm. Most importantly, this treatment combination did not significantly prolong OS which was the primary endpoint in this study. As a result of these findings, Kelly *et al.* (2012) did not recommend adding BEV in combination with DP for approval in treating male patients with metastatic castration-resistant prostate cancer.

3.2.2 ECOG 2100

ECOG 2100, another randomized phase 3 trial, studied the efficacy of bevacizumab (BEV) plus paclitaxel (PAC) versus PAC alone as the initial treatment for patients with metastatic breast cancer. 722 patients were enrolled in the trial, 368 of whom received BEV plus PAC and 354 of whom received PAC alone. The trial used PFS as the primary endpoint and OS as the secondary endpoint. With the median follow-up of 41.6 months in the combination-therapy group and 43.5 months in the control group, the results in the final analysis showed that PFS was significantly improved for patients in the BEV plus PAC arm compared with PAC alone. The median PFS was 11.8 months for patients given BEV plus PAC and 5.9 months for patients given PAC alone. The hazard ratio for progression was 0.60 ($P < 0.001$) indicating the risk was 40% lower for the patients in the combination-therapy group. Despite the superior

results in PFS, OS did not reach statistically significance. The median for OS was similar among the two groups with 26.7 months and 25.2 months in the BEV plus PAC and PAC alone group respectively, and the hazard ratio for OS was only 0.88 ($P = 0.16$) (Miller *et al.*, 2007). Serious adverse events were more common with the experimental arm than the control arm in this study. For example, grade 3 or higher neuropathy (23.6% vs. 17.6%), infection (9.3% vs. 2.9%) and fatigue (8.5% vs. 4.9%) were more frequent among patients who treated with BEV plus PAC (Miller *et al.*, 2007).

Adding BEV to PAC significantly improved PFS, but the median OS was only prolonged by 1 month. Since the prolongation of survival is the ultimate goal in cancer research, this finding along with the increased toxicity in the treatment arm with BEV, despite the primary outcome being PFS and OS being the secondary outcome, Miller *et al.* (2007) did not recommend adding BEV to PAC as a potential treatment for patients with metastatic breast cancer.

3.2.3 BOLERO-2

The third example is BOLERO-2, a phase 3 randomized trial for patients with hormone-receptor-positive advanced breast cancer. The trial compared everolimus (EVE) plus exemestane (EXE) versus EXE plus placebo in 724 patients, and these patients were allocated to treatment in an approximate 2:1 ratio. The primary endpoint was PFS, and OS was one of the secondary endpoints in this trial (Baselga *et al.*, 2012). The primary analysis of the trial, after 18 months median follow-up, showed that EVE plus EXE significantly prolonged PFS. The median PFS for patients who received EVE plus EXE was 7.8 months and 3.2 months for patients who received EXE plus placebo. The hazard ratio for progression or death in the treatment group relative to

the control group was 0.45 ($P < 0.001$), which indicated the risk was 55% lower for the patients in the EVE plus EXE group than the patients in the EXE plus placebo group. Unlike CALGB 90401 and ECOG 2100, OS results was immature at the time of the results of the analysis was published (Yardley *et al.*, 2013). Due to the superior performance of EVE plus EXE in PFS, Baselga *et al.* (2012) recommended the addition of EVE to EXE as a potential treatment choice for patients with hormone-receptor positive advanced breast cancer. Later in the same year, the U.S. FDA approved EVE in combination with EXE to treat certain postmenopausal women with advanced hormone-receptor positive advanced breast cancer (Food and Drug Administration, 2012).

After the new treatment has been approved, there were some questions about whether this observed benefit in PFS is a truly clinically meaningful benefit for patients. Two years after the primary analysis, Piccart *et al.* (2014) found out that OS was not statistically significant at 39 months median follow-up. The median OS for patients given EVE plus EXE was 31 months and 26.6 months for patients given EXE plus placebo. The hazard ratio for OS was 0.89 ($P = 0.14$), indicating that there was only 11% risk reduction for patients in the experimental arm in terms of OS, which is a drastically reduced effect compared with the PFS results. Even though the results of OS was not clinically significant, adding EVE into the therapy did prolong the median survival by 4 months. This is an improvement in survival by approximately one sixth of the median OS for patients who treated with EXE plus placebo. Moreover, similar to CALGB 90401 and ECOG 2100, adverse events were more common with the experimental arm than with the control arm in this trial. The majority of adverse events experienced with the EVE therapy occurred soon after the initiation of the therapy. Mild to moderate severity adverse effects were considered generally manageable by the study authors with dose reduction or discontinuation

of the therapy (Rugo *et al.*, 2014); however, grade 3 or 4 treatment-related toxicity, such as stomatitis (8% in the experimental arm vs. 1% in the control arm), anemia (6% vs. 1%), dyspnea (4% vs. 1%), hyperlycemia (4% vs. <1%), and fatigue (4% vs. 1%) were also more frequent with patients in the EVE plus EXE arm (Baselga et al., 2011). This finding made the decision of whether or not to treat with EVE becomes a question of personal choice for patients.

3.2.4 Why Did This Scenario Occur?

All three trials described demonstrate situations where a new therapy significantly improved PFS, but not OS. There are multiple reasons that could explain this scenario. The first one could be the potential treatment effect of the new therapy on OS may be confounded or diminished by the effect of subsequent therapies given to patients after progression. Second, the number of events for OS was smaller than the number of events for PFS in each trial. For example, among the 722 patients in the ECOG 2100 trial, at the time of the final analysis, 624 patients had an event for PFS but only 483 patients had an event for OS (Miller *et al.*, 2007). The power for the test of significance were reduced for OS. Another possibility is that the PFS results were biased due to informative censoring. For example, in BOLERO-2, due to adverse events, a higher percentage of patients discontinued EVE in the EVE plus EXE group than EXE in the EXE plus placebo group (19% vs. 4%) (Baselga *et al.*, 2011). These patients were considered to be informative censored from the study, because they were in declining health and it was felt that the treatment was not working or was adding too much toxicity, even though the patients did not meet the objective criteria for progression, they were likely to progress very soon after they left the study. Following discontinuation, objective progression evaluation was to continue as planned, although

it was likely that some patients altered their treatment or follow-up schedule. Hence, in this case, the treatment effect on PFS in the experimental group may be overestimated, since PFS can only be observed on patients who continued in the study.

When the results of OS did not agree with the results of PFS, there is less concern when the results for OS were presented at the same time as the results for PFS, such as the case in CALGB 90401 and ECOG 2100. But with BOLERO-2, the PFS results were published first, and due to the significant benefit in PFS, the new therapy that was proposed in the trial was approved by FDA before the OS results were available. Although, the OS results that were presented 2 years later still showed a slight improvement with the new therapy, this is concerning, not only for the interpretation of the BOLERO-2 results, but also for interpretation of future trials. Therefore, when the results for OS are not available, it is crucial to determine whether the improvement in PFS would translate into a corresponding improvement in OS in randomized cancer trials.

3.3 Summary

There are numerous examples of clinical trials which demonstrated statistically significant treatment effects for a surrogate marker such as PFS or TTP, however, no significant effect on the true outcome of interest, OS, was observed. This occurred despite the fact that PFS/TTP has been validated as a surrogate marker for OS in many scenarios. It is possible that this lack of concordant results was due to informative censoring. Although it is known that informative censoring may bias the observed results, it is not clear what impact informative censoring has on the surrogacy of one marker in relation to a true outcome. The simulation study presented in this thesis was undertaken to investigate this question, and to better understand the effect of

informative censoring on surrogacy.

Chapter 4

Simulation Design

A simulation study was conducted in order to investigate how informative censoring might affect the treatment effect on PFS/TTP and how this might affect our interpretation of the treatment effect on OS. TTP was used as the surrogate endpoint for this simulation. In the context of this study, TTP and PPS will be used interchangeably, because the cause of death for the majority of patients who enroll in a cancer clinical trial is due to tumor progression. In many cases where patients have died without previously being observed to have objectively progressed, the patient is too sick to have objective measurements and would have clinically progressed. Hence, for many clinical trials, investigators use PFS instead of TTP. In the context of this simulation, the surrogate marker has defined to be TTP since all simulated patients have a time after progression which occurs before death. Use of a PFS endpoint would be expected to yield similar results.

In this study, 1000 simulated clinical trials were generated, with 1000 patients in each trial. Among the 1000 patients, 500 patients were assigned to group 1, simulating as if these patients received a placebo, or control treatment, and 500 were assigned

to group 2, simulating as if these patients received the experimental treatment. A supportive analysis was performed investigating the effect of smaller sample sizes, with $n = 500$ total patients in each trial, and no important differences were observed. Hence, only data for the $n = 1000$ patient simulated trials are presented for simplicity. For determining TTP, each patient is simulated to have a TTP event time and a TTP censoring time. A censoring indicator was used where a 0 indicates that the patient was censored from the study for TTP, and the indicator 1 indicates that the patient had an event for TTP. Two independent random variables from the exponential distribution with separate rate parameters were generated, one for the censoring distribution and one for the progression time. If a patient's event time is smaller than his or her censoring time, this patient will be recorded as having an event for TTP at the TTP event time, and the censoring indicator set to 1. If a patient's censoring time is smaller than his or her event time, this patient will be recorded as censored for TTP at the TTP censoring time, and the censoring indicator set to 0. Hence, the outcome for TTP for a patient will be the minimum value between that patient's TTP event time and TTP censoring time. The rate parameter for each exponential distribution was carefully designed so that the TTP censoring rate can be controlled at a certain proportion, e.g. 0.10.

Post progression survival times and censoring times were simulated in a similar fashion, and the OS event and censoring times were set to be equal to the TTP times plus the post progression survivals times. Note that all of the censoring that was described above was considered non-informative censoring.

The simulation is designed so that the new therapy has the same treatment effect on OS and on TTP and both treatment arms have the same proportion of non-informative censoring. To investigate the effect of informative censoring on the treatment effect on TTP, informative censoring were added to arm 2, the experimental arm. For a

certain proportion of patients in arm 2, their censoring indicator was switched from 1 to 0. In this way patients who had an event were recorded in the database as being censored. This is similar to a patient being censored informatively, just prior to an event of interest occurring. In this way, this proportion (e.g. 0.10) of patient would be informatively censored, and would be deemed as censored, just prior to the occurrence of the actual event. The treatment effect on TTP and on OS were measured using log hazard ratios from the Cox regression model.

Variables which were simulated under different situations included HR_{OS} and the initial value of HR_{TTP} before informative censoring was applied, the proportion of censoring for OS in both treatment arms, the proportion of non-informative censoring for TTP in both arms, and the proportion of informative censoring for TTP in arm 2, the experimental arm. The values for these variables varied in the simulation are indicated in the table below:

Table 4.1: The Values of the Variables Varied During the Simulation

Variables	Values
HR_{OS} and the initial value of HR_{TTP}	1, 0.9, 0.7
Censoring rate for OS in both arms	0.15, 0.3, 0.6
Non-informative censoring rate for TTP in both arms	0.05, 0.1, 0.2
Informative censoring rate for TTP in arm 2	0, 0.05, 0.1, 0.2

The values that were evaluated for each variable represent plausible values that one might see in a clinical trial. For the censoring values, they represent the different levels of the proportion of censoring for OS and for TTP. For the HR values, $HR_{OS} = HR_{TTP} = 1$ indicates the new therapy has no treatment effect on OS and TTP, $HR_{OS} = HR_{TTP} = 0.9$ identifies small treatment effect, and $HR_{OS} = HR_{TTP} = 0.7$ represents large treatment effect.

The median event time for TTP was set to vary between 0.6 to 1.0, and the median

event time for OS was set to between 0.8 to 1.2. Note that time is described as standard units, and could refer to days, months or years, without any loss of generality. For the 1000 simulated trials, the primary outcome that was considered is RE, and the secondary outcomes were IRE and STE. The hazard ratio of TTP was also evaluated to examine the effect of informative censoring on the treatment effect. RE and STE were calculated using the methods introduced in Section 2.3.3 and 2.3.5, respectively. IRE represents the difference between the log hazard ratio of OS and the log hazard ratio of TTP and was calculated using the formula $logHR_{OS} - logHR_{TTP}$. 95% confidence intervals were also constructed for RE, IRE, and HR_{TTP}, and they were calculated using the 0.025 and the 0.975 percentiles among the 1000 simulated trials for each outcome.

Chapter 5

Simulation Results

Using the values of the variables indicated in Table 4.1, twenty-seven different scenarios were evaluated. For each scenario, one result table were generated assuming different rates of informative censoring as described. Each table contains the simulation outcomes for HR_{TTP}, RE, IRE, and STE for the different levels of informative censoring for TTP in the experimental arm.

5.1 Main Findings

The simulation outcomes for HR_{TTP}, RE, and STE, when the HR_{OS} and the initial HR_{TTP} value was 1.00, the censoring rate for OS was 0.30, and non-informative censoring rate was 0.10, are given in Table 5.1. In the table, when informative censoring was increased from 0 to 0.05 to 0.10 to 0.20, the median HR_{TTP} went from 0.999 to 0.943 to 0.887 to 0.777. Similarly, the median RE went from 0.869 to 0.394 to 0.051 to 0.002, the median IRE went from 0.001 to 0.058 to 0.119 to 0.251, and STE went from 0.897 to 0.846 to 0.795 to 0.696. In other words, due solely to the effect of

informative censoring, the HR_{TTP} went from no treatment effect at all, to a treatment effect indicating one treatment had a decreased hazard of over 22%. Not only would this often change a non-significant result to a significant result, but it affected the ability of TTP to be used as a surrogate, noted by the STE going from 0.897 to 0.696. For simplicity, additional results are presented in Appendix B.

Table 5.1: $HR_{OS} = HR_{TTP} = 1.00$ before informative censoring,
censoring for OS = 0.30, non-informative censoring for TTP in both arms = 0.10.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.10, Median HR for OS=1.000 (95% CI =0.870, 1.160).					
0.00	0.10	0.999 (0.882, 1.145)	0.869 (-8.011, 12.75)	0.001 (-0.102, 0.090)	0.897
0.05	0.15	0.943 (0.833, 1.078)	0.394 (-8.259, 7.828)	0.058 (-0.046, 0.149)	0.846
0.10	0.20	0.887 (0.780, 1.019)	0.051 (-7.335, 3.813)	0.119 (0.015, 0.210)	0.795
0.20	0.30	0.777 (0.685, 0.891)	0.002 (-1.093, 0.435)	0.251 (0.148, 0.341)	0.696

5.1.1 The Effect on the Hazard Ratio of Time to Progression

Figure 5.1 shows the association between informative censoring and the median HR_{TTP} for different values of non-informative censoring rate. Before informative censoring was added in arm 2, $HR_{TTP} = 1$. HR_{TTP} became smaller than HR_{OS} after informative censoring was applied, and the greater the informative censoring, the smaller the HR_{TTP}. For the same proportion of informative censoring, the higher

the proportion of non-informative censoring, the smaller the HR_{TTP}. However, the influence of informative censoring on HR_{TTP} were much larger than the influence of non-informative censoring. These associations hold regardless of the level of the non-informative censoring rate for TTP.

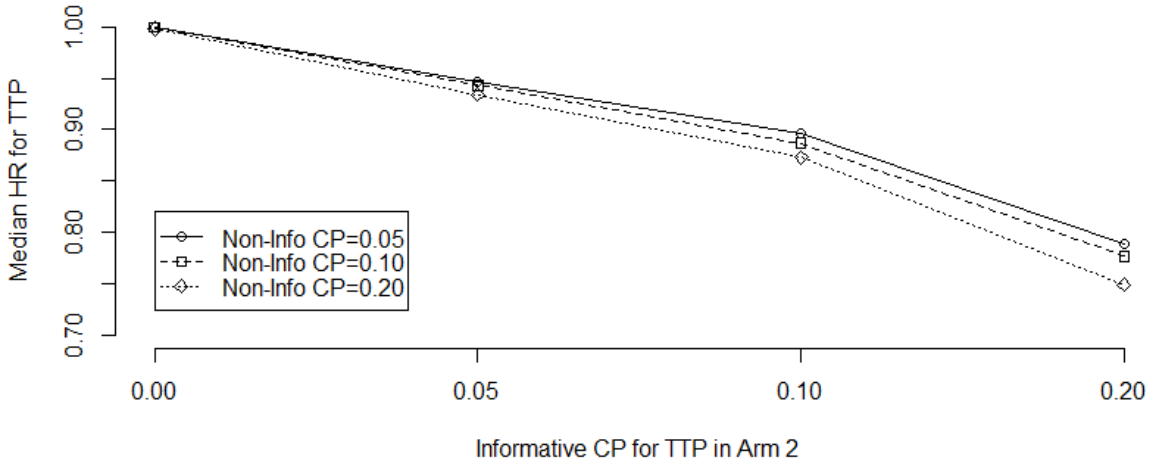

Figure 5.1: Informative Censoring Probability (CP) vs. the median HR for TTP with different non-informative CP for TTP, when the initial HR(OS)=HR(TTP)=1.00 and CP for OS=0.30.

5.1.2 Relative Effect

In Table 5.1, when both HR_{OS} and HR_{TTP} were approximately 1, the median RE was still not very close to 1. RE decreases rapidly as informative censoring rate for TTP increases, and its 95% CI were extremely wide. The values of RE became more stable with smaller CI as the initial treatment effect on OS and TTP increases. Table 5.2 to 5.5 shows the simulation results when HR_{OS} and the initial HR_{TTP} were both 0.9 or both 0.7. In these 4 tables, as informative censoring increases, RE decreased much more slowly, and its CI were much narrower, compared to the results when $HR_{TTP} = 1$. This result is demonstrated graphically in Figure 5.2 which graphed the

effect of informative censoring on the median RE with different initial HR values for OS and TTP.

The faster decrease in RE which occurred as informative censoring increased when both HR_{OS} and HR_{TTP} are close to 1, is likely due to the fact that the log scale of HR_{OS} and HR_{TTP} were used to quantify the treatment effect on OS and on TTP. Remember that $RE = \frac{logHR_{OS}}{logHR_{TTP}}$, thus, when $HR_{TTP} = 1$, $logHR_{TTP} = 0$, and therefore small changes in the denominator relate to large changes in RE.

Figure 5.2 also shows that when the informative censoring rate for TTP are kept the same, RE increases when the treatment effect on OS and TTP increases. Moreover, by comparing the RE values in Table 5.2 with Table 5.3 and Table 5.4, one can see that the percentage of censoring for OS has no effect on RE when the values of other variables remain the same.

Table 5.2: $HR_{OS} = HR_{TTP} = 0.90$ before informative censoring,
censoring for OS = 0.15, non-informative censoring for TTP in both arms = 0.10.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.10, Median HR for OS=0.896 (95% CI =0.793, 1.034).					
0.00	0.10	0.899 (0.793, 1.030)	1.005 (-1.143, 3.249)	-0.001 (-0.085, 0.074)	0.935
0.05	0.15	0.848 (0.749, 0.970)	0.662 (-0.585, 1.271)	0.055 (-0.027, 0.131)	0.883
0.10	0.20	0.798 (0.704, 0.912)	0.489 (-0.313, 0.840)	0.118 (0.035, 0.191)	0.830
0.20	0.30	0.699 (0.617, 0.798)	0.308 (-0.134, 0.534)	0.250 (0.167, 0.326)	0.727

Table 5.3: $HR_{OS} = HR_{TTP} = 0.90$ before informative censoring,
censoring for OS = 0.30, non-informative censoring for TTP in both arms = 0.10.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.10, Median HR for OS=0.899 (95% CI =0.783, 1.041).					
0.00	0.10	0.899 (0.793, 1.030)	0.990 (-1.380, 3.697)	0.001 (-0.103, 0.091)	0.913
0.05	0.15	0.848 (0.749, 0.970)	0.652 (-0.749, 1.385)	0.058 (-0.045, 0.149)	0.862
0.10	0.20	0.798 (0.704, 0.912)	0.474 (-0.341, 0.915)	0.120 (0.015, 0.208)	0.810
0.20	0.30	0.699 (0.617, 0.798)	0.299 (-0.152, 0.586)	0.252 (0.148, 0.342)	0.710

Table 5.4: $HR_{OS} = HR_{TTP} = 0.90$ before informative censoring,
censoring for OS = 0.60, non-informative censoring for TTP in both arms = 0.10.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.10, Median HR for OS=0.895 (95% CI =0.743, 1.105).					
0.00	0.10	0.899 (0.793, 1.030)	0.998 (-3.536, 5.375)	0.000 (-0.159, 0.159)	0.847
0.05	0.15	0.848 (0.749, 0.970)	0.656 (-1.797, 1.925)	0.058 (-0.104, 0.217)	0.799
0.10	0.20	0.798 (0.704, 0.912)	0.489 (-0.837, 1.231)	0.120 (-0.044, 0.282)	0.751
0.20	0.30	0.699 (0.617, 0.798)	0.313 (-0.402, 0.745)	0.251 (0.090, 0.413)	0.658

Table 5.5: $HR_{OS} = HR_{TTP} = 0.70$ before informative censoring, censoring for OS = 0.30, non-informative censoring for TTP in both arms = 0.10.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.10, Median HR for OS=0.698 (95% CI =0.606, 0.812).					
0.00	0.10	0.698 (0.615, 0.800)	1.000 (0.752, 1.324)	0.000 (-0.103, 0.091)	0.951
0.05	0.15	0.660 (0.583, 0.754)	0.862 (0.635, 1.117)	0.057 (-0.047, 0.150)	0.899
0.10	0.20	0.620 (0.546, 0.709)	0.751 (0.537, 0.971)	0.119 (0.014, 0.209)	0.847
0.20	0.30	0.543 (0.479, 0.621)	0.589 (0.413, 0.758)	0.251 (0.145, 0.344)	0.744

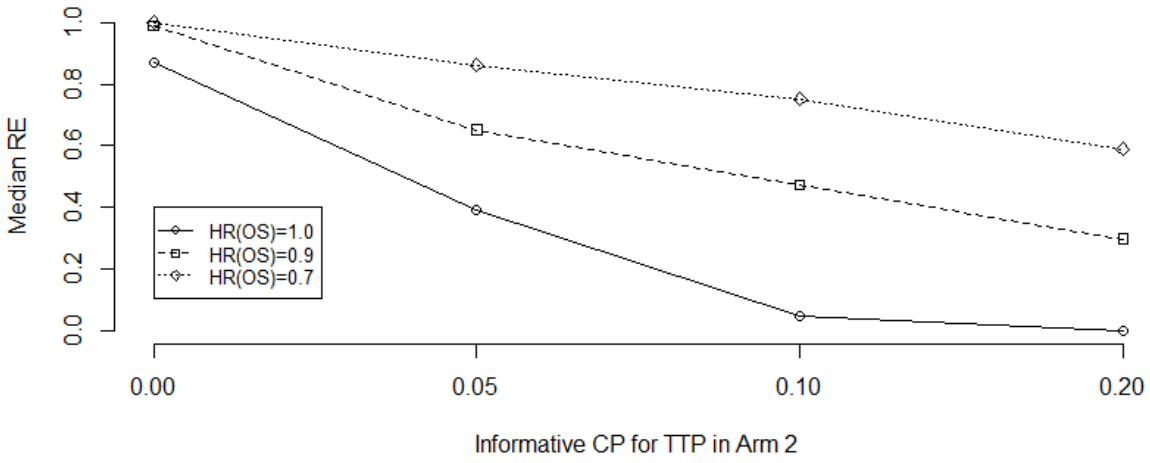

Figure 5.2: Informative CP vs. the median RE with different initial HR values for OS and TTP, when non-informative CP for TTP=0.10 and CP for OS =0.30.

5.1.3 The Effect on Surrogate Threshold Effect

The output for the STE values in Table 5.1 to 5.5 show that STE decreases as informative censoring for TTP increases. From Table 5.1, 5.3, and 5.5, one can observe that when the other variables are constant, the STE value increases as HR_{OS} and the initial value of HR_{TTP} decreases. That is, the greater the treatment effect on both endpoints, the greater the value of STE. This association can also be seen in Figure 5.3.

Figure 5.4 shows the effect of informative censoring on the STE with different levels of censoring for OS. This graph shows that the STE value decreases as the percentage of OS censoring increases. This observation can be explained using the positive association between the censoring rate for OS and the variability of $logHR_{OS}$. The calculation of STE involves the variance of $logHR_{OS}$, and the greater the variance, the smaller the STE. Therefore, as the censoring for OS increases, the variance of $logHR_{OS}$ increases, which decreases the STE value at the same time.

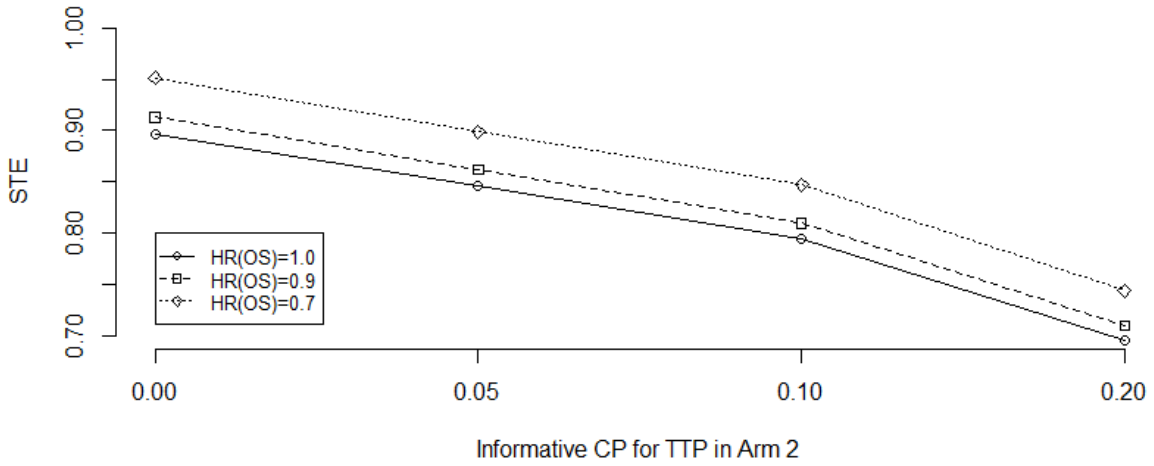

Figure 5.3: Informative CP vs. the STE with different initial HR values for OS and TTP, when non-informative CP for TTP=0.10 and CP for OS =0.30.

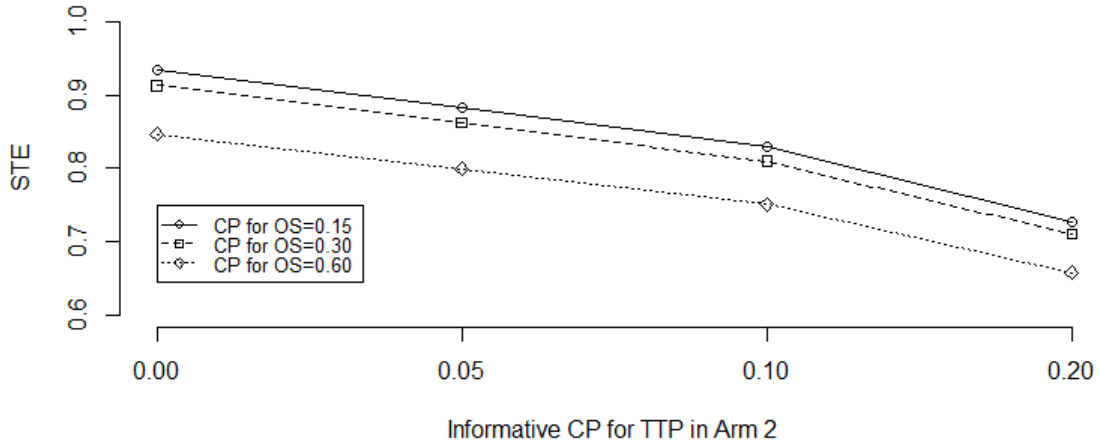

Figure 5.4: Informative CP vs. the STE with different CP for OS, when the initial HR(OS)=HR(TTP)=0.90 and non-informative CP for TTP=0.10.

5.1.4 The Effect on the Difference Between the Log Hazard Ratio for OS and for TTP

Figure 5.5 shows the effect of informative censoring on IRE with different levels of non-informative censoring. IRE was positively associated with the informative censoring rate for TTP, and for the same level of informative censoring, IRE increases by a little bit as the non-informative censoring rate for TTP increases. Regardless of the level of OS censoring and the initial value of HR_{OS} and HR_{TTP}, similar values of the median IRE and the 95% CI for IRE were observed. This implies that the OS censoring rate, and the initial treatment effect on OS and TTP have minimal effect on the IRE values.

RE expressed $logHR_{OS}$ relative to $logHR_{TTP}$, and IRE represented the difference between $logHR_{OS}$ and $logHR_{TTP}$. As mentioned previously, RE is not very stabilized when $HR_{TTP} \approx 1$. Hence, to fix this problem, IRE was used as an alternative statistics for RE.

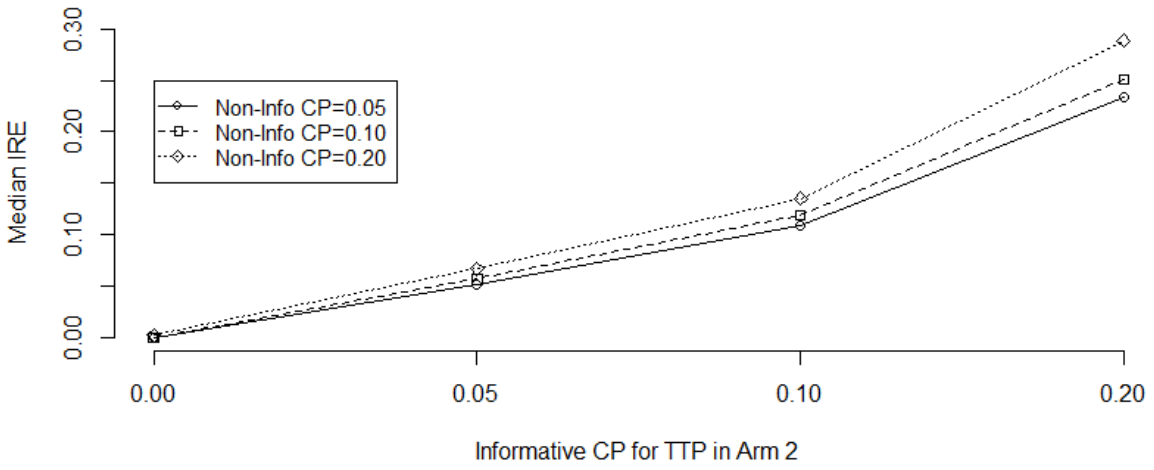

Figure 5.5: Informative CP vs. the median IRE with different non-informative CP for TTP, when the initial HR(OS)=HR(TTP)=0.70 and CP for OS=0.30.

Chapter 6

Discussion

Using OS, the universally agreed clinical endpoint, as the primary endpoint for a clinical cancer trial could be time consuming and delay the implementation of an effective new therapy. Thus, early markers such as PFS and TTP, have been increasingly used as a surrogate endpoint for OS in phase 3 clinical trials in the past few years. For a surrogate to be clinically useful, investigators must be able to predict the treatment effect on OS based on the treatment effect on the surrogate endpoint. However, some clinical trials have shown the new therapy has significantly improved the surrogate, without a corresponding benefit in OS. Although PFS is known to not be a good surrogate in all cases, it becomes increasingly poor when the results are influenced by informative censoring. If one treatment arm has a greater percentage of informative censoring than the other arm, the estimate of the surrogate endpoint could be biased, which may also bias the results of the validation methods for surrogacy such as PE, RE and STE. As a result, not only are attempts at validating surrogate markers affected by informative censoring, but also inferences from clinical trials which use a surrogate marker as the primary outcome.

A simulation study was conducted to examine how informative censoring could bias the results of a surrogate endpoint, say TTP or PFS, and how that would affect the surrogacy of that endpoint. In the simulation, twenty-seven result tables were constructed, each with four sets of clinical trials having different levels of informative censoring, where the level of censoring for OS, informative censoring for TTP in both arms, HR for OS, and the initial HR for TTP were varied. Each set contained 1000 replicated clinical trials, with 1000 patients in each trial. For each set of trials, the median HR_{PFS}, median RE, median IRE, and STE were calculated. The 95% CI for HR_{PFS}, RE, and IRE were also constructed for each set of clinical trials.

The key outcome observed in the simulation study was that the values of HR_{TTP}, RE, and STE decrease and the values of IRE increase, as the informative censoring rate in arm 2 increases regardless of the level of non-informative censoring, censoring for OS, HR_{OS}, and the initial value of HR_{TTP}. Hence, informative censoring for the surrogate endpoint is indeed, having an effect on the treatment effect as well as the values for the surrogate validation methods, RE, IRE, and STE. The simulation results showed that by having informative censoring in arm 2, the experimental arm, could overestimate the treatment effect on TTP for the new therapy. In some circumstances, a treatment with no treatment effect at all (i.e. $HR_{OS} = HR_{TTP} = 1$) could be shown to have a $HR_{TTP} = 0.78$ with as little as 20% of patients having informative censoring. As the rate of informative censoring increases, the difference between HR_{TTP} and HR_{OS} also increases. In other words, the greater the informative censoring rate for TTP in the experimental arm, the greater the overestimation of the treatment effect on TTP.

This bias was observable looking at all the statistics evaluated. For instance, with RE, a value of 1 was observed in most cases, indicating that TTP would be considered a good potential surrogate for OS. However, as informative censoring was added in

arm 2 , RE decreased, indicating that the treatment effect on TTP can no longer accurately predict the treatment effect on OS. The simulation results also showed that the values of RE can be very unstable when the treatment has little or no effect on both the surrogate and true endpoints. The stability and sensitiveness of RE improves as the treatment effect on OS and TTP increase. The median RE values look the most reasonable and the 95% confidence intervals for RE were the narrowest when HR_{OS} and the initial value of HR_{TTP} were 0.70. Based on this finding, when the treatment effect on both endpoints are small, RE is not an optimal statistical method to use in validating surrogacy for time-to-event endpoints. Other than informative censoring and treatment effect, the values of RE were not affected too much by the other variables in the simulation study.

Similarly, STE was negatively associated with the amount of informative censoring for the surrogate. That is, the greater the proportion of informative censoring for TTP, the smaller the HR value need to be observed for TTP in order to predict a HR value for OS that is smaller than 1. In other words, the greater the amount of informative censoring for TTP, the greater the treatment effect on TTP needed to ensure a non-zero treatment benefit in OS. Aside from informative censoring, the STE values were also noticeably affected by the censoring for OS. When the values of other variables remain constant, STE decreases as the censoring rate for OS increases. This observation has showed that the calculation of STE can be affected if the proportion of censoring for OS is large.

The IRE values were most stable compared to the values of the other two statistical methods in the simulation. The only variable that can greatly affect the values of IRE was informative censoring. As mentioned earlier, IRE increases as the informative censoring rate for TTP increases. When the level of informative censoring remain the same, the values of the median IRE and the 95% confidence intervals for IRE were

very similar regardless of the values of other variables.

In short, this study demonstrated that informative censoring can greatly influence the ability to validate a surrogate marker, and additionally can bias the ability to determine the efficacy of a new therapy from a clinical trial using a surrogate marker as the primary outcome. And the magnitude of the effect depends primarily on the proportion of patients who are informatively censored and secondarily on the treatment effect on OS and on the surrogate before informative censoring was applied.

Appendix A

Hazard Ratio and the Cox Proportional Hazards Regression Model

A.1 Hazard Ratio

In survival analyses that examine and model time-to-event data, hazard ratios are often used to express treatment effects in studies comparing the new treatment with the standard treatment. The hazard ratio is the ratio of the hazard rate of the event (e.g. death or tumor progression) occurring in the treatment arm and the hazard rate of the event occurring in the control arm. Let the hazard rate, denoted by $h(t)$, be the instantaneous event rate for an individual at time t, given that the event has not occurred to the individual prior to the time t. Let the random variable T denote the survival time, with the cumulative distribution function $F(t) = P(T < t)$, the survival function $S(t) = 1 - F(t)$, and the probability density function $f(t) = dF(t)/dt$. Then

the hazard rate $h(t)$ can be expressed as:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t} \quad (A.1)$$

$$= \frac{f(t)}{S(t)}. \quad (A.2)$$

Let HR_{OS} be the hazard ratio of death for the experimental group relative to the control group, then a HR_{OS} that is greater than 1 indicates that the patients in the experimental group has a higher risk of death than the patients in the control group. Equivalently, a HR_{OS} that is less than 1 indicates that the patients in the experimental group has a smaller risk of death compare to the patients in the control group. For example, a HR_{OS} of 1.56 could be interpreted as the risk of the patients who were given the new drug has 56% higher risk in death than the patients who were given placebo. And a HR_{OS} of 0.87 indicates that the patients who received the new drug has 13% decrease in the risk of death compared to the patients who received a placebo.

A.2 Cox Proportional Hazards Regression Model

The Cox proportional hazards regression model is a semi-parametric method which can be used to investigate the effects of several explanatory variables on a time-to-event endpoint. This model is useful in estimating the hazard rate at a given time. Let x_i be the explanatory variables for observation i, and let death be the event of interest. Using the Cox model, the hazard rate for observation i at time t can be estimated by:

$$h_i(t) = \lambda_0(t) exp[\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}], \quad (A.3)$$

or, equivalently,

$$log[h_i(t)] = log[\lambda_0(t)] + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}. \tag{A.4}$$

$\lambda_0(t)$ represents a baseline hazard function, which is the hazard rate for death when all of the x_i's are zero, and it is normally unspecified. The hazard ratio for observation i relative to observation j can be expressed as:

$$\begin{aligned} HR_{i,j} &= \frac{h_i(t)}{h_j(t)} && \text{(A.5)} \\ &= \frac{\lambda_0(t) exp[\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}]}{\lambda_0(t) exp[\beta_1 x_{j1} + \beta_2 x_{j2} + ... + \beta_k x_{jk}]} && \text{(A.6)} \\ &= exp[\beta_1(x_{i1} - x_{j1}) + \beta_2(x_{i2} - x_{j2}) + ... + \beta_k(x_{ik} - x_{jk})]. && \text{(A.7)} \end{aligned}$$

When the only explanatory variable that is considered is the treatment group, the hazard ratio of death for observation i relative to observation j can be estimated as:

$$HR_{i,j} = \frac{h_i(t)}{h_j(t)} = exp[\beta_1(x_i - x_j)], \tag{A.8}$$

where x_i and x_j represents the two different treatment groups.

Appendix B

More Tables from the Simulation Results

Table B.1: $HR_{OS} = HR_{TTP} = 1.00$ before informative censoring,
censoring for OS = 0.15, non-informative censoring for TTP in both arms = 0.05.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.05, Median HR for OS=0.999 (95% CI =0.879, 1.149).					
0.00	0.05	0.999 (0.885, 1.135)	0.901 (-5.961, 7.352)	0.002 (-0.077, 0.071)	0.923
0.05	0.10	0.947 (0.839, 1.077)	0.442 (-7.810, 10.22)	0.055 (-0.023, 0.125)	0.874
0.10	0.15	0.896 (0.793, 1.019)	0.077 (-4.751, 5.822)	0.111 (-0.033, 0.181)	0.826
0.20	0.25	0.789 (0.698, 0.897)	0.004 (-1.193, 0.394)	0.237 (0.159, 0.306)	0.728

Table B.2: $HR_{OS} = HR_{TTP} = 1.00$ before informative censoring,
censoring for OS = 0.15, non-informative censoring for TTP in both arms = 0.10.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.10, Median HR for OS=0.998 (95% CI =0.881, 1.150).					
0.00	0.10	0.999 (0.882, 1.145)	0.866 (-6.449, 8.749)	0.001 (-0.082, 0.076)	0.915
0.05	0.15	0.943 (0.833, 1.078)	0.398 (-5.947, 8.595)	0.059 (-0.024, 0.133)	0.863
0.10	0.20	0.887 (0.780, 1.019)	0.064 (-6.983, 4.285)	0.120 (0.037, 0.195)	0.811
0.20	0.30	0.777 (0.685, 0.891)	0.009 (-1.069, 0.374)	0.252 (0.170, 0.328)	0.710

Table B.3: $HR_{OS} = HR_{TTP} = 1.00$ before informative censoring,
censoring for OS = 0.15, non-informative censoring for TTP in both arms = 0.20.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.20, Median HR for OS=0.998 (95% CI =0.878, 1.147).					
0.00	0.20	0.998 (0.884, 1.150)	0.759 (-4.701, 6.923)	0.002 (-0.098, 0.083)	0.896
0.05	0.25	0.934 (0.828, 1.076)	0.349 (-6.326, 6.458)	0.0664 (-0.032, 0.149)	0.837
0.10	0.30	0.873 (0.771, 1.009)	0.049 (-4.976, 2.110)	0.135 (0.038, 0.224)	0.780
0.20	0.40	0.749 (0.659, 0.866)	0.007 (-0.847, 0.347)	0.287 (-0.190, 0.376)	0.667

Table B.4: $HR_{OS} = HR_{TTP} = 1.00$ before informative censoring,
censoring for OS = 0.30, non-informative censoring for TTP in both arms = 0.05.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.05, Median HR for OS=0.998 (95% CI =0.870, 1.150).					
0.00	0.05	0.999 (0.885, 1.135)	0.899 (-6.835, 9.994)	0.002 (-0.094, 0.085)	0.907
0.05	0.10	0.947 (0.839, 1.077)	0.456 (-9.616, 12.80)	0.055 (-0.042, 0.139)	0.859
0.10	0.15	0.896 (0.793, 1.019)	0.065 (-5.357, 5.216)	0.111 (0.013, 0.196)	0.811
0.20	0.25	0.789 (0.698, 0.897)	0.007 (-1.193, 0.453)	0.237 (0.140, 0.323)	0.715

Table B.5: $HR_{OS} = HR_{TTP} = 1.00$ before informative censoring,
censoring for OS = 0.30, non-informative censoring for TTP in both arms = 0.20.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.20, Median HR for OS=0.998 (95% CI =0.865, 1.159).					
0.00	0.20	0.998 (0.884, 1.150)	0.764 (-5.693, 8.492)	-0.001 (-0.110, 0.095)	0.879
0.05	0.25	0.934 (0.828, 1.076)	0.332 (-7.947, 6.798)	0.065 (-0.046, 0.163)	0.821
0.10	0.30	0.873 (0.771, 1.009)	0.049 (-5.089, 1.647)	0.133 (0.020, 0.223)	0.764
0.20	0.40	0.749 (0.659, 0.865)	0.008 (-0.879, 0.396)	0.287 (-0.170, 0.383)	0.654

Table B.6: $HR_{OS} = HR_{TTP} = 1.00$ before informative censoring,
censoring for OS = 0.60, non-informative censoring for TTP in both arms = 0.05.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.05, Median HR for OS=0.998 (95% CI =0.824, 1.234).					
0.00	0.05	0.999 (0.885, 1.135)	1.010 (-9.259, 17.84)	-0.001 (-0.153, 0.154)	0.850
0.05	0.10	0.947 (0.839, 1.077)	0.489 (-14.13, 13.77)	0.053 (-0.101, 0.209)	0.805
0.10	0.15	0.896 (0.793, 1.018)	0.100 (-7.309, 5.360)	0.109 (-0.046, 0.266)	0.761
0.20	0.25	0.789 (0.698, 0.897)	0.010 (-1.627, 0.665)	0.236 (0.080, 0.390)	0.671

Table B.7: $HR_{OS} = HR_{TTP} = 1.00$ before informative censoring,
censoring for OS = 0.60, non-informative censoring for TTP in both arms = 0.10.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.10, Median HR for OS=0.995 (95% CI =0.823, 1.231).					
0.00	0.10	0.999 (0.882, 1.145)	1.020 (-11.44, 22.56)	0.000 (-0.161, 0.165)	0.838
0.05	0.15	0.943 (0.833, 1.078)	0.451 (-8.932, 13.27)	0.057 (-0.106, 0.218)	0.791
0.10	0.20	0.887 (0.780, 1.020)	0.103 (-7.840, 3.126)	0.120 (-0.046, 0.282)	0.743
0.20	0.30	0.777 (0.685, 0.891)	0.022 (-1.093, 0.435)	0.252 (0.148, 0.341)	0.651

Table B.8: $HR_{OS} = HR_{TTP} = 1.00$ before informative censoring,
censoring for OS = 0.60, non-informative censoring for TTP in both arms = 0.20.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.20, Median HR for OS=0.998 (95% CI =0.865, 1.159).					
0.00	0.20	0.998 (0.884, 1.150)	0.839 (-8.622, 15.10)	0.001 (-0.166, 0.166)	0.819
0.05	0.25	0.934 (0.828, 1.076)	0.418 (-10.58, 10.63)	0.065 (-0.098, 0.237)	0.765
0.10	0.30	0.873 (0.771, 1.009)	0.061 (-5.817, 2.094)	0.134 (-0.034, 0.309)	0.713
0.20	0.40	0.749 (0.659, 0.865)	0.008 (-1.133, 0.564)	0.288 (0.117, 0.453)	0.609

Table B.9: $HR_{OS} = HR_{TTP} = 0.90$ before informative censoring,
censoring for OS = 0.15, non-informative censoring for TTP in both arms = 0.05.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.05, Median HR for OS=0.895 (95% CI =0.787, 1.031).					
0.00	0.05	0.899 (0.796, 1.022)	1.000 (-0.953, 2.874)	-0.001 (-0.076, 0.064)	0.940
0.05	0.10	0.851 (0.754, 0.970)	0.687 (-0.597, 1.197)	0.052 (-0.024, 0.118)	0.890
0.10	0.15	0.805 (0.713, 0.917)	0.510 (-0.267, 0.837)	0.108 (0.033, 0.175)	0.842
0.20	0.25	0.710 (0.629, 0.809)	0.324 (-0.138, 0.549)	0.234 (0.159, 0.303)	0.742

Table B.10: $HR_{OS} = HR_{TTP} = 0.90$ before informative censoring, censoring for OS = 0.15, non-informative censoring for TTP in both arms = 0.20.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.20, Median HR for OS=0.894 (95% CI =0.790, 1.030).					
0.00	0.20	0.899 (0.795, 1.037	0.974 (-1.837, 4.251)	-0.002 (-0.099, 0.080)	0.931
0.05	0.25	0.842 (0.745, 0.969)	0.640 (-0.396, 1.364)	0.062 (-0.034, 0.147)	0.871
0.10	0.30	0.786 (0.694, 0.908)	0.456 (-0.267, 0.828)	0.132 (0.037, 0.220)	0.813
0.20	0.40	0.675 (0.593, 0.779)	0.282 (-0.110, 0.504)	0.284 (0.192, 0.372)	0.697

Table B.11: $HR_{OS} = HR_{TTP} = 0.90$ before informative censoring, censoring for OS = 0.30, non-informative censoring for TTP in both arms = 0.05.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.05, Median HR for OS=0.897 (95% CI =0.782, 1.041).					
0.00	0.05	0.899 (0.796, 1.022)	0.995 (-1.094, 3.779)	-0.001 (-0.096, 0.085)	0.917
0.05	0.10	0.851 (0.754, 0.970)	0.672 (-0.746, 1.387)	0.052 (-0.043, 0.138)	0.869
0.10	0.15	0.805 (0.713, 0.917)	0.495 (-0.341, 0.943)	0.109 (0.012, 0.194)	0.821
0.20	0.25	0.710 (0.629, 0.809)	0.316 (-0.165, 0.602)	0.235 (0.139, 0.321)	0.724

Table B.12: $HR_{OS} = HR_{TTP} = 0.90$ before informative censoring, censoring for OS = 0.30, non-informative censoring for TTP in both arms = 0.20.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.20, Median HR for OS=0.899 (95% CI =0.776, 1.045).					
0.00	0.20	0.899 (0.795, 1.037)	0.952 (-3.094, 4.218)	0.001 (-0.113, 0.094)	0.908
0.05	0.25	0.842 (0.745, 0.969)	0.629 (-0.635, 1.534)	0.065 (-0.047, 0.163)	0.850
0.10	0.30	0.786 (0.694, 0.908)	0.442 (-0.363, 0.884)	0.134 (0.017, 0.235)	0.792
0.20	0.40	0.675 (0.593, 0.779)	0.273 (-0.163, 0.555)	0.287 (0.172, 0.387)	0.679

Table B.13: $HR_{OS} = HR_{TTP} = 0.90$ before informative censoring, censoring for OS = 0.60, non-informative censoring for TTP in both arms = 0.05.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.05, Median HR for OS=0.896 (95% CI =0.742, 1.110).					
0.00	0.05	0.899 (0.796, 1.022)	0.996 (-3.732, 6.075)	-0.001 (-0.153, 0.153)	0.854
0.05	0.10	0.851 (0.754, 0.970)	0.684 (-1.928, 1.904)	0.053 (-0.099, 0.206)	0.809
0.10	0.15	0.805 (0.713, 0.917)	0.507 (-0.904, 1.274)	0.109 (-0.043, 0.263)	0.764
0.20	0.25	0.710 (0.629, 0.809)	0.322 (-0.434, 0.762)	0.235 (0.081, 0.391)	0.674

Table B.14: $HR_{OS} = HR_{TTP} = 0.90$ before informative censoring, censoring for OS = 0.60, non-informative censoring for TTP in both arms = 0.20.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.20, Median HR for OS=0.897 (95% CI =0.743, 1.101).					
0.00	0.20	0.899 (0.795, 1.037)	0.958 (-6.866, 6.285)	-0.001 (-0.163, 0.162)	0.840
0.05	0.25	0.842 (0.745, 0.969)	0.628 (-1.450, 2.075)	0.064 (-0.100, 0.230)	0.785
0.10	0.30	0.786 (0.694, 0.908)	0.449 (-0.701, 1.198)	0.133 (-0.036, 0.303)	0.733
0.20	0.40	0.675 (0.593, 0.779)	0.280 (-0.341, 0.681)	0.286 (0.119, 0.449)	0.627

Table B.15: $HR_{OS} = HR_{TTP} = 0.70$ before informative censoring, censoring for OS = 0.15, non-informative censoring for TTP in both arms = 0.05.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.05, Median HR for OS=0.695 (95% CI =0.610, 0.797).					
0.00	0.05	0.699 (0.617, 0.794)	1.000 (0.822, 1.245)	-0.004 (-0.077, 0.063)	0.967
0.05	0.10	0.663 (0.586, 0.751)	0.883 (0.709, 1.062)	0.049 (-0.023, 0.118)	0.917
0.10	0.15	0.626 (0.553, 0.711)	0.776 (0.609, 0.928)	0.104 (0.032, 0.172)	0.867
0.20	0.25	0.552 (0.487, 0.629)	0.611 (0.454, 0.739)	0.230 (0.158, 0.298)	0.766

Table B.16: $HR_{OS} = HR_{TTP} = 0.70$ before informative censoring, censoring for OS = 0.15, non-informative censoring for TTP in both arms = 0.10.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.10, Median HR for OS=0.695 (95% CI =0.613, 0.800).					
0.00	0.10	0.698 (0.615, 0.800)	1.000 (0.802, 1.289)	-0.003 (-0.087, 0.074)	0.977
0.05	0.15	0.660 (0.583, 0.754)	0.870 (0.690, 1.081)	0.054 (-0.030, 0.130)	0.924
0.10	0.20	0.620 (0.546, 0.709)	0.758 (0.588, 0.928)	0.115 (0.030, 0.192)	0.871
0.20	0.30	0.543 (0.479, 0.621)	0.598 (0.441, 0.725)	0.248 (0.162, 0.326)	0.765

Table B.17: $HR_{OS} = HR_{TTP} = 0.70$ before informative censoring, censoring for OS = 0.15, non-informative censoring for TTP in both arms = 0.20.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.20, Median HR for OS=0.697 (95% CI =0.615, 0.801).					
0.00	0.20	0.699 (0.615, 0.803)	0.999 (0.778, 1.334)	0.000 (-0.094, 0.086)	1.000
0.05	0.25	0.654 (0.576, 0.753)	0.843 (0.651, 1.087)	0.066 (-0.027, 0.153)	0.941
0.10	0.30	0.612 (0.537, 0.707)	0.726 (0.555, 0.908)	0.134 (0.042, 0.223)	0.881
0.20	0.40	0.525 (0.461, 0.608)	0.556 (0.407, 0.687)	0.287 (0.195, 0.377)	0.761

Table B.18: $HR_{OS} = HR_{TTP} = 0.70$ before informative censoring, censoring for OS = 0.30, non-informative censoring for TTP in both arms = 0.05.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.05, Median HR for OS=0.697 (95% CI =0.604, 0.809).					
0.00	0.05	0.699 (0.617, 0.794)	1.000 (0.775, 1.300)	-0.001 (-0.097, 0.083)	0.940
0.05	0.10	0.663 (0.586, 0.751)	0.874 (0.653, 1.106)	0.052 (-0.046, 0.137)	0.891
0.10	0.15	0.626 (0.553, 0.711)	0.770 (0.560, 0.975)	0.109 (0.011, 0.194)	0.843
0.20	0.25	0.552 (0.487, 0.629)	0.606 (0.428, 0.770)	0.234 (0.136, 0.321)	0.745

Table B.19: $HR_{OS} = HR_{TTP} = 0.70$ before informative censoring, censoring for OS = 0.30, non-informative censoring for TTP in both arms = 0.20.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.20, Median HR for OS=0.701 (95% CI =0.604, 0.815).					
0.00	0.20	0.699 (0.615, 0.803)	0.990 (0.730, 1.367)	0.003 (-0.108, 0.102)	0.975
0.05	0.25	0.656 (0.577, 0.753)	0.841 (0.614, 1.122)	0.067 (-0.047, 0.169)	0.917
0.10	0.30	0.613 (0.539, 0.709)	0.724 (0.509, 0.953)	0.135 (0.020, 0.238)	0.858
0.20	0.40	0.526 (0.463, 0.609)	0.553 (0.379, 0.722)	0.288 (0.173, 0.395)	0.741

Table B.20: $HR_{OS} = HR_{TTP} = 0.70$ before informative censoring, censoring for OS = 0.60, non-informative censoring for TTP in both arms = 0.05.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.05, Median HR for OS=0.698 (95% CI =0.576, 0.868).					
0.00	0.05	0.699 (0.617, 0.794)	0.991 (0.530, 1.474)	0.003 (-0.150, 0.160)	0.859
0.05	0.10	0.663 (0.586, 0.751)	0.865 (0.453, 1.266)	0.057 (-0.097, 0.211)	0.814
0.10	0.15	0.626 (0.553, 0.711)	0.765 (0.381, 1.100)	0.113 (-0.042, 0.270)	0.770
0.20	0.25	0.552 (0.487, 0.629)	0.603 (0.283, 0.862)	0.238 (0.083, 0.397)	0.680

Table B.21: $HR_{OS} = HR_{TTP} = 0.70$ before informative censoring, censoring for OS = 0.60, non-informative censoring for TTP in both arms = 0.10.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.10, Median HR for OS=0.699 (95% CI =0.579, 0.862).					
0.00	0.10	0.698 (0.615, 0.800)	0.993 (0.509, 1.493)	0.003 (-0.151, 0.157)	0.866
0.05	0.15	0.660 (0.583, 0.754)	0.859 (0.437, 1.262)	0.060 (-0.097, 0.220)	0.818
0.10	0.20	0.620 (0.546, 0.709)	0.752 (0.368, 1.084)	0.122 (-0.039, 0.279)	0.770
0.20	0.30	0.543 (0.479, 0.621)	0.589 (0.278, 0.840)	0.254 (0.095, 0.412)	0.676

Table B.22: $HR_{OS} = HR_{TTP} = 0.70$ before informative censoring, censoring for OS = 0.60, non-informative censoring for TTP in both arms = 0.20.

Informative Censoring for TTP in Arm 2	Censoring for TTP in Arm 2	Median HR for TTP (95% CI)	Median RE (95% CI)	Median IRE (95% CI)	STE
Censoring for TTP in arm 1 = 0.20, Median HR for OS=0.702 (95% CI =0. 579, 0.862).					
0.00	0.20	0.699 (0.615, 0.803)	0.981 (0.515, 1.517)	0.007 (-0.151, 0.168)	0.879
0.05	0.25	0.656 (0.577, 0.753)	0.838 (0.425, 1.254)	0.068 (-0.096, 0.232)	0.826
0.10	0.30	0.613 (0.539, 0.709)	0.719 (0.359, 1.067)	0.136 (-0.028, 0.302)	0.772
0.20	0.40	0.526 (0.463, 0.609)	0.553 (0.267, 0.800)	0.291 (0.121, 0.452)	0.666

Bibliography

Baselga, J., Campone, M., Piccart, M. *et al.* (2012). Everolimus in Postmenopausal Hormone-Receptor-Positive Advanced Breast Cancer. *New England Journal of Medicine*, **366**, 520-529.

Bast, R. C., Thigpen, J. T., Arbuck, S.G. *et al.* (2007). Clinical trial endpoints in ovarian cancer: report of an FDA/ASCO/AACR Public Workshop. *Gynecologic Oncology*, **107**(2),173-176.

Berger, V. W. (2005). Censored Observations. In *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd.

Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, **69**, 89-95.

Booth, C.M. and Eisenhauer, E.A. (2012). Progression-Free Survival: Meaningful or Simply Measurable? *Journal of Clinical Oncology*, **30**, 1030-1033.

Burzykowski, T. and Buyse, M. (2006). Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics*, **5**, 173-186.

Burzykowski, T., Buyse, M. Piccart, M. J. *et al.* (2008). Evaluation of Tumor Response, Disease Control, Progression-Free Survival, and Time to Progression As Potential Surrogate End Points in Metastatic Breast Cancer. *Journal of Clinical Oncology*, **26**, 1987-1992.

Buyse, M. and Molenberghs, G. (1998). Criteria for the Validation of Surrogate Endpoints in Randomized Experiment. *Biometrics*, **54**, 1014-1029.

Buyse, M., Molenberghs, G., Burzykowski, T. *et al.* (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, **1**(1), 49-67.

Buyse, M., Thirion, P., Carlson, R. W. *et al.* (2000). Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. *The Lancet*, **356**, 373-378.

Buyse, M., Burzykowski, T., Carroll, K. *et al.* (2007). Progression-Free Survival Is a Surrogate for Survival in Advanced Colorectal Cancer. *Journal of Clinical Oncology*, **25**, 5218-5224.

Buyse, M. (2008). Validation of surrogate markers and endpoints in clinical trials. Seminar power point, Harvard School of Public Health.

Buyse, M., Sargent, D. J., Grothey, A. *et al.* (2010). Biomarkers and surrogate end points–the challenge of statistical validation. *Nature Reviews Clinical Oncology*, **7**(6), 309-317.

Buyse, M., Michiels, S., Squifflet, P. *et al.* (2011). Leukemia-free survival as a surrogate end point for overall survival in the evaluation of maintenance therapy

for patients with acute myeloid leukemia in complete remission. *Haematologica*, **96**(8), 1106-1112.

Collette, L., Burzykowski, T., Carroll, K. J. *et al.* (2005). Is prostate-specific antigen a valid surrogate end point for survival in hormonally treated patients with metastatic prostate cancer? Joint research of the European Organisation for Research and Treatment of Cancer, the Limburgs Universitair Centrum, and AstraZeneca Pharmaceuticals. *Journal of Clinical Oncology*, **23**, 6139-6148.

Denne, J. S., Stone, A. M., Bailey-Iacona, R. and Chen, T. T. (2013). Missing data and censoring in the analysis of progression-free survival in oncology clinical trials. *Journal of Biopharmaceutical Statistics*, **23**, 951-970.

Eisenhauer, E. A., 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**, 228 - 247.

Fleming, T. R., Prentice, R. L., Pepe, M. S., and Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine*, **13**, 955-968.

Fleming, T. R. and DeMets, D.L. (1996). Surrogate end points in clinical trials: are we being misled. *Annals of Internal Medicine*, **125**(7), 605-613.

Food and Drug Administration. (2007). FDA Guidance for Industry Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatory Information/Guidances/ucm071590.pdf, accessed May 2007.

Food and Drug Administration. (2012). FDA approves Afinitor for advanced breast cancer. Available at: http://www.fda.gov/NewsEvents/Newsroom/Press Announcements/ucm312965.htm, accessed July 2012.

Freedman, L. S., Graubard, B. I., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, **11**, 167-178.

GASTRIC. (2013). Role of chemotherapy for advanced/recurrent gastric cancer: An individual-patient-data meta-analysis. *European Journal of Cancer*, **49**(7), 1565-1577.

Kelly, W.K., Halabi, S., Carducci, M. *et al.* (2012). Randomized, Double-Blind, Placebo-Controlled Phase III Trial Comparing Docetaxel and Prednisone With or Without Bevacizumab in Men With Metastatic Castration-Resistant Prostate Cancer: CALGB 90401. *Journal of Clinical Oncology* , **30**(13), 1534-1540.

Laporte, S., Squifflet, P., Baroux, N. *et al.* (2013). Prediction of survival benefits from progression-free survival benefits in advanced non-small-cell lung cancer: evidence from a meta-analysis of 2334 patients from 5 randomised trials. *BMJ Open*, **3**.

Lesko, L.J. and Atkinson A.J. (2001). Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. *Annual Review of Pharmacology and Toxicology*, **41**, 347-366.

Lin, D. Y., Fleming, T. R., and DeGruttola, V. (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine*, **16**(13), 1515-1527.

Michiels, S., LeMaitre, A., Buyse, M. *et al.* (2009). Surrogate endpoints for overall survival in locally advanced head and neck cancer: meta-analyses of individual patient data. *The Lancet Oncology*, **10**(4), 341-350.

Miller, K., Wang, M., Gralow, J. *et al.* (2007). Paclitaxel plus Bevacizumab versus Paclitaxel Alone for Metastatic Breast Cancer. *New England Journal of Medicine*, **357**, 2666-2676.

Oba, K., Paoletti, X., Alberts, S. *et al.* (2013). Disease-free survival as a surrogate for overall survival in adjuvant trials of gastric cancer: a meta-analysis. *Journal of the National Cancer Institute*, **105**(21), 1600-1607.

Oxnard, G. R., Morris, M. J., Hodi, F. S. *et al.* (2012). When Progressive Disease Does not mean treatment Failure: reconsidering the Criteria for Progression. *Journal of the National Cancer Institute*, **104**, 1534-1541.

Paoletti, X., Oba, K., Bang, Y. J. *et al.* (2013). Progression-Free Survival as a Surrogate for Overall Survival in Advanced/Recurrent Gastric Cancer Trials: A Meta-Analysis. *Journal of the National Cancer Institute*, **105**(21), 1667-1670.

Piccart, M., Hortobagyi, G. N., Campone, M. *et al.* (2014). Everolimus Plus Exemestane for Hormone Receptor-Positive (HR+), Human Epidermal Growth Factor Receptor-2-Negative (HER2-) Advanced Breast Cancer (BC): Overall Survival Results From BOLERO-2. 9th European Breast Cancer Conference, abstract 1LBA.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definitions and operational criteria. *Statistics in Medicine*, **8**, 431-440.

Rugo, H. S., Pritchard, K. I., Gnant, M. *et al.* (2014). Incidence and time course of everolimus-related adverse events in postmenopausal women with hormone receptor-positive advanced breast cancer: insights from BOLERO-2 *Annals of Oncology*, **25**, 808-815.

Sargent, D. J., Wieand, H. S., Haller, D. G. *et al.* (2005). Disease-Free Survival Versus Overall Survival As a Primary End Point for Adjuvant Colon Cancer Studies: Individual Patient Data From 20,898 Patients on 18 Randomized Trials. *Journal of Clinical Oncology*, **23**, 8664-8670.

Tang, P. A., Bentzen, S. M., Chen, E. X. *et al.* (2007). Surrogate end points for median overall survival in metastatic colorectal cancer: Literature-based analysis from 39 randomized controlled trials of first-line chemotherapy. *Journal of Clinical Oncology*, **25**, 4562-4568.

Yardley, D. A., Noguchi, S., Pritchard K. I. *et al.* (2013). Everolimus Plus Exemestane in Postmenopausal Patients with HR+ Breast Cancer: BOLERO-2 Final Progression-Free Survival Analysis. *Advances in Therapy*, **30**(10), 870-884.