UNIVERSITA' DEGLI STUDI DI TORINO:

DIPARTIMENTO DI SCIENZE MATEMATICHE "Giuseppe Luigi Lagrange"

DOTTORATO DI RICERCA IN : Pure and applied mathematics

CICLO: XXX

TITOLO DELLA TESI: Evaluation of Surrogate Endpoints with Applications in Respiratory Clinical Trials

TESI PRESENTATA DA: Maryam Zolghadr

Tutor(s): Prof. Mauro Gasparini

Coordinatore del Dottorato prof. Riccardo Adami

ANNI ACCADEMICI: 01/11/2014 - 31/10/2017

SETTORE SCIENTIFICO-DISCIPLINARE DI AFFERENZA*: SECS-S/01 STATISTICA

Thesis for the PhD degree in pure and applied mathematics

Evaluation of Surrogate Endpoints with Applications in Respiratory Clinical Trials

Maryam Zolghadr

Supervisor: Prof. Mauro Gasparini[1]
Co-supervisors: Prof. Ziad Taib[2] and Alexandra Jauhiainen [3]

UNIVERSITÀ DEGLI STUDI DI TORINO

Department of Mathematical Sciences
Politecnico di Torino and Universita degli studi di Torino. Italy, Spring 2018

[1]Dept. of Mathematical Sciences "G. L. Lagrange", Politecnico di Torino, Turin, Italy
[2]AstraZeneca, Molndal, Sweden
[3]AstraZeneca, Molndal, Sweden

Evaluation of Surrogate Endpoints with Applications in Respiratory Clinical Trials

Maryam Zolghadr

Department of Mathematical Sciences
Politecnico di Torino and Universita degli studi di Torino.
10129 Turin, Italy
Phone: +39 011 090 6666

Typeset with LaTeX.
Department of Mathematical Sciences
Printed in Turin, Italy 2018

Acknowledgment

I would like to express my gratitude to my supervisors Prof. Mauro Gasparini, Alexandra Jauhiainen and Prof. Ziad Taib for the continuous support of my PhD study and research.

Beside my supervisors, I would like to thank Gaelle for fruitful discussions and her insightful comments.

I would also like to thank my friends Mahsa, Saeed, Aida, Tayyab, Lorenzo, Abouzar, Marjan and Nasim for being wonderful friends and always make me happy.

Last but not least, I would like to thank my brother and parents for supporting me throughout my life.

Maryam Zolghadr
Turin, September 17, 2018

Contents

CHAPTER 1

Introduction

1.1 Introduction

In clinical studies, the effect of a new medical therapy is often determined by comparing the treatment group to the control group with respect to some outcome measurements. These outcomes are called clinical endpoints. For example, in most cancer trials, the survival time is a gold standard endpoint. A new drug is considered effective if the survival time of patients in the new treatment group is significantly longer than in the control group. To demonstrate treatment effects, trials based on the clinical endpoints usually require a large number of subjects and extended follow up period, which might be impractical or even unethical in certain circumstances. Therefore, there is a need to identify alternative outcome measurements that can provide cost effective ways of assessing therapeutic effects. Later in this chapter, we will call these outcomes surrogate endpoints and explain more about advantages of using them as a replacement for the clinical endpoint.

1.2 Clinical Trials

Clinical trials are research experiments on volunteer patients (humans) that investigate whether new drugs, therapies and medical devices are safe and beneficial for humans or not. In clinical studies, drugs, therapies and medical devices are called interventions (treatments). Clinical trials may be used to compare the current intervention with the new one and to explore which one works best for specific group of patients. In general, results from clinical trials aim to improve medical knowledge and health care services. For more details see ClinicalTrials.gov (2017).

1.3 Clinical Endpoints

Outcome measures in clinical trials to determine directly whether the treatment being studied is beneficial, are called clinical or true endpoints, see e.g. AIDSinfo (2017).

In other words, a true endpoint is a charactristic or a variable that reflects how a patient feels, functions or survive. Some examples of true endpoints are survival time, relief of symptoms and stroke. For more examples on clinical endpoints see e.g. Bushe et al. (2010), Neaton et al. (1994) and Troughton et al. (2000). We will use clinical endpoint and true endpoint interchangeably in this thesis.

In practice, measuring clinical endpoints is not usually an easy task. For example, recording survival times usually requires a long follow up of patients and occurrence of some clinical endpoints might be rare among patients. As a consequence, a new drug or therapy needs to stay out of market for a long time until readouts of the relevant trials become available. For this reason, there is a need to search for reliable alternatives to replace true endpoints in order to perform the study faster and easier.

1.4 Biomarkers

According to Downing (2000), a biological marker also known as a biomarker is defined as a characteristic that is objectively measured and evaluated as an indication of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.

Aronson (2005) summarized the main advantages of biomarkers in clinical studies for drug development and clinical decision making as follows:

- It is usually cheaper and easier to measure Biomarkers rather than true endpoints. For example, it is easier to measure a patient's systolic blood pressure (SBP) rather than using echocardiography to determine left ventricular function, and it is much easier to do echocardiography than to measure occurrence of stroke and finally mortality in the long term.

- Biomarkes can be measured prior to a true endpoint. Usually, it can take several years to measure a true endpoint, while biomarkers can be measured much faster than the true endpoint. For example, it may take many years that a patient would experience stroke. However, measuring blood pressure can be done much faster and easier.

- Smaller sample size is another benefit of using biomarkers in clinical studies instead of true endpoints. For example, to measure the effect of a new drug on blood pressure, a sample of size 100 or 200 may be enough to do the study. However, to determine the effect of a new drug on the prevention of death from stroke a much larger sample should be taken to proceed with the study.

- Sometimes, measuring true endpoints may be involved with ethical problems. For example, in paracetamol overdose it is unethical to wait for evidence of liver damage before deciding whether or not to treat a patient. Instead a pharmacological biomarker, the plasma paracetamol concentration, is used to predict whether treatment is required.

Biomarkers have the potential to enhance the research and development process of new treatments by providing new approaches to measure disease activity. They maybe used to diagnose a disease, monitor progression of a disease, and to show how the body is responding towards a new given treatment. Any change in biomarkers during treatment period, may be used to predict clinical benefit (harm) from the treatment. Therefore, they are increasingly used in medicine, and many potential biomarkers are proposed every year. More details on biomarkers and their influence on drug development can be found in Katz (2004) and U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER) (2014).

1.5 Surrogate Endpoints

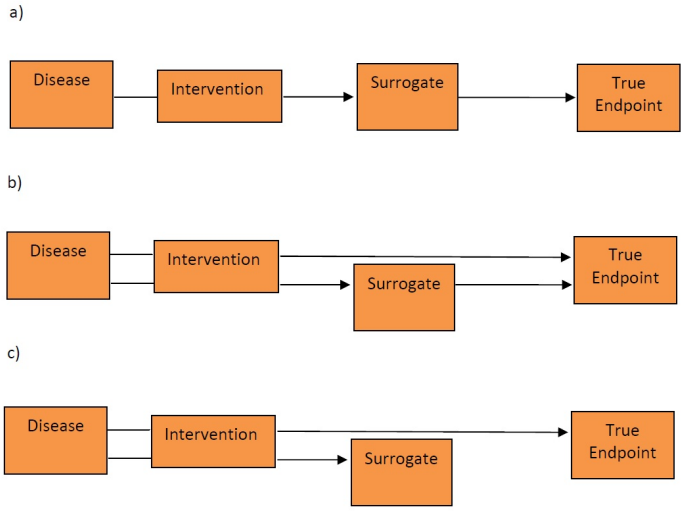
A surrogate endpoint is defined by U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER) (2014), as a biomarker intended to substitute for a clinical meaningful endpoint. Given an intervention to a patient, useful surrogate endpoints are expected to predict the effect of the intervention on a clinical endpoint. In other words, a surrogate endpoint is an intermediate laboratory measure or physical sign that is not direct measure of the clinical benefit, however it reflects the outcome of interest.

It is important to note that, not all biomarkers are surrogate endpoints. In fact, surrogate endpoints are a small subset of well defined biomarkers that very well evaluates clinical relevance. For a biomarker to be considered as a surrogate endpoint, it must be able to predict clinical benefit (or harm or lack of benefit or harm) consistently and accurately. One could use the guidelines suggested by Austin Bradford Hill for choosing biomarkers that could be considered as good candidates for surrogate endpoints, see Hill (1965) and Legator and Morris (2003) for more details.

In clinical studies, when a clinical endpoint is inaccessible due to cost, time, or difficulty of measurement, clinical researchers favor a surrogate endpoint. A useful surrogate endpoint has to bear some nice properties. Piantadosi (2005) summarized some characteristics for a useful surrogate endpoint, as follows:

- It can be measured simply, with less cost and before the true endpoint, Ellenberg and Hamilton (1989).
- It will result in the same inference about the disease and intervention as the true endpoint.
- It is responsive to interventions.
- It has to be correlated with the true endpoint. However, note that the correlation alone in not sufficient, Fleming and DeMets (1996).
- It has to lie on the causal pathway for the true endpoint. Figure 1.1 shows possible relations between disease, an intervention, a candidate biomarker as a surrogate endpoint for the true endpoint, and the true

Figure 1.1: Possible relations between disease, intervention, surrogate and true endpoint.



endpoint. Figure 1.1: part (a), displays the ideal setting for surrogacy. According to Fleming and DeMets (1996), a biomarker has full surrogate value if "it is in the only causal pathway of the disease process, and the intervention's entire effect on the true clinical outcome is mediated through its effect on the surrogate". Figure 1.1: part (b), shows that some effect of the intervention on the clinical endpoint is mediated through the biomarker. This implies that the biomarker has partial surrogate value for the true endpoint. Eventually, figure 1.1: part (c), indicates that the biomarker has no surrogate value for the true endpoint, since the intervention effect on the clinical endpoint is independent of the intervention effect on the biomarker.

To use a surrogate endpoint as a replacement for a true endpoint, the surrogate endpoint first has to be validated, so that conditions a) in figure 1.1 can be fulfilled. For example, in cardiovascular disease, blood pressure is a valid surrogate endpoint for stroke. Table 1.1 shows more examples on valid surrogate endpoints, frequently used in clinical trials, along with associated

true endpoints. An important advantage of a validated surrogate endpoint is that to predict the clinical efficacy of a drug more rapidly. This is the basis for accelerated approval of a drug (therapy), Gellad and Kesselheim (2003).

Table 1.1: Examples of surrogate endpoints, along with associated diseases and true endpoints.

Disease	Surrogate Endpoint	True Endpoint
HIV infection	Viral load	AIDS (or death)
Cancer	Tumor size	Mortality
Osteoporosis	Bone density	Fractures
Prostate cancer	Disease progression	PSA level
Diabetes mellitus	HbA1c in blood	Overall survival (OS)

In the next section, we will briefly review methods and techniques that are currently used for the validation of surrogate biomarkers.

1.6 Validation of Surrogate Endpoints in Clinical Trials

To validate a surrogate endpoint, it has to be shown that the effect of an intervention on a true endpoint can be explained fairly well by the effect of the intervention on a surrogate. Proving the existence of such a relation between the surrogate and the true endpoint is of interest for clinical researchers.

Prentice (1989) proposed a definition for a surrogate endpoint along with the operational criteria for the validation of surrogate endpoints, known as Prentice's definition and criteria. He then motivated using of Prentice's criteria for evaluation of surrogate endpoints by examples from cancer, ophthalmologic and cardiovascular clinical studies.

Following Prentice, numerous statistical methodologies for the validation of surrogate endpoints have been developed, that were influenced by his work. Freedman et al. (1992) studied scenarios where surrogate and true endpoints are both binary random variables. He further suggested an estimator for the proportion of treatment effect explained by a surrogate (PTE). Fleming (1992) discussed benefits and problmes that stem from using surrogate endpoints. Validation of binary and normal endpoints in a single and

multiple trials were explored extensively by Buyse and Molenberghs (1998), Burzykowski et al. (2005), Buyse et al. (2016). Buyse and Molenberghs (1998) also introduced two quantitative measures. First, the relative effect (RE) that stands for the effect of an intervention on a true endpoint relative to the effect of the intervention on a surrogate endpoint. Second, the adjusted association that represents the association between the surrogate and true endpoints after accounting for the intervention effect. Further quantitative measures to measure the strength of surrogacy for a candidate surrogate are likelihood reduction factor (LRF) and proportion of information gain (PIG). The likelihood reduction factor was suggested by Alonso et al. (2004) and is based on Kent (1983) idea about the generalized correlation, and Qu and Case (2007) introduced the proportion of information gain base on the Kullback-Leibler information gain.

Most existing approaches for the validation of surrogate endpoints, in abovementioned studies, are based on parametric models, called parametric approaches. The model we choose to validate a surrogate endpoint depends highly on the type of variables we deal with i.e. if the surrogate and true endpoints are discrete or continuous random variables. Further, if these variables are truncated or not. Table 1.2 shows different type of endpoints (either clinical or surrogate) that are used in clinical studies.

Table 1.2: Examples of different types of endpoints in clinical studies.

Variable Type	Example
Continuous measurements	Blood pressures, and tumor size.
Time to event	Time to recurrence of cancer, and time to death.
Counts	Number of skin lesions, and number of uses of rescue inhaler for asthma.
Binary or dichotomous endpoints	Failure/ success, and cured/ not cured.
Ordered categories (Ordinal)	Pain levels: absent, mild, severe, and Cholesterol levels: low, normal, high.
Unordered categories	Categories of adverse experiences. For example cardiac adverse events (including hypertension, heart failure, left ventricular systolic dysfunction, and QT prolongation.)

Failing to choose a correct model for surrogate and true endpoints or to fulfill the assumptions underlying the chosen model, could cause poor estimates and consequently incorrect inferences about the surrogacy of biomarkers. This is known as a drawback of using parametric models to evaluate surrogate endpoint.

Miao et al. (2012) suggested a non-parametric approach to validate surrogate endpoints that exploits the Kullback-Leibler divergence and permutation test. In Miao's approach, surrogate and true endpoints could come from any distribution. Therefore, Miao's approach very well affords to overcome limitations of using the parametric models to validate surrogate endpoint.

The aim in most of the methods and studies we mentioned earlier are to find out whether the surrogate is a valid replacement for the clinical endpoint. However, there has been a little or no discussion on methods that search for surrogacy interval/region (the region where the surrogate is valid on it.). The method to find such a region will be introduced in 5.

1.7 Structure of the thesis

The current thesis develops a novel approach using the notion of equivalence test for the validation of binary surrogate endpoints and to construct a surrogacy region, in a single trial setting. However, it should be stressed that our methodology is generic in the sense that it can be applied to other types of endpoints rather than binary endpoints.

This thesis consists of nine chapters and is aimed at researchers from different disciplines with different backgrounds. Therefore, Chapter 2 is mainly designed to introduce a reader to some useful statistical concepts. Chapter 3 lays out the theoretical basis for the validation of surrogate endpoints in a single trial. In Chapter 4, we will briefly talk about asthma as a common respiratory disease. Then, we will introduce three asthma trials along with the true and proposed surrogate endpoints. Finally we will apply the validation techniques we introduced in Chapter 3 to validate the surrogate endpoint in asthma trials. Chapter 5 introduces the concept of equivalence test. The application of equivalence test in surrogacy evaluation and construction of a surrogacy region is completely new, and will be covered in Chapter 5. Chapter 6 is devoted to further discussion and some possible future works. Chapters 7, 8 and 9 contains the related scripts in R that generate tables, plots and figures in the earlier chapters.

CHAPTER 2

Definitions, Theorems, Notations

2.1 Introduction

Chapter 2 provides essential statistical notions that will be useful in understanding the subsequent chapters. Hence, a reader with a solid background knowledge in statistics may skip sections 2.2 and 2.3 and jump to section 2.4, where we will introduce necessary notations that will be used throughout this study.

2.2 Definitions

2.2.1 Bonferroni Correction

Consider performing several hypotheses tests simultaneously. The first guess is to test each hypothesis at some level of significance α separately. The level of significance α is also known as type I error or false positive rate, which is defined as the error of rejecting a null hypothesis (H_0) when H_0 is actually true. In the hypothesis testing, the interest is to keep this error as low as possible.

Assume we want to test 10 independent hypotheses, where $\alpha = 0.05$ in each test. Then, the false positive rate for 10 multiple tests is as follows:

$$\begin{aligned} &P(\text{reject } H_0 \text{ in at least one test } | H_0 \text{ is true}) \\ &\qquad = 1 - P(\text{not reject } H_0 \text{ in all tests } | H_0 \text{ is true}) \\ &\qquad\qquad = 1 - (1 - 0.05)^{10} \approx 0.4 \end{aligned} \tag{2.1}$$

Therefore, we have approximately 40% chance of rejecting the true null hypothesis by mistake, having considered 10 tests simultaneously. As the number of tests increases, the probability of rejecting the true null hypothesis gets larger, which is not desirable. To deal with this problem, there are methods that adjust for α when the number of tests n increases, so that the false positive rate in n simultaneous tests remains below the desired significance level α.

The Bonferroni correction sets the significant cut-off at α/n for each test. In the above example the significant cut-off is 0.005. As a result the false positive rate for 10 multiple independent tests can be calculated as follows:

$$\begin{aligned} P(\text{reject } H_0 \text{ in at least one test } | H_0 \text{ is true}) &= 1 - (1 - 0.005)^{10} \\ &\approx 0.048 \end{aligned} \tag{2.2}$$

Therefore, the false positive rate for 10 simultaneous hypotheses fell below the desired $\alpha = 0.05$. For more details on Bonferroni method see Bender (2001) and the references therein. For cases when the tests are not independent we can still use Bonferroni method, see Sidak (1967) and Games (1977) for more details.

2.2.2 Confusion Matrix

Consider a classification problem with two available classes $Y = \{P, N\}$ where P stands for positive and N for negative class labels. Therefore, in any given sample each observation is mapped to only one of the available classes. A classifier is a model that predicts classes for observations of a given sample.

Consider $Y_{pred} = \{P, N\}$ for the class predictions produced by the classifier. Now given any classifier and observations of a given sample, we can construct a two-way contingency table which is called two-by-two confusion matrix that show the dispositions of the observations.

Table 2.1 shows the confusion matrix for a sample of size n, where n_1 is the total negatives, n_2 is the total positives and $n = n_1 + n_2$. True positive (TP) is an observation truly positive and correctly classified as positive, and false negative (FN) is an observation truly positive but incorrectly classified as negative. If an observation is truly negative and it is correctly classified as negative, it is called true negative (TN), and if it is truly negative but is incorrectly classified as positive, it is named false positive (FP).

Table 2.1: Confusion matrix

Classifier	True Classes, Y	
Predicted Class, Y_{pred}	P (Positive)	N (Negative)
P	True Positives (TP)	False Positives (FP)
N	False Negatives (FN)	True Negatives (TN)
Total sample size, n	n_1	n_2

The following measurements can be derived from the confusion matrix:

$$TPR = \frac{TP}{n_1} \tag{2.3}$$

$$FPR = \frac{FP}{n_2} \tag{2.4}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.5}$$

$$Accuracy = \frac{TP + TN}{n_1 + n_2} \tag{2.6}$$

where TPR and FPR stand for true positive and false positive rates respectively which are sample analogous to power of the test $1-\beta$ and type I error α, in statistical hypothesis testing. Alternative names to true positive rate (TPR) is sample sensitivity, recall and statistical power. Another measurement that can be derived from confusion matrix is specificity which is $1-FPR$.

$$TNR = Specificity = \frac{TN}{n_2} \tag{2.7}$$

2.2.3 Distance (Dissimilarity) Functions

In science and mathematics, distance or dissimilarity is defined as the numerical or quantitative degree of how far away two or more objects are from each other. The opposite to distance function is called similarity function that quantifies how close two elements are. In physics and mathematics two objects may represent two points and the distance between these two points is the physical length of the line that connected the points together. For more details distance and similarity functions see e.g. Deza and Deza (2016) and McCune et al. (2002).

Distance functions and their applications are widely used in different disciplines. These functions have been developed in different fields such as chemistry, physics, mathematics, computer science and statistics due to their needs. However we could divide them in two major categories as follows.

A distance function $d : X \times X \to [0, \infty)$ that satisfies three conditions for any $x, y, z \in X$:

- $d(x, y) > 0$
- $d(x, y) = 0 \Leftrightarrow x = y$
- $d(x, y) = d(y, x)$
- $d(x, y) \leq d(x, z) + d(z, y)$

is a distance metric (measure), see e.g. Goshtasby (2012) for more details on distance measures. Distance functions that do not satisfy all metric properties are usually called divergence measures as an alternative to distance metrics.

Measuring the distance between objects are widely used in statistics, probability and information theory to quantify distance between random variables, probability or sample distributions and two or more individual sample points. Kullback and Leibler (1951) proposed Kullback-Leibler (KL) divergence to measure the difference between two probability distributions over the same variable X. Nowadays, Kullback-Leibler divergence is widely used in statistics and computer science. In the following section we will discuss about KL divergence in more details.

2.2.4 Kullback-Leibler (KL) Divergence

Consider two probability distributions P and Q for a random variable X with the same support set W. Then the Kullback-Leibler divergence of Q from P, is a measure of information lost when Q is used to approximate P, which is denoted as follows:

$$d_{KL}(P||Q) = \mathbf{E}_P \left[\log \frac{dP}{dQ} \right] \tag{2.8}$$

Typically P represents the true probability distribution of observations, data or population and Q represents an approximation or estimation of P.

$d_{KL}(P||Q)$ is always greater or equal to 0, and it is equal to 0 if and only if $P = Q$. Another alternative notation for $d_{KL}(P||Q)$ is $d_{KL}(P, Q)$ which we will use throughout this study.

In the discrete case, where P and Q are probability distributions of a discrete random variable X, KL divergence can be written as follows:

$$d_{KL}(P, Q) = \sum_{x \in W} p(x) \log \frac{p(x)}{q(x)} \tag{2.9}$$

where $p(.)$ and $q(.)$ represent probability mass fictions. And when X is a continuous random variable, KL divergence is represented as follows:

$$d_{KL}(P, Q) = \int_W p(x) \log \frac{p(x)}{q(x)} \, dx \tag{2.10}$$

where $p(.)$ and $q(.)$ are probability density functions.

As mentioned earlier, not all functions that quantify distance are metrics. The Kullback-Leibler divergence measures the distance between two distributions, however it is not distance metric. The reason is that KL divergence

is not symmetric $d_{KL}(P,Q) \neq d_{KL}(Q,P)$. Moreover, it does not satisfies the triangular inequality in distance metric definition. For more details on KL divergence see e.g. Kullback and Leibler (1951), MacKay (2005) and Bishop (2006).

Example 1. Consider two normal distributions with probability distribution functions $f_X(x;\mu_1,1)$ and $f_X(x;\mu_2,1)$, with different means and the same variance. To calculate the Kullback-Leibler divergence, we can proceed by calculating the log-ratio as follows:

$$\begin{aligned}\log\left(\frac{f_X(x;\mu_1)}{f_X(x;\mu_2)}\right) &= \log f_X(x;\mu_1) - \log f_X(x;\mu_2) \\ &= (-\log\sqrt{2\pi} - \frac{x^2}{2} - \frac{\mu_1^2}{2} + x\mu_1) - (-\log\sqrt{2\pi} - \frac{x^2}{2} - \frac{\mu_2^2}{2} + x\mu_2) \\ &= \frac{\mu_2^2 - \mu_1^2}{2} + x(\mu_1 - \mu_2)\end{aligned} \tag{2.11}$$

Then, the KL divergence is the expectation of the log-ratio in equation (2.11) with respect to X, where $X \sim N(\mu_1, 1)$.

$$\begin{aligned}d_{KL}(f_{\mu_1}, f_{\mu_2}) &= \mathbf{E}_{f_{\mu_1}}\left(\frac{\mu_2^2 - \mu_1^2}{2} + X(\mu_1 - \mu_2)\right) \\ &= -\frac{(\mu_1^2 - \mu_2^2)}{2} + \mu_1(\mu_1 - \mu_2) \\ &= \frac{(\mu_1 - \mu_2)^2}{2}\end{aligned} \tag{2.12}$$

Example 2. Let X as a Bernoulli random variable. Assume two different probability mass functions $f_X(x;p)$ and $f_X(x;q)$ for, X with different means p and q. Then the log-ratio can be calculated as follows:

$$\log\left(\frac{f_X(x;p)}{f_X(x;q)}\right) = x\log p + (1-x)\log(1-p) - x\log q - (1-x)\log(1-q) \tag{2.13}$$

Eventually, the KL divergence can be calculated as the expected value of the

log-ratio with respect to X, where $X \sim \mathsf{Bern}(p)$

$$\begin{aligned} d_{KL}(f_p, f_q) &= \mathbf{E}_{f_p}\left[\log\left(\frac{f_X(x;p)}{f_X(x;q)}\right)\right] \\ &= p\log\frac{p}{q} + (1-p)\log\left(\frac{1-p}{1-q}\right) \end{aligned} \tag{2.14}$$

2.2.5 Likelihood Ratio Test

Likelihood ratio test is an approach to compare two different models. Commonly the comparison is between the unrestricted model and a simpler one. The unrestricted model is usually called a saturated model and the simpler one is called the reduced model.

Let $f_X(x;\theta)$ be the probability density or mass function of X, where θ represents one or more unknown parameters. Further, assume $\boldsymbol{x} = (x_1, ..., x_n)$ be a random sample from $f_X(x, \theta)$. We also denote the whole parameter space as $\Theta = \Theta_0 \cup \Theta_1$ such that $\Theta_0 \cap \Theta_1 = \emptyset$, which is the set of all possible values for θ. Now consider the following hypothesis:

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_1, \tag{2.15}$$

at significance level α.
Then, we can construct the likelihood ratio test based on the sample $\boldsymbol{x} = (x_1, ..., x_n)$ as follows:

- Calculate the supremum of the likelihood function $L(\theta; \boldsymbol{x}) = f(\boldsymbol{x}; \theta)$, with respect to θ, where $\theta \in \Theta_0$. Then, we denote it as $L(\hat{\theta}_{H_0})$, such that $\hat{\theta}_{H_0} = \arg\max_{\theta\in\Theta_0} L(\theta; \boldsymbol{x})$.

- Calculate the supremum of the likelihood function $L(\theta; \boldsymbol{x})$, with respect to θ, where $\theta \in \Theta$. We denote it as $L(\hat{\theta}_{H_0\cup H_1})$, such that $\hat{\theta}_{H_0\cup H_1} = \arg\max_{\theta\in\Theta_0\cup\Theta_1} L(\theta; \boldsymbol{x})$.

- Then the likelihood ratio is obtained as follows:

$$\Lambda(\underline{x}) = \frac{\sup\{L(\theta; \boldsymbol{x}), \theta \in \Theta_0\}}{\sup\{L(\theta; \boldsymbol{x}), \theta \in \Theta\}} = \frac{L(\hat{\theta}_{H_0})}{L(\hat{\theta}_{H_0\cup H_1})} \tag{2.16}$$

The less likely the null hypothesis is, the smaller $\Lambda(\boldsymbol{x})$ will be. Thefore, to test H_0 against H_1, the critical region for the likelihood ratio test is as follows:

$$C = \{\boldsymbol{x} : \Lambda(\boldsymbol{x}) \leq k_1\} \tag{2.17}$$

where k_1 is always between 0 and 1.

Consider the following transformation on Λ

$$G^2(\boldsymbol{x}) = -2 \log \Lambda(\boldsymbol{x}) \tag{2.18}$$

where $G^2(\boldsymbol{x})$ is called the likelihood ratio statistic and is an increasing function in $\boldsymbol{x}$. Then the critical region can be rewritten as

$$C = \{\boldsymbol{x} : G^2(\boldsymbol{x}) \geq k_2\} \tag{2.19}$$

For a large sample size n, it can be proved that the likelihood ratio statistics is asymptotically Chi-Square with k degrees of freedom:

$$G^2(\boldsymbol{X}) \xrightarrow{D} \chi^2_k \quad as \quad n \longrightarrow \infty, \tag{2.20}$$

where k is the difference between the number of parameters in two different models under $H_0 \cup H_1$ and H_0 hypothesis.

Therefore, the critical region based on the asymptotic distribution of $G^2(\boldsymbol{X})$ is

$$C = \{\boldsymbol{x} : G^2(\boldsymbol{x}) \geq \chi^2_k(1-\alpha)\} \tag{2.21}$$

Example 3. Let X_i for $i = 1, ..., n$ be a random sample from a Poisson distribution with a parameter θ. We want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ at significance level 0.05.
The p.m.f for X_i is

$$p(x; \theta) = \frac{e^{-\theta}\theta^x}{x!}, \quad x = 0, 1, ... \tag{2.22}$$

Then, the log-likelihood function is as follows:

$$l(\theta; \boldsymbol{x}) = -n\theta + \sum_{i=1}^{n} x_i \log \theta - \log \prod_{i=1}^{n} x_i! \tag{2.23}$$

Therefore, solving $\dfrac{\partial l(\theta;\boldsymbol{x})}{\partial\theta}=0$ for θ will result in $\hat{\theta}=\bar{x}$, and the likelihood ratio statistics is

$$G^2(\boldsymbol{x})=2n\left[\theta_0-\bar{x}+\bar{x}\log\left(\frac{\bar{x}}{\theta_0}\right)\right] \tag{2.24}$$

The distribution of $G^2(\boldsymbol{X})$ under H_0 is approximately Chi-square with one degree of freedom, and $\chi^2_1(0.95)=3.84$. Therefore, the critical region of the test can be constructed as follows:

$$C=\left\{\boldsymbol{x}:2n\left[\theta_0-\bar{x}+\bar{x}\log\left(\frac{\bar{x}}{\theta_0}\right)\right]\geq 3.84\right\} \tag{2.25}$$

For more details on likelihood ratio test see Casella and Berger (2002).

2.2.6 Logistic Regression Models for Binary Responses

In problems where the response variable takes one of only two values that representing the presence or absence of an attribute of interest, logistic regression models are suitable tools to deal with the situation. In such problems, the response or dependent variable is a Bernoulli or Binomial random variable, and the predictors or explanatory variables could be either discrete or continuous random variables or a mixture of both.

Consider we observe independent binary responses, and we aim to draw inferences about the probability of an event in the population. Suppose that, the probability of an event occurs for each individual, in the population, is equal to p_i. Let n denotes the number of observations in the sample, and y_1,...y_n as realizations of independent random variables Y_1,...Y_n, that take the values 0 and 1, such that $Y_i=1$ indicates that an event occurs for the ith subject, otherwise, $Y_i=0$. Then, we have $\mathbf{E}(Y_i)=p_i$, and the joint probability (likelihood function) of the data can be written as follows:

$$L(p)=\prod_{i=1}^{n}p_i^{y_i}(1-p_i)^{1-y_i}=p_i^{\sum_{i=1}^{n}y_i}(1-p_i)^{n-\sum_{i=1}^{n}y_i} \tag{2.26}$$

applying the log transform on the likelihood function (2.26) we get

$$l(p)=log(L)=\sum_{i=1}^{n}y_i log\left(\frac{p_i}{1-p_i}\right)+n\sum_{i=1}^{n}log(1-p_i) \tag{2.27}$$

that is called the log-likelihood function.

Now consider the vector of predictors as $\boldsymbol{x_i} = (x_1, ..., x_p)$ for $i = 1, .., n$. We would like to have the probabilities p_i depend on a vector of observed predictors $\boldsymbol{x_i}$. The logistic regression model equates the logit transform of p_i as the linear function of predictors as follows:

$$logit(p_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = \boldsymbol{x_i'\beta} \tag{2.28}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients. Therefore,

$$\mathbf{E}(Y_i) = \frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}} \tag{2.29}$$

Then, the log-likelihood for n observations can be rewritten as follows:

$$l(\beta) = \sum_{i=1}^{n} y_i \boldsymbol{x_i'\beta} - n \sum_{i=1}^{n} log(1 + e^{\boldsymbol{x_i'\beta}}) \tag{2.30}$$

The likelihood equations result from setting $d\, l(\boldsymbol{\beta})/d\boldsymbol{\beta} = 0$. Since

$$\frac{d\, l(\boldsymbol{\beta})}{d\beta_j} = \sum_{i=1}^{n} y_i x_{ij} - n \sum_{i=1}^{n} x_{ij} \frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}} \tag{2.31}$$

the likelihood equations are

$$\sum_{i=1}^{n} y_i x_{ij} - n \sum_{i=1}^{n} \hat{p}_i x_{ij} = 0 \quad j = 0, 1, ..., p \tag{2.32}$$

where $\hat{p_i} = \dfrac{e^{\boldsymbol{x_i'\hat{\beta}}}}{1 + e^{\boldsymbol{x_i'\hat{\beta}}}}$.

Setting the equations in Eq. (2.32) to zero results in a system of $p + 1$ nonlinear equations each with $p + 1$ unknown variables. The solution to the system is a vector with elements, $\hat{\beta}_j$ for $j = 0, 1, ..., p$, for which the observed data have the highest probability of occurrence, and are called the maximum likelihood estimates (MLEs) for β_j for $j = 0, 1, ..., p$. This nonlinear system of equations cannot be solved analytically. It is common to use a numerical algorithm, such as Newton-Raphson algorithm to obtain the MLEs. More details on the estimation of β_j, for $j = 0, 1, ..., p$ and their variances can be found in Agresti (2013, Chapter 5).

Finally, we can plug in the estimated values of β_j for $j = 0, 1, ..., p$ in $\hat{p_i} = \frac{e^{\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}}}{1 + e^{\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}}}$, and gets $\hat{p_i}$, for $i = 1, ..., n$.

To test the hypothesis of the form $H_0 : \beta_j = 0 \quad vs. \quad H_1 : \beta_j \neq 0$ for a single logistic regression coefficient, we can use the Wald statistic as follows:

$$z = \frac{\hat{\beta}_j - \beta_{j|H_0}}{\hat{var}(\hat{\beta}_j)} \tag{2.33}$$

This statistic has approximately a standard normal distribution in large samples.

Eventually, using the Wald statistic in Eq. (2.33), the $100(1 - \alpha)\%$ confidence interval for β_j can be calculated as follows:

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\hat{var}(\hat{\beta}_j)} \tag{2.34}$$

where $z_{1-\alpha/2}$ is the normal critical value for a two-sided test of size α, and $\hat{var}(\hat{\beta}_j)$ is the estimated variance of $\hat{\beta}_j$.

2.2.7 Permutation Test

A permutation test has been widely used for a long time to construct the correct distribution of a test statistic under a null hypothesis. It is a non-parametric method that generates the sampling distribution of a test statistic by re-sampling the observed data without replacement. If the data are permuted over all possible arrangements of the data, it is called an exact permutation test. However, it is called approximate permutation test if the data are permuted over only a subset of all possible arrangements of the data. See Berry et al. (2011) for more details.

When it is not easy to compute the distribution of the test statistics, the permutation test can be helpful. Results of permutation tests are valid however, they are computationally intensive compare to the standard parametric (analytic) methods. In the following, the permutation test to compare the mean in two different groups is explained in details. Let n be the total sample size.

1. Compute the mean for each group and calculate the difference in means, dif_{obs}.

2. Permute (shuffle) randomly the n observations between the two groups.

3. For the permuted sample, calculate the mean in each group and the difference in means, dif_{perm}

4. Repeat steps 2 and 3 for L times.

5. Then the $p - value$ permutation is

$$p - value_{perm} = \frac{\text{number of} |dif_{perm}| \geq |dif_{obs}|}{L} \tag{2.35}$$

Note that in practice, scientists normally use approximate permutation test, since generating all possible permuted samples are computationally intensive and costly. The procedure we described in steps 1 to 5 above is an approximate permutation test.

Example 4. Let $n = 500$ be the total sample size. we consider two groups, 0 and 1. Then, we generate observations from two different scenarios. In each group:

1. First scenario: Generate observations from the normal distribution with the same mean and the same variance equal to 1.

```
group<-rep(c(0,1), c(200,300))
S1<-rnorm(500)
```

2. Second scenario: Generate observations from the normal distributions with different means and the same variance equal to 1.

```
S2<-rnorm(500, mean=group/2)
```

Our aim is to compare the means of two different groups $(0, 1)$, in these two scenarios, using the $t-$test and the permutation test. Table 2.2 and figure 2.1 show results of performing $t-$test and permutation test on S1 and S2, where OMG0 and OMG1 are the observed mean in group 0 and 1, respectively. The number of permuted samples in each scenario is $L = 1000$.

Table 2.2: Results of $t-$test and permutation test to compare group means in two different scenarios.

Data	OMG0	OMG1	$t-$value	df	$p-$value$_{t-test}$	$p-$value$_{perm}$
S1	-0.043	0.046	-0.960	498	0.338	0.351
S2	0.086	0.490	-3.312	498	0.001	0.00001

Figure 2.1: Histograms of difference in means generated from $L = 1000$ permutations of $S1$ and $S2$.

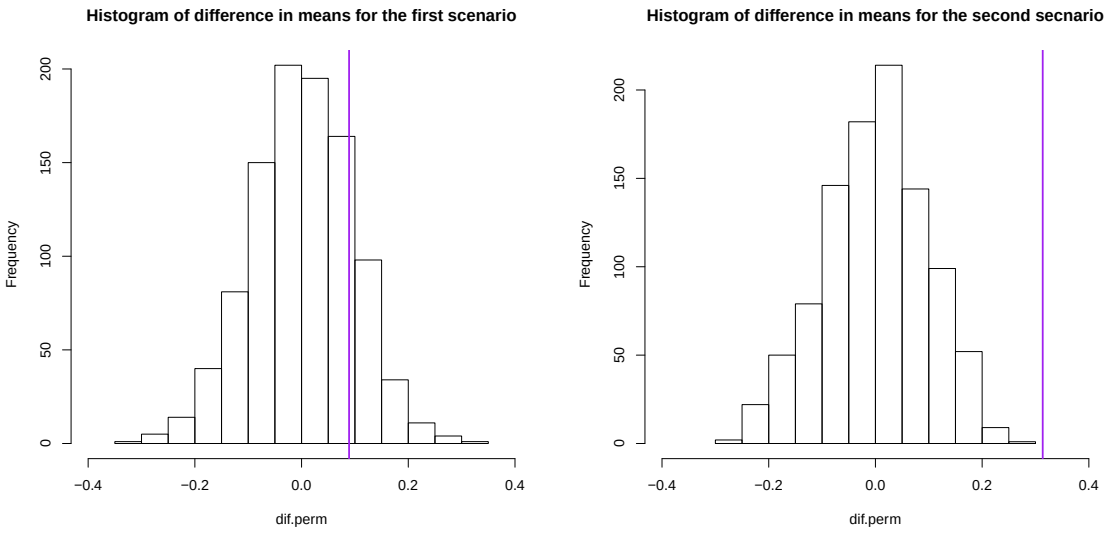


Figure 2.1 shows the approximate (estimate) distributions of difference in means generated from $L = 1000$ number of permutations, for the first and second scenarios. The vertical line represent the observed difference in means for each scenario. According to Figure 2.1, means are not different in the first scenario however, they are different in the second scenario. We can also see in table 2.2 that, the results from $t-$test and permutation test are consistent in both scenarios. Full scripts that generate figure 2.1 and table 2.2 can be found in Ch. 7.

2.2.8 The Support of a Random Variable

The support of a continuous random variable X is defined as the set of all values that probability density function (pdf) is strictly positive, symbolically $supp(X) = \{x \in \mathbb{R} : f_X(x) > 0\}$.

Example 5. Assume X is a continuous uniform random variable with the following density function.

$$f_X(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{o.w.} \end{cases} \tag{2.36}$$

Then, the support of X is denoted as $supp(X) = [0, 1]$.

In a discrete case, the support of a random variable X is defined as $supp(X) = \{x \in \mathbb{R} : p_X(x) > 0\}$, where $p_X(.)$ is the probability mass function (pmf) of the random variable X.

Example 6. Assume X is a discrete uniform random variable with the following probability mass function.

$$p_X(x) = \begin{cases} 1/2 & x = 0 \\ 1/2 & x = 1 \\ 0 & \text{o.w.} \end{cases} \tag{2.37}$$

Then, the support of X is $supp(X) = \{0, 1\}$. More details on the support of a random variable can be found in Folland (1999).

2.3 Theorems

2.3.1 Fieller's Theorem

Fieller's theorem Fieller (1940) is used to construct confidence sets (limits) for a ratio of normal means, see for example Casella and Berger (2002, Chapter 9).

Consider $(X_1, Y_1), ..., (X_n, Y_n)$ as a random sample from a bivariate normal distribution as follows:

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \right] \quad \text{for} \quad i = 1, ..., n \tag{2.38}$$

Then, a confidence set for $\theta = \dfrac{\mu_X}{\mu_Y}$ can be constructed as follows:

- Let $Z_{\theta i} = Y_i - \theta X_i$ for $i = 1, ..., n$ and therefore, $\bar{Z}_\theta = \bar{Y} - \theta \bar{X}$. Then, it can be shown that $\bar{Z}_\theta$ is a normally distributed random variable with mean 0 and variance

$$V_\theta = \frac{\sigma_Y^2 - 2\theta\rho\sigma_Y\sigma_X + \theta^2\sigma_X^2}{n} \tag{2.39}$$

  where $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$

- Then, $\hat{V}_\theta$ is the estimate for V_θ that can be calculated as

$$\begin{aligned} \hat{V}_\theta &= \frac{\sum_{i=1}^n (Z_{\theta i} - \bar{Z}_\theta)^2}{n(n-1)} \\ &= \frac{S_Y^2 - 2\theta S_X S_Y + \theta^2 S_X^2}{n-1}. \end{aligned} \tag{2.40}$$

  where

$$\begin{cases} S_Y^2 = \dfrac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n} & \text{for} \quad i = 1, ..., n \\ S_X^2 = \dfrac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} & \text{for} \quad i = 1, ..., n \\ S_{XY} = \dfrac{\sum_{i=1} n(Y_i - \bar{Y})(X_i - \bar{X})}{n} & \text{for} \quad i = 1, ..., n \end{cases} \tag{2.41}$$

- Finally, it can be shown that $\mathbf{E}(\hat{V}_\theta) = V_\theta$, $\hat{V}_\theta$ and $\bar{Z}_\theta$ are independent of each other, and $(n-1)\hat{V}_\theta/V_\theta \sim \chi^2_{n-1}$. Therefore, $\bar{Z}_\theta/\sqrt{\hat{V}_\theta} \sim t_{n-1}$. Then, the confidence set for θ can be defined as follows:

$$\left\{\theta : \frac{\bar{z}_\theta^2}{\hat{v}_\theta} \leq t^2_{n-1,\alpha/2}\right\} \tag{2.42}$$

that can be rewritten as:

$$\left\{\theta : \left(\bar{x}^2 - \frac{t^2_{n-1,\alpha/2}}{n-1}S_x^2\right)\theta^2 - 2\theta\left(\bar{x}\bar{y} - \frac{t^2_{n-1,\alpha/2}}{n-1}S_{xy}\right) + \left(\bar{y}^2 - \frac{t^2_{n-1,\alpha/2}}{n-1}S_y^2\right) \leq 0\right\} \tag{2.43}$$

This set characterizes a parabola in θ, and solving this quadratic equation for θ will result in the endpoints of the confidence set.

2.4 Notations

In this thesis, we will use Z, T and S as abbreviations for treatment, true and surrogate endpoints respectively, where Z, T and S are random variables. Surrogate and true endpoints are random variables since it is not possible to exactly predict their values for each patients. We also think of a treatment Z as a random variable since we are dealing with a randomized clinical trial. In a randomized clinical trial, subjects participating in the trial are randomly assigned to the new treatment or to the group receiving the standard treatment as a control group.

We let $f_S(s)$ for $s \in A$ denotes the probability distribution of the random variable S. Further, $f_T(t)$ represents the probability distribution of the random variable T, for $t \in B$, and $f_Z(z)$ indicates the probability distribution of the random variable Z, for $z \in C$. Note that A, B and C are the supports of random variables S, T and Z respectively.

In the same way, $f_{S|Z}(s|z)$ indicates the probability distribution of S conditional on $Z = z$, for $s \in A$ and any given $z \in C$, $f_{T|Z}(t|z)$ indicates the probability distribution of T conditional on values of Z, for $t \in B$ and any given $z \in C$, $f_{T|S}(t|s)$ indicates the probability distribution of T conditional on $S = s$, for $t \in B$ and any given $s \in A$, and $f_{T|S,Z}(t|s, z)$ is the probability distribution of T conditional on values of S and Z, for $t \in B$ and any given $s \in A$ and $z \in C$.

Moreover, $f_{T,S,Z}(t, s, z)$ indicates the joint probability distribution of a triplet (T, S, Z), for $T \in B$, $S \in A$ and $Z \in C$. Throughout this thesis, we will follow the same rule to derive further notations corresponding T, S and Z.

In a given sample of size n, $S_1, ..., S_n$ represent surrogate endpoints as random variables, that are independent and identically distributed. Moreover, $s_1, ..., s_n$ indicate observed values of the surrogate endpoints $S_1, ..., S_n$. In the same fashion, $T_1, ..., T_n$ and $Z_1, ..., Z_n$ stand for true endpoints and treatments as independent and identically distributed random variables respectively. Further, $t_1, ..., t_n$ and $z_1, ..., z_n$ correspond to observed values of $T_1, ..., T_n$ and $Z_1, ..., Z_n$. Note that, the same notations, as we introduced earlier for marginal, joint and conditional probability distributions of S, T and Z, are applied to S_i, T_i, Z_i, for $i = 1, ..., n$.

CHAPTER 3

Evaluation of Surrogate Endpoints

3.1 Introduction

Candidate surrogate endpoints are generally suggested based on biological considerations however, their validation depends on statistical methods. In the past two decades, there is fast-growing literature on statistical approaches for the evaluation of surrogate markers. These approaches fall into two major frameworks: one is developed for the evaluation of surrogate markers using data from a single large clinical (a single trial setting) and the other is based on meta-analysis of multiple clinical trials (a multiple trial setting). This current thesis concentrates on statistical evaluation of surrogate endpoints in a single trial setting. The statistical evaluation of surrogate markers in a single setting was effectuated by Prentice (1989).

In the current chapter, we will explain the Prentice (1989) definition of surrogacy along with four operational criteria to validate candidate surrogate endpoints. Then, we will discuss parametric and non-parametric methods in a single trial setting to verify the operational criteria.

3.2 Prentice's Definition and Operational Criteria of Surrogacy

Prentice (1989) defined a surrogate endpoint as "a response variable, for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint" Prentice (1989). Symbolically, S should satisfy:

$$f_{S|Z}(s|z) = f_S(s) \Leftrightarrow f_{T|Z}(t|z) = f_T(t), \tag{3.1}$$

Prentice's definition itself cannot be used in practice to evaluate a surrogate endpoint. To verify the Prentice's definition (3.1) directly, we deal with equivalence of two statistical tests. Large number of trials are needed in order to show that a triplet (T, S, Z) satisfies (3.1). However, we still might not be able to prove (3.1) for all the trials due to the lack of sensitivity.

Therefore, Prentice suggested that a biomarker S is regarded as a valid surrogate for a clinical endpoint T if the triplet (T, S, Z) satisfy the following conditions:

$$f_{S|Z}(s|z) \neq f_S(s), \tag{3.2}$$

$$f_{T|Z}(t|z) \neq f_T(t), \tag{3.3}$$

$$f_{T|S}(t|s) \neq f_T(t), \tag{3.4}$$

$$f_{T|S,Z}(t|s, z) = f_{T|S}(t|s), \tag{3.5}$$

The first and second criteria imply that the treatment has prognostic relevance for the surrogate and the true endpoints. The third criterion states that the surrogate endpoint has a significant impact on the true endpoint. Thus, the true endpoint cannot be independent of the surrogate. Note that, in order to consider a biomarker as a candidate for a surrogate endpoint the first three Prentice's criteria have to be verified in the first place.

Eventually, the last Prentice's criterion implies that the full effect of the treatment upon the true endpoint is captured by the surrogate. The last Prentice's criterion bears the essential concept of surrogacy, in the sense that, once we have information available about the surrogate, knowing about the treatment does not give us any new information for predicting the true endpoint, Buyse et al. (2000). This implies the importance role of the last Prentice's criterion for the validation of the surrogate endpoint.

Lemma 1. *Buyse and Molenberghs (1998) showed that Prentice's criteria* (3.4)-(3.5) *are sufficient conditions for Prentice's definition* (3.1)*, in the case of binary endpoints.*

Proof. First assume $f_{S|Z}(s|z) = f_S(s)$ is correct, then need to show $\Rightarrow$ holds in (3.1). By definition, we have

$$f_{T|Z}(t|z) = \sum_s f_{T,S|Z}(t,s|z) = \sum_s f_{T|S,Z}(t|s,z) f_{S|Z}(s|z). \tag{3.6}$$

We can now replace $f_{S|Z}(s|z)$ by $f_S(s)$. And if (3.5) holds, then (3.6) can be written as follows:

$$f_{T|Z}(t|z) = \sum_s f_{T|S}(t|s) f_S(s) = f_T(t). \tag{3.7}$$

So far we showed $\Rightarrow$ holds in (3.1).

Now consider $f_{T|Z}(t|z) = f_T(t)$ is correct, and we are going to show that $\Leftarrow$ holds in (3.1). If (3.5) holds, then we have

$$f_{T|Z}(t|z) = \sum_s f_{T|S,Z}(t|s,z) f_{S|Z}(s|z) = \sum_s f_{T|S}(t|s) f_{S|Z}(s|z). \tag{3.8}$$

On the other hand we have

$$f_T(t) = \sum_s f_{T|S}(t|s) f_S(s). \tag{3.9}$$

As we assumed $f_{T|Z}(t|z) = f_T(t)$, the left hand sides of (3.8) and (3.9) are equal. Consequently we can write the right hand sides of (3.8) and (3.9) equal, which yields to:

$$\sum_s f_{T|S}(t|s)[f_{S|Z}(s|z) - f_S(s)] = 0. \tag{3.10}$$

For a binary surrogate endpoint, (3.10) can be written as follows:

$$[f_{S|Z}(S=1|z) - f_S(S=1)][f_{T|S}(t|S=1) - f_{T|S}(t|S=0)] = 0. \tag{3.11}$$

If $f_{T|S}(t|S=1) - f_{T|S}(t|S=0) \neq 0$, which is equivalent to $f_{T|S}(t|s) \neq f_T(t)$ in (3.4) holds, we could conclude that $f_{S|Z}(s|z) = f_S(s)$. Thus we proved $\Leftarrow$ holds in (3.1). $\square$

As a result, to evaluate a binary surrogate endpoint, it is sufficient to show that Prentice's criteria (3.4)-(3.5) are satisfied. For endpoints other than binary, Burzykowski et al. (2005) and Buyse et al. (2000) noted that Prentice's criteria (3.2)-(3.5) are informative conditions and tend to be verified for valid surrogate endpoints, however "they should not be regarded as strict criteria." We will introduce other measures like proportion treatment explained (PE) later in this chapter, which can be used as complements to conditions (3.2)-(3.5).

In the following sections, we will present different methods that try to verify Prentice's criteria for the validation of a surrogate endpoint in a single trial.

3.3 Parametric Methods for Surrogacy Evaluation in a Single Trial

One approach to verify Prentice's criteria for the validation of a surrogate endpoint is using certain parametric models. For example, Burzykowski et al. (2005) discus the case where both endpoints are normal variables, and linear regression can be used to model relationships between S, T and Z. For example the relationship between S and Z in $f_{S|Z}(S|Z)$ can be modeled as a linear regression, where S is a response variable and Z is an explanatory variable. Then validating Equation (3.2) is equivalent to test for the significance of regression coefficient of Z in the linear regression between S and Z. However, choosing proper models to represent relationships between S, T and Z depends on the type of variables.

In general, surrogate S and true endpoint T could be any type of variables. They are not necessarily normal variables. Thus, validating the existence of relationship between variables cannot be done only by applying simple regression models. Therefore, one possibility would be to extend the relationship to generalized linear models to incorporate non-normal variables. To read more on generalized linear model see de Jong and Heller (2008).

The translation of Prentice's criteria to the generalized linear models can be explained through the following models and Table 3.1:

$$g(\mathbf{E}(S_i|Z_i = z_i)) = \mu_{S|Z} + \alpha z_i \tag{3.12}$$
$$g(\mathbf{E}(T_i|Z_i = z_i)) = \mu_{T|Z} + \beta z_i \tag{3.13}$$
$$g(\mathbf{E}(T_i|S_i = s_i)) = \mu_{T|S} + \gamma s_i \tag{3.14}$$
$$g(\mathbf{E}(T_i|Z_i = z_i, S_i = s_i)) = \mu_{T|Z,S} + \beta_S z_i + \gamma_Z s_i \tag{3.15}$$
$$g(\mathbf{E}(T_i|Z_i = z_i, S_i = s_i)) = \mu_{T|Z,S} + \beta_S z_i + \gamma_Z s_i + \delta z_i s_i \tag{3.16}$$

where g is a link function and $\mathbf{E}$ indicates the expectation. Let a triplet (T_i, S_i, Z_i) represents a random vector, and (t_i, s_i, z_i) represents its observed value, for ith observation in the sample, for $i = 1, ..., n$. If based on some prior information, we would know that there exists no interaction between S and Z, then we may use model without interaction term (3.15) instead of model (3.16).

Table 3.1: Prentice's criteria translation to generalized linear models

Prentice's criterion	Quantity	Null hypothesis [1]
$f_{S\|Z}(s\|z) \neq f_S(s)$	Effect of Z on S	$\alpha = 0$
$f_{T\|Z}(t\|z) \neq f_T(t)$	Effect of Z on T	$\beta = 0$
$f_{T\|S}(t\|s) \neq f_T(t)$	Effect of S on T	$\gamma = 0$
$f_{T\|Z,S}(t\|z, s) = f_{T\|S}(t\|s)$	Effect of Z on T, given S	$\beta_S = \delta = 0$

Prentice's criteria and surrogacy can be validated through tests of hypotheses for the parameters of the models 3.12 to 3.16. See Buyse and Molenberghs (1998) for more details on the verification of Prentices criteria using generalized linear models. Specifically, the first, second and third Prentice's criteria can be verified by showing that the hypothesis tests of $\alpha = 0$, $\beta = 0$ and $\gamma = 0$ are significant. To test for the significance of the effect of Z on T given S, we need to test for the significance of β_S and δ at the same time. To make sure that the false positive rate for multiple testing $\beta_S = \delta = 0$ is controlled at the 0.05 level, we need to use Bonferroni correction to control the false positive rates of each individual test, $\beta_S = 0$ and $\delta = 0$.

We will set the false positive rates for each test in Table 3.1 to 0.05, unless otherwise stated. The false positive rate for testing the null hypotheses in

[1] Null hypothesis for testing significance of coefficients in the general linear models.

Table 3.1 are as follows:

$$P(\text{Reject the null hypothesis}|\alpha = 0) \tag{3.17}$$
$$P(\text{Reject the null hypothesis}|\beta = 0) \tag{3.18}$$
$$P(\text{Reject the null hypothesis}|\gamma = 0) \tag{3.19}$$
$$P(\text{Reject the null hypothesis}|\beta_S = \delta = 0) \tag{3.20}$$

We test the hypotheses in the same order as in Table 3.1. If we find no evidence to reject $\alpha = 0$, we would not proceed with the other hypotheses. Under these circumstance, we must look for another potential candidate as a surrogate endpoint. Clearly Z should have a significant effect on T otherwise, the process of searching for a surrogate to replace the true endpoint does not make any sense.

Note that, in order to validate the fourth Prentice's criterion using the general linear model, we need to show that the null hypothesis of the form $\beta_S = \delta = 0$ is true. However, what we actually showed here is that the null hypothesis of the form $\beta_S = \delta = 0$ cannot be rejected. In a regular hypothesis test framework, we usually collect evidence to reject the null hypothesis, and failing to reject the null hypothesis is not equivalent to conclude that the null hypothesis is true. In here, we face the situation where we are interested to prove the null hypothesis rather than to reject it, which is much common in the (bio)-equivalence setting. Therefore, using the usual hypothesis test framework is not the correct way of verifying the fourth Prentice's criterion. We will discuss a possible solution to this problem thoroughly in Chapter 5.

3.3.1 Logistic Regression for the Validation of Prentice's Criteria in a Single Trial with Binary Endpoints

When we deal with binary variables, validation of Prentice's criteria using generalized linear models boils down to the use of Logistic regression, if the link function $\eta = logit$ is used. Logistic regression can be used as a tool to model relationship of binary variable as an outcome to explanatory features, see e.g. Agresti (2013).

In equations (3.12) to (3.16) assume T_i and S_i are binary endpoints. Then we can express the effect of Z on T through the following logistic model:

$$\ln\left(\frac{P(S_i = 1|Z_i = z_i)}{P(S_i = 0|Z_i = z_i)}\right) = \mu_{S|Z} + \alpha z_i, \tag{3.21}$$

where we are interested in the test of hypothesis for $\alpha = 0$. In order to verify the first Prentice's criterion α needs to deviate from 0.

In a similar way, we could use logistic regression to model the relation between Z and T, and S and T as follows:

$$\ln\left(\frac{P(T_i = 1|Z_i = z_i)}{P(T_i = 0|Z_i = z_i)}\right) = \mu_{T|Z} + \beta z_i, \tag{3.22}$$

$$\ln\left(\frac{P(T_i = 1|S_i = s_i)}{P(T_i = 0|S_i = s_i)}\right) = \mu_{T|S} + \gamma s_i. \tag{3.23}$$

Therefore, in order to (3.3) and (3.4) are satisfied, β and γ need to deviate from 0.

Finally, the effect of Z and S on T in (3.5) could be represented through the following logistic model.

$$\ln\left(\frac{P(T_i = 1|Z_i = z_i, S_i = s_i)}{P(T_i = 0|Z_i = z_i, S_i = s_i)}\right) = \mu_{T|Z,S} + \beta_S Z_i + \gamma_z s_i + \delta z_i s_i, \tag{3.24}$$

where β_S is the effect of Z on T adjusted for S, γ_Z is the effect of S on T adjusted for Z, and δ is the interaction effect of S and Z on T. The last Prentice's criterion in equation (3.5) implies that the full effect of Z on T is captured by S. This is equivalent to show that $\beta_S = \delta = 0$ in (3.24).

It was suggested by Freedman et al. (1992) to do the hypotheses test of $\beta_S = 0$ and $\delta = 0$ sequentially. Thus, to verify (3.5) we first need to test for δ; if not significant, model (3.24) reduces to:

$$\ln\left(\frac{P(T_i = 1|Z_i = z_i, S_i = s_i)}{P(T_i = 0|Z_i = z_i, S_i = s_i)}\right) = \mu_{T|Z,S} + \beta_S z_i + \gamma_Z s_i, \tag{3.25}$$

thereafter we need to test model (3.25) versus model (3.23) .

3.3.2 Odds Ratio for Surrogacy Evaluation in a Single Trial with Binary Endpoints

The odds ratio (OR) is a measure of association between an exposure and a binary outcome. It is most commonly used in case control studies which allows to compare the intervention group of a study relative to the comparator drug or placebo group. An odds ratio indicates the odds of an outcome occurring given a particular exposure, compared to the odds of the outcome

occurring in the absence of that exposure. In logistic regression, the exponential function of the coefficient of the model is the odds ratio associated with a one unit increase in the exposure (explanatory variable). See e.g. Agresti (2013) for more details on odds ratios.

Therefore, testing for the significance of coefficients α, β and γ in the logistic models (3.21), (3.22) and (3.23) is equivalent to test for the significance of odds ratios $OR_{SZ} \neq 1$, $OR_{TZ} \neq 1$ and $OR_{ST} \neq 1$, in the marginal contingency tables for (S, Z), (T, Z) and (T, S) . Finally, showing that $\delta = 0$ and $\beta_S = 0$ in the models (3.24) and (3.25) is equivalent to show that, $OR_{TZ|S=0} = OR_{TZ|S=1} = 1$ in the three way contingency table of (T, S, Z).

So far, we reviewed the parametric models to verify Prentice's criteria. However there are limitations to the parametric model based approach, e.g.

- The results highly depend on choosing the correct parametric models.
- Verification of the fourth Prentice's criterion by testing for the parameters of parametric models is not a direct test of equality of the two conditional distributions.
- To validate the fourth Prentice's criterion showing that the null hypothesis is true, using the evidence we collected to reject the null hypothesis, raises a conceptual problem. Hence, the last criterion is useful to reject a poor surrogate endpoint however, it is inadequate to validate a good surrogate endpoint. Moreover, the non-significance of the test does not mean that the full effect of Z on T can be explained S.

3.3.3 Freedman's Proportion Treatment Explained, PE

The fact that the hypothesis tests for β_s and δ, in the logistic model (3.24), are not statistically significant indicates that, some treatment effect Z on the true endpoint T is explained by the surrogate S. However it cannot be proven that the full effect of Z on T is captured by S. Therefore, Freedman et al. (1992) proposed a quantitative measure called the proportion of treatment explained (PE) to measure the proportion of a treatment effect on a clinical endpoint that can be explained by a surrogate. PE can be used as a complement to the fourth Prentice's criterion, when Prentice's fourth criterion is fulfilled.

The proportion of treatment explained can be mathematically expressed as follows:

$$PE = 1 - \frac{\beta_S}{\beta}, \tag{3.26}$$

where β and β_S are the parameters in the models (3.13) and (3.15). If there is any evidence on the existence of an interaction effect between S and Z, one might need to replace model (3.15) by model (3.16). In both cases, if a surrogate completely explains treatment effect on true endpoint, $\beta_S = 0$ and as a consequence $PE = 1$. If $PE < 1$, a surrogate explains only part of the treatment effect on the true endpoin, thus the surrogate is called an incomplete surogate.

Since PE is the ratio of two parameters, the confidence limits for PE can be constructed using the Fieller's theorem, as explained in Section 2.3.1:

$$\text{Lower limit} = \frac{f_1 - \sqrt{D}}{f_2}, \quad \text{Upper limit} = \frac{f_1 + \sqrt{D}}{f_2}, \tag{3.27}$$

where

$$f_0 = \hat{\beta_S}^2 - z^2_{1-\alpha/2} \hat{var}(\hat{\beta_S}) \tag{3.28}$$

$$f_1 = \hat{\beta_S}\hat{\beta} - z^2_{1-\alpha/2} \hat{cov}(\hat{\beta_S}, \hat{\beta}) \tag{3.29}$$

$$f_2 = \hat{\beta}^2 - z^2_{1-\alpha/2} \hat{var}(\hat{\beta}) \tag{3.30}$$

$$D = f_1^2 - f_0 f_2 \tag{3.31}$$

and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the normal distribution. Note that, if the number of observations n were not large enough $z_{1-\alpha/2}$ should be replaced by Student's t-distribution with $n-1$ degrees of freedom, $t_{1-\alpha/2}(n-1)$.

According to Freedman et al. (1992), a large proportion of the treatment effect on the true endpoint is explained by the surrogate if the lower confidence limit of PE is larger than some proportion, say 0.5 or 0.75. However, there are some limitations to the proportion treatment explained:

- Defining PE as a proportion would cause some conceptual problems where β_S and β take different signs, and as a consequence PE could take values greater than one. Thus, PE is not restricted to the unit interval which makes interpretation of PE difficult. See e.g. Choi et al. (1993) and Volberding et al. (1990) for more details.

- PE confidence limit could be wide, see e.g. Lin et al. (1997). Taking large sample sizes and having strong effect of Z on T may overcome this issue. However note that usually having strong effect of Z on T is actually a rare situation to happen in real problems.

3.4 Non-Parametric Methods for Surrogacy Evaluation in a Single Trial

3.4.1 An Entropy Based Non-Parametric Method for Surrogacy Evaluation

Miao et al. (2012) proposed a new approach to validate surrogate endpoints based on the Kullback-Leibler divergence measure and the permutation test. This method directly verifies the fourth Prentice's criterion. It does not make any distributional assumptions on the endpoints, and it is robust to model misspecification. In the following we will describe the method.

Consider the following hypotheses:

$$H_0 : f_{T|S,Z}(t|s,z) = f_{T|S}(t|s) \quad vs. \quad H_1 : f_{T|S,Z}(t|s,z) \neq f_{T|S}(t|s), \quad (3.32)$$

The Kullback-Leibler divergence quantifies the difference between $f_{T|S,Z}(t|s,z)$ and $f_{T|S}(t|s)$ as follows:

$$d_{KL}(f_{T,S,Z|H_0}(t,s,z), f_{T,S,Z|H_1}(t,s,z)) = \int\limits_{(t,s,z)} \log\left(\frac{f_{T,S,Z}(t,s,z) f_S(s)}{f_{S,Z}(s,z) f_{T,S}(t,s)}\right) dF_{T,S,Z}(t,s,z), \quad (3.33)$$

where $dF_{T,S,Z}(t,s,z) = f_{T,S,Z}(t,s,z)\, d(t,s,z)$, and $f_{T,S,Z|H_0}(t,s,z)$ and $f_{T,S,Z|H_1}(t,s,z)$ are the joint probability density functions of (T,S,Z) under the null and alternative hypotheses in equation (3.32).

Therefore, testing the hypotheses in equation (3.32) is equivalent to test the following hypotheses:

$$H_0 : d_{KL} = 0 \quad vs. \quad H_1 : d_{KL} \neq 0, \quad (3.34)$$

where d_{KL} is the right hand side of the equality sign in equation (3.33).

We can estimate the KL divergence by replacing the true probability density functions in equation (3.33) with their empirical estimates from the sample.

$$\hat{d}_{KL} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\hat{f}_{T,S,Z}(t_i,s_i,z_i)\hat{f}_S(s_i)}{\hat{f}_{S,Z}(s_i,z_i)\hat{f}_{T,S}(t_i,s_i)}\right) \quad (3.35)$$

where (t_i, s_i, z_i) shows the observed clinical endpoint, surrogate endpoint and treatment for the ith patient, n is the sample size and $\hat{d}_{KL}$ is called a test

statistics. In equation (3.35), $\hat{f}_{T,S,Z}(t_i, s_i, z_i)$ represent the estimated joint density function of (T_i, S_i, Z_i) at (t_i, s_i, z_i), $\hat{f}_S(s_i)$ shows estimated marginal density of S_i at s_i, $\hat{f}_{S,Z}(s_i, z_i)$ is the estimated joint density of (S, Z) at (s_i, z_i), and $\hat{f}_{T,S}(t_i, s_i)$ is the joint density function of (T_i, S_i) at (t_i, s_i). The chance of S being a good surrogate for T increases as $\hat{d}_{KL}$ gets closer to 0.

To calculate the $p - value$ for this test, we estimate the sampling distribution of $\hat{d}_{KL}$ under the null hypothesis using the permutation test. We permute the observations (t_i, s_i, z_i), for $i = 1, ..., n$, in a way that agrees with the null hypothesis. The null hypothesis states that, given $S = s$, the probability distribution of T is the same for all values of Z. That is to say, given the information about the surrogate, whether the patient received new drug or active control, has no effect on the outcome of the clinical endpoint. Thus to generate a permuted sample of the form (t_i, s_i, z_i^*) from the original sample (t_i, s_i, z_i), for $i = 1, ..., n$, we proceed as follows:

- We fix t_i and s_i, for each patient i where $i = 1, ..., n$.
- Then in each level (value) of the surrogate S, we resample z_i, for $i = 1, ..., n$, without replacement.

For each permuted sample with observed values (t_i, s_i, z_i^*), for $i = 1, ..., n$, we can calculate $\hat{d}_{KL}^*$ as follows:

$$\hat{d}_{KL}^* = \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\hat{f}_{T,S,Z^*}(t_i, s_i, z_i^*) \hat{f}_S(s_i)}{\hat{f}_{S,Z^*}(s_i, z_i^*) \hat{f}_{T,S}(t_i, s_i)} \right) \tag{3.36}$$

Note that the possible number of different permutations and consequently permuted samples are huge. However, following Pesarin (2001), Westfall and Young (1993) and Bickel and Van Zwet (1987), we limit the number of permutations to L random permuted samples. Then, we compute $\hat{d}_{KL}^*$ for each sample and construct the approximate distribution of $\hat{d}_{KL}$ under H_0.

Then the $p - value$ for the test, based on L permuted samples, can be constructed as follows:

$$p - value_{perm} = \frac{1 + \sum_{l=1}^{L} I(\hat{d}_{KL,l}^* \geq \hat{d}_{KL})}{1 + L}, \tag{3.37}$$

where I denotes the indicator function, $\hat{d}_{KL,l}^*$ is the estimate of KL divergence using the ith permuted sample, and $\hat{d}_{KL}$ is the estimate of KL divergence

using the original sample (t_i, s_i, z_i) for $i = 1, ..., n$. Then, we reject the null hypothesis if $p - value_{perm} < \alpha$, where α is the pre-specified and fixed false positive rate for the test, $\alpha = P(d_{KL} \neq 0 | d_{KL} = 0)$. Note that adding 1 unit to the numerator and denominator guarantees that the $p - value_{perm}$ differs from 0.

In the case of binary endpoints, simply replace the probability density functions with probability mass functions, and the integral with the sum over total number of observations in equations (3.33), (3.35) and (3.36). The rest of the procedure remains the same for binary endpoints.

3.4.2 Asymptotic Distribution of KL Divergence Estimator for Categorical Endpoints, Based on the Original Data

Another approach that Miao et al. (2012) introduced to validate the fourth Prentice's criterion, where surrogate endpoints are count or binary, poisson or multinomial random variables, is to use the asymptotic distribution of $\hat{d}_{KL}$ in (3.35).

Consider a sample of n patients. For each patient i, for $i = 1, ..., n$, we collected the information on (T_i, S_i, Z_i) variables, so that (t_i, s_i, z_i) represents the observed value of (T_i, S_i, Z_i) for the ith patient. Also assume T, S and Z are discrete (categorical) random variables with $1 \leq h \leq a$, $1 \leq j \leq b$, $1 \leq k \leq 2$ categories respectively. We can now construct a three way contingency table for the variables (T, S, Z), based on the total number of n observations in the sample.

Let $p_{hjk} = P(T = t_h, S = s_j, Z = z_k)$ denote the joint probability of (T, S, Z), where (t_h, s_j, z_k) are the values of (T, S, Z) at the hth, jth and kth levels of T, S and Z respectively. Then the number of observations in the cell corresponds to the hth, jth and kth levels of T, S and Z in the contingency table is a binomial random variable, represented as follows:

$$n_{hjk} \sim \mathsf{Bin}(n, p_{hjk}) \tag{3.38}$$

Therefore, the joint probability distribution of the number of observations for all cells, in the contingency table, follows the multinomial distribution:

$$\{n_{hjk} : 1 \leq h \leq a, 1 \leq j \leq b, 1 \leq k \leq 2\} \sim \mathsf{Mult}(n, \{p_{hjk} : 1 \leq h \leq a, 1 \leq j \leq b, 1 \leq k \leq 2\}), \tag{3.39}$$

where $\sum_h \sum_j \sum_k n_{hjk} = n$ and $\sum_h \sum_j \sum_k p_{hjk} = 1$.

Under this set up, $\hat{d}_{KL}$ in (3.35) can be constructed as follows:

$$\begin{aligned}
\hat{d}_{KL} &= \sum_{k=1}^{2}\sum_{j=1}^{b}\sum_{h=1}^{a} \hat{p}_{hjk} \log\left(\frac{\hat{p}_{hjk} \quad \hat{p_{.j.}}}{\hat{p}_{.jk} \quad \hat{p}_{hj.}}\right) \\
&= \frac{1}{n}\log\left(\prod_{k=1}^{2}\prod_{j=1}^{b}\prod_{h=1}^{a}\left(\frac{\hat{p}_{hjk} \quad \hat{p}_{.j.}}{\hat{p}_{.jk} \quad \hat{p}_{hj.}}\right)^{n_{hjk}}\right) \\
&= \frac{1}{n}\log\left(\frac{\prod_{k=1}^{2}\prod_{j=1}^{b}\prod_{h=1}^{a}(\frac{n_{hjk}}{n})^{n_{hjk}}}{\prod_{k=1}^{2}\prod_{j=1}^{b}\prod_{h=1}^{a}\left(\frac{\frac{n_{.jk}}{n} \quad \frac{n_{hj.}}{n}}{\frac{n_{.j.}}{n}}\right)^{n_{hjk}}}\right). \qquad (3.40)
\end{aligned}$$

Then

$$\begin{aligned}
2n\hat{d}_{KL} &= -2\log\left(\frac{\prod_{k=1}^{2}\prod_{j=1}^{b}\prod_{h=1}^{a}\left(\frac{\frac{n_{.jk}}{n} \quad \frac{n_{hj.}}{n}}{\frac{n_{.j.}}{n}}\right)^{n_{hjk}}}{\prod_{k=1}^{2}\prod_{j=1}^{b}\prod_{h=1}^{a}(\frac{n_{hjk}}{n})^{n_{hjk}}}\right) \qquad (3.41) \\
&= -2\log\frac{L_0}{L_1}.
\end{aligned}$$

In equation (3.41), L_0 is the maximum value of the joint probability of observations n_{hjk}, for all the cells, in the contingency table of (T, S, Z), under the null hypothesis in (3.32), where $\hat{p_{hjk}} = \frac{\frac{n_{.jk}}{n} \quad \frac{n_{hj.}}{n}}{\frac{n_{.j.}}{n}}$ is the restricted maximum likelihood estimate (MLE) for p_{hjk}, for $1 \leq h \leq a$, $1 \leq j \leq b$, $1 \leq k \leq 2$. L_1 is the maximum value of the joint probability of n_{hjk}, under an alternative hypothesis in equation (3.32), where $\hat{p}_{hjk} = \frac{n_{hjk}}{n}$ is the unrestricted MLE, for $1 \leq h \leq a$, $1 \leq j \leq b$, $1 \leq k \leq 2$.

Therefore, $2n\hat{d}_{KL}$ is the likelihood ratio statistic, and approximately chi-square distributed with degrees of freedom (df) equal to the degrees of freedom under the alternative hypothesis $df_{|H_1}$ minus the degrees of freedom under the null hypothesis $df_{|H_0}$.

Under the alternative hypothesis we have a total of $2ab$ parameters and a constraint $\sum_{k=1}^{2}\sum_{j=1}^{b}\sum_{h=1}^{a} p_{hjk} = 1$. Thus, the total number of independent (free) parameters are $df = 2ab - 1$. By free parameter we mean those p_{hjk}, for $1 \leq h \leq a$, $1 \leq j \leq b$, $1 \leq k \leq 2$, that have to be estimated.

Under the null hypothesis, we have $p_{hjk} = \frac{p_{hj.}\ \ p_{.jk}}{p_{.j.}}$, that means p_{hjk} can be determined through the marginal probabilities of $p_{hj.}$, $p_{.jk}$ and $p_{.j.}$, for $1 \leq h \leq a$, $1 \leq j \leq b$, $1 \leq k \leq 2$. The total number of parameters $p_{hj.}$, for $1 \leq h \leq a$, $1 \leq j \leq b$, are ab, and we have one constraint $\sum_{j=1}^{b}\sum_{h=1}^{a} p_{hj.} = 1$. Thus, the number of free parameters are equal to $ab - 1$. Having all $p_{.jk}$, for $1 \leq h \leq a$, $j \leq b$, $1 \leq k \leq 2$, we can simply sum over k for $1 \leq k \leq 2$ and get $p_{.j.}$, for $1 \leq j \leq b$. The total number of $p_{.jk}$, for $1 \leq j \leq b$ and $1 \leq k \leq 2$, are $2b$, where b of them can be determined through $p_{.j.}$, for $1 \leq j \leq b$. Consequently, only $(2b - b)$ out of $2b$ parameters are free and should be estimated. Thus, the total number of free parameters under the null hypothesis is $df_{|H_0} = ab - 1 + b$.

The asymptotic distribution of $2n\hat{d_{KL}}$ is chi-square with $(2ab-1)-(ab-1+b)$ degrees of freedom, symbolically represented as follows:

$$2n\hat{d}_{KL} \xrightarrow{d} \chi^2_{b(a-1)} \tag{3.42}$$

As a result, the $p-value$ to test the hypotheses in equation (3.34), based on the asymptotic distribution of $2n\hat{d}_{KL}$, can be calculated as follows:

$$p - value_{asymptotic} = P(\chi^2_{b(a-1)} > 2n\hat{d}_{KL_{obs}}) \tag{3.43}$$

where $\hat{d}_{KL_{obs}}$ is the observed value of a random variable $\hat{d_{KL}}$ from the sample.

Therefore, we reject the null hypothesis in equation (3.34) if $p-value_{asymptotic}$ is less than the false positive rate for testing the hypotheses in equation (3.34).

Consequently, we reject the null hypothesis in equation (3.32), if we reject the null hypothesis in equation (3.34), leading to the conclusion that $f_{T|S,Z}(t|s,z) \neq f_{T|S}(t|s)$

In Chapter 4, we will introduce three asthma trials. Validation of the suggested surrogate endpoint in these trials using parametric and non-parametric methods introduced earlier, is the major concentration of this chapter.

CHAPTER 4

Surrogate Endpoint in Respiratory Clinical Trials

4.1 Introduction

Respiratory diseases are common type of illnesses that influence the lungs and other parts of the respiratory system. They affect tens of millions of people and are causing significant numbers of death globally. They may be caused by infection, by smoking tobacco, or by breathing in secondhand tobacco smoke, radon, asbestos, or other forms of air pollution. In this chapter, we will present asthma as one of the most common respiratory diseases. Then, we introduce three asthma trials along with the suggested candidate as a surrogate endpoint in these trials. The rest of Chapter 4 is devoted to validate the proposed candidate surrogate marker for the true endpoint, by applying different methods introduced in Chapter 3.

4.2 Asthma

Asthma is a chronic reversible inflammatory disorder of the airways which is characterized by recurrent attacks of breathlessness, wheezing, coughing, chest tightness or pain that inflames and narrows the airways. These symptoms mostly happen at night and early in the morning and the level of severity is not the same for all patients. Asthma may happen in people of all ages but it most often appears for the first time during childhood. For more details see e.g. Reddel et al. (2009)

According to the U.S. Department of Health & Human Services (2000), the airways are tubes that carry air into and out of lungs. Asthma patients have very sensitive and inflamed airways. In asthma patients, the airways may over-react to certain inhaled substances. This strong response could cause the muscles around the airways become to tighten. Tightening of the muscles around the airways makes the airways narrower which reduces the flow of air into and out of the lungs. Moreover, the strong reaction of airways could make the cells in the airways to produce more mucus than usual, which may narrow the airways in long term. Mucus is a sticky, thick liquid that keeps the airways moist and further traps any dust and dirt in the inhaled air. Eventually asthma symptoms could appear every time the airways are provoked by certain inhaled substances.

There are many different factors that makes asthma symptoms appear or amplify them. Doctors can determine these factors by running some tests on asthma patients. Some of these factors are as follows:

- Allergens like pets' dander, dust, and pollen from trees, grasses and flowers.
- Stress.
- Physical activity.
- Viral upper respiratory infections such as cold and flu.
- Certain foods and medicines.
- Extreme weather condition like cold air or extremely dry, wet or windy weather.

- Irritants in the air like smoke, air pollution, chemical fumes and strong odors.

Global Asthma Network (2014), reported that the estimated number of people with asthma in the world may be as many as 334 million based on the best data available. Asthma cannot be cured but the symptoms can be controlled by correct and regular treatments. Even when a patient looks asthma free, the disease still exists and could appear anytime. If an asthma patient does not receive necessary and adequate medication to control the disease properly, the symptoms may get worse and there is high risk of asthma attack, hospitalization and death. Therefore, it is very important for an asthma patient to keep record of asthma symptoms in a diary to see how well treatments are controlling the disease, consulting with a doctor and taking proper and effective medicines regularly to control the disease. Moreover, there is high demands for a cheaper and faster production therapy compared to the current ones that could hopefully prevent the progression of the symptoms in asthma. Many studies has been done on different Biomarkers for asthma with a hope of finding a potential surrogate endpoints that could be use as a replacement for true endpoints. For example, Zissler et al. (2016) introduce some of the current and futures biomarkes in allergic asthma, and Verrills et al. (2011) identify diagnostic biomarkers for asthma and chronic obstructive pulmonary disease (COPD).

4.3 Research Questions

In the previous sections we introduced asthma as a chronic lung disease and the importance of searching for potential surrogate endpoints among available biomarkers for Asthma. We still need to go through two more definitions, severe exacerbation and diary events variable, before we introduce our candidate as a surrogate endpoint and the true endpoint in Asthma study, and discuss possible research questions.

Severe exacerbation is defined as a sudden worsening of disease symptoms that require urgent action to prevent a serious outcome, such as hospitalization or death. As for the diary event variable, in our study, doctors defined diary events variable by threshold and slope criteria using the following diary variables: lung function, rescue medication, asthma symptoms score (0-3) and night-time awakenings. See Fuhlbrigge et al. (2017) for more details on diary events variable and exacerbation.

Now consider three trials on asthma patients: STEAM Rabe et al. (2006), of size $n_1 = 608$, STEP Scicchitano et al. (2004), of size $n_2 = 1793$ and STAY O'Byrne et al. (2005), of size $n_3 = 1447$ patients. In each trial, treatment Z is 1 for the new drug or 0 for active control (current drug):

$$Z = \begin{cases} 0 & \text{Pulmicort+Bricanyl.} \\ 1 & \text{Symbicort SMART.} \end{cases} \tag{4.1}$$

Then, we define a pair of true and surrogate endpoints (T, S), as follows:

$$T = \begin{cases} 0 & \text{no severe exacerbation occurs during the trial.} \\ 1 & \text{one or more severe exacerbation occurs.} \end{cases} \tag{4.2}$$

$$S = \begin{cases} 0 & \text{no diary events occurs during the trial.} \\ 1 & \text{one or more diary events occurs.} \end{cases} \tag{4.3}$$

We now have all the necessary ingredients to develop our research questions. The most important questions are as follows:

- How much of the treatment effect on severe exacerbation can be captured by the treatment effect on diary events variable?
- Can diary events variable be a legitimate replacement for severe exacerbation?
- Can we use the treatment effect on diary events variable to predict the treatment effect on severe exacerbation, where the data on sever exacerbation is missing or not yet observed?

These are some of the questions someone could ask regarding surrogate and true endpoints. In the remaining chapters we will intend to answer these questions.

4.4 Logistic Regression and Likelihood Ratio Test (LRT) for the Validation of Prentice's Criteria in the Asthma Trials

The data of the STEAM trial are shown in Table 4.1.

Table 4.1: Three way contingency table for data from the STEAM asthma trial.

		Z	
T	S	0	1
0	0	194	252
	1	74	42
1	0	5	5
	1	28	8

Results from validation of Prentice's criteria using logistic regression in the STEAM trial are shown in Tables 4.2 and 4.3.

In table 4.2 results from testing the coefficients of logistic regression models (3.21), (3.22) and (3.23) show that, α, β and γ are significant, implying that the first three Prentices criteria are satisfied.

As for the verification of the last Prentice's criterion, table 4.3 shows there is no evidence on the significance of δ and β_S in the model, provides evidence that some effect of Z on T is mediated through S.

Another approach to check for the validity of the fourth Prentice's criterion is the likelihood ratio test. Table 4.4 presents the results of LRT that checks for the validity of the last Prentice's criterion in the first trial. As we see the results agree with the one from testing the coefficients in the corresponding logistic models.

Table 4.2: Results of using logistic regression to validate first three Prentice's criteria in the STEAM asthma trial.

Coefficient	Estimate	Std. error	$p-value$
$\mu_{S\|Z}$	-0.67	0.12	< 0.001
α	-0.97	0.20	< 0.001
$\mu_{T\|Z}$	-2.09	0.18	< 0.001
β	-1.02	0.34	0.002
$\mu_{T\|S}$	-3.80	0.32	< 0.001
γ	2.63	2.63	< 0.001

Table 4.3: Results of using logistic regression to validate fourth Prentice's criterion in the STEAM asthma trial.

Coefficient	Estimate	Std. error	$p-value$
$\mu_{T\|Z,S}$	-3.66	0.45	< 0.001
γ_Z	2.69	0.50	< 0.001
β_S	-0.26	0.64	0.68
δ	-0.42	0.78	0.59
$\mu_{T\|Z,S}$	-3.52	0.36	< 0.001
γ_Z	2.52	0.38	< 0.001
β_S	-0.55	0.36	0.13

Table 4.4: Result of using LRT to validate fourth Prentice's criterion in the STEAM trial.

Model Comparison	df.	$p-value$
Model with interaction (3.24) vs Model without interaction term (3.25)	1	0.58
Model without interaction (3.25) vs model without Z (3.23)	1	0.12

Note that, we could use different methods to compare parametric models, and the results of using these different methods agreed with each other in our problem.

In the same way, results from applying logistic regression on the STEP and STAY trials both imply that the first three Prentices criteria are satisfied, but not the fourth criterion. Tables 4.5, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11 and 4.12 show the data and results in the STEP and STAY trials.

Table 4.5: Three way contingency table for data from the STEP asthma trial.

		Z	
T	S	0	1
0	0	469	572
	1	215	189
1	0	55	45
	1	154	94

Table 4.6: Results of using logistic regression to validate first three Prentice's criteria in the STEP asthma trial.

Coefficient	Estimate	Std. error	$p-value$
$\mu_{S\|Z}$	-0.35	0.07	< 0.001
α	-0.43	0.10	< 0.001
$\mu_{T\|Z}$	-1.18	0.08	< 0.001
β	-0.51	0.12	< 0.001
$\mu_{T\|S}$	-2.34	0.10	< 0.001
γ	1.85	0.13	< 0.001

Table 4.7: Results of using logistic regression to validate fourth Prentice's criterion in the STEP asthma trial.

Coefficient	Estimate	Std. error	$p-value$
$\mu_{T\|Z,S}$	-2.14	0.14	< 0.001
γ_Z	1.81	0.18	< 0.001
β_S	-0.40	0.21	0.06
δ	0.03	0.27	0.90
$\mu_{T\|Z,S}$	-2.15	0.12	< 0.001
γ_Z	1.82	0.13	< 0.001
β_S	-0.38	0.13	0.003

Table 4.8: Result of using LRT to validate fourth Prentice's criterion in the STEP trial.

Model Comparison	df.	$p-value$
Model with interaction (3.24) vs Model without interaction term (3.25)	1	0.90
Model without interaction (3.25) vs model without Z (3.23)	1	0.003

Table 4.9: Three way contingency table for data from the STAY asthma trial.

		Z	
T	S	0	1
0	0	380	483
	1	198	136
1	0	42	35
	1	113	60

Table 4.10: Results of using logistic regression to validate first three Prentice's criteria in the STAY asthma trial.

Coefficient	Estimate	Std. error	$p-value$
$\mu_{S\|Z}$	-0.30	0.07	< 0.001
α	-0.66	0.11	< 0.001
$\mu_{T\|Z}$	-1.32	0.09	< 0.001
β	-0.56	0.14	< 0.001
$\mu_{T\|S}$	-2.42	0.12	< 0.001
γ	1.76	0.15	< 0.001

Table 4.11: Results of using logistic regression to validate fourth Prentice's criterion in the STAY asthma trial.

Coefficient	Estimate	Std. error	$p-value$
$\mu_{T\|Z,S}$	-2.20	0.16	< 0.001
γ_Z	1.64	0.20	< 0.001
β_S	-0.42	0.24	0.07
δ	-0.16	0.31	0.59
$\mu_{T\|Z,S}$	-2.25	0.14	< 0.001
γ_Z	1.71	0.15	< 0.001
β_S	-0.32	0.15	0.03

Table 4.12: Result of using LRT to validate fourth Prentice's criterion in the STAY trial.

Model Comparison	df.	$p-value$
Model with interaction (3.24) vs Model without interaction term (3.25)	1	0.59
Model without interaction (3.25) vs model without Z (3.23)	1	0.032

4.5 Odds Ratio for the Validation of Prentice's Criteria in the Asthma Trials

Tables 4.14 and 4.15 show the result of testing for the odds ratios in the first trial. Table 4.14 represents the odds ratios, their confidence intervals and the corresponding $p-values$ for the test of independence.

Table 4.13: Marginal contingency tables for data from the STEAM trial.

	Z		S	
	0	1	0	1
T				
0	268	294	446	116
1	33	13	10	36
S				
0	199	257		
1	102	50		

Results from tables 4.14 and 4.15 indicate that all Prentice's criteria are satisfied which is in consistent with the results we got from testing the parameters in the logistic regression models.

Table 4.14: Estimates of odds ratios to validate the first three Prentice's criteria in the STEAM trial.

Odds ratio	Estimate	C.I.
OR_{SZ}	0.38	$(0.26, 0.56)$
OR_{TZ}	0.36	$(0.18, 0.70)$
OR_{TS}	13.84	$(6.67, 28.71)$

Table 4.15: Estimates of odds ratios to validate the fourth Prentice's criterion in the first asthma trial.

Odds ratio	Estimate	95% confidence interval
$OR_{TZ\|S=0}$	0.77	$(0.22, 2.70)$
$OR_{TZ\|S=1}$	0.50	$(0.21, 1.20)$

In the similar way, using odds ratios in the STEP and STAY trials to validate Prentice's criteria, we observed that the first, second and third Prentice's criteria are fulfilled but not the last one. The results for the STEP and STAY trials are shown in tables 4.16, 4.17, 4.18 and 4.19.

Table 4.16: Estimates of odds ratios to validate the first three Prentice's criteria in the STEP trial.

Odds ratio	Estimate	C.I.
OR_{SZ}	0.65	$(0.54, 0.79)$
OR_{TZ}	0.60	$(0.47, 0.76)$
OR_{TS}	6.39	$(4.93, 8.27)$

Table 4.17: Estimates of odds ratios to validate the fourth Prentice's criterion in the STEP asthma trial.

Odds ratio	Estimate	95% confidence interval
$OR_{TZ\|S=0}$	0.67	$(0.44, 1.01)$
$OR_{TZ\|S=1}$	0.69	$(0.50, 0.96)$

Table 4.18: Estimates of odds ratios to validate the first three Prentice's criteria in the STAY trial.

Odds ratio	Estimate	C.I.
OR_{SZ}	0.51	$(0.41, 0.64)$
OR_{TZ}	0.57	$(0.43, 0.76)$
OR_{TS}	5.80	$(4.31, 7.81)$

Table 4.19: Estimates of odds ratios to validate the fourth Prentice's criterion in the STAY asthma trial.

Odds ratio	Estimate	95% confidence interval
$OR_{TZ\|S=0}$	0.65	$(0.41, 1.05)$
$OR_{TZ\|S=1}$	0.77	$(0.53, 1.13)$

4.6 Proportion Treatment Explained (PE) for the Asthma Trials

Proportion treatment explained along with their confidence sets for the STEAM, STEP and STAY trials are shown in table (4.20). Confidence limits are calculated using Fieller's theorem explained in Sections 2.3.1 and 3.3.3.

Table 4.20: Estimates of PE in the STEAM, STEP and STAY asthma trials

Trial	Estimate	CI
STEAM	0.46	$(0.19, 1.40)$
STEP	0.26	$(0.09, 0.59)$
STAY	0.42	$(0.21, 0.91)$

As we mentioned in Sec 3.3.3, there are some limitation in using PE for the validation of surrogate endpoints. In the STEAM and STAY trials the confidence intervals for PE are wide and the lower limits are less than 0.5. Moreover the confidence upper limit for PE in the first trial exceeds 1. Therefore, PE and its confidence interval are not informative measures for the validation of surrogate endpoints in the STEAM and STAY trials.

The confidence interval for PE in the STEP trial is not wide. If we use this confidence interval along with the result of using logistic regression and odds ratio for the validation of surrogate endpoint in the STEP trial, we can conclude that the surrogate poorly explains the effect of the treatment on the clinical endpoint.

4.7 Non-Parametric Methods for Validation of Prentice Fourth Criterion in the Asthma Trials

Table 4.21 shows $p - value_{asympt}$ and $p - value_{perm}$ for the STEAM, STEP and STAY asthma trials. Note that n is the total sample size in each trial. Moreover, we fixed the total number of permutations to $L = 1000$. The $\chi^2_2(0.95) = 5.99$.

Table 4.21: Results for $2n\hat{d}_{KL}$, $p - value_{asympt}$ and $p - value_{perm}$

Trial	n	$2n\hat{d}_{KL}$	$p - value_{asympt}$	$p - value_{perm}$
STEAM	608	2.72	0.26	0.29
STEP	1793	8.58	0.01	0.01
STAY	1447	4.90	0.09	0.10

In the STEAM trial, results of $p - value_{asympt}$ and $p - value_{perm}$ do not imply any significant evidence to reject the null hypothesis of the form $f_{T|S,Z}(t|s,z) = f_{T|S}(t|s)$. For the STEP trial, we can conclude the existence of significant difference between two distributions $f_{T|S,Z}(t|s,z)$ and $f_{T|S}(t|s)$, based on $p - value$s. For the STAY trial, based on the $p - value$s, we found no significant difference between two distributions $f_{T|S,Z}(t|s,z)$ and $f_{T|S}(t|s)$. Earlier in sections 4.4 and 4.5, the result from logistic regression and odds ratio did not give us enough evidence to verify the fourth Prentice's criterion. However, results from using asymptotic distribution of $2n\hat{d}_{KL}$ and permutation test to approximate the distribution of $\hat{d}_{KL}$ both verify the fourth Prentice's criterion. This is actually an interesting result. We may conclude, it is possible that the reason that the logistic regression and odds ratio could not validate the Prentice's fourth criterion is due to the model

misspecification. The related scripts that generate the results in this chapter can be found in Chapter 8.

CHAPTER 5

The Equivalence Test and its Application in Surrogacy Evaluation

5.1 Introduction

In the previous chapters, we explored evaluation of surrogate endpoints in single trial studies, using parametric methods and an entropy based non-parametric method. However, there are weaknesses to both approaches that we are going to discuss in the following.

In the parametric approach, the results highly depend on choosing correct parametric models. Moreover, validating the fourth Prentice's criterion is equivalent to proving the null hypothesis in testing parameters of the chosen model. Proving the null hypothesis is not what we usually look for in the common hypothesis testing framework.

An entropy based non-parametric approach solves the problem of choosing a correct parametric model, since it does not consider any distributional assumption on the endpoints. However, the second limitation on proving the null hypothesis, still stands.

In reality, perfect surrogacy is difficult to achieve. Therefore, it is of interest to find a region (bound) where the surrogate is valid on it.

The randomized clinical trial where the goal is to determine one treatment is superior to one another is called superiority test (trial), and the standard hypothesis test can be applied to show the superiority of the treatment. Often researchers inappropriately use a non-significant standard hypothesis test for superiority as proof of no difference between the two treatments. The correct approach to demonstrate that the two treatments are more likely the same in efficacy is to design an equivalence test. The equivalence test aims to show that two treatments (items) are not much different from each other. In this chapter, we will introduce equivalence tests along with relevant procedures to design and analyze them. Further, we introduce the notion of surrogacy region and the application of equivalence test in surrogacy context. To this end, we will reformulate the Prentice's fourth criterion, and apply equivalence test to show that the coefficient in the logistic model is not much far from zero. By using equivalence test instead of standard hypothesis test, we correctly treat scenarios where we cannot reject the null hypothesis in surrogacy context.

5.2 Equivalence Test

At the present, it is getting more difficult to develop a new treatment with higher efficacy compared to the ones that are currently existing in the market. Therefore, the concentration of clinical studies is to develop treatments that are equivalent in efficacies with the ones currently in use, and pose much benefits in terms of less side effects, less cost, easy in application, and less drug interactions.

In the standard hypothesis testing framework for example t-test, the burden of proof rests on that the two treatments are different. Then, if we do not find enough evidence to support the difference of two treatments, the equality cannot be rejected.

To show the equivalence between two treatments, it is tempting to simply perform the standard t-test. Then, if the difference is statistically significant, it seems clear that the two treatments are not equivalent. And if the difference is not statistically significant, it seems to make sense to conclude that the two treatments are equivalent. However, this approach is incorrect, and leads to invalid conclusions. To better grasp the issue, consider the courtroom, where a person analyzing data is the judge. The hypothesis test is the trial, the alternative hypothesis is like the prosecution, and the null hypothesis is the defendant. If the evidence presented by the prosecution cannot prove that the defendant is guilty beyond the reasonable doubt (say, with 95% certainty), the prosecution has not still proved that the defendant is innocent. However, based on the evidence in hand, the prosecution cannot reject the possibility of the defendant being innocent. Therefore, the judge announces the verdict as not guilty. It does not mean that the defendant is innocent, since that has not been proven. It means that the prosecution failed to convince the judge to disregard the assumption of innocence.

In equivalence studies the goal is to show equivalency. Thus, the burden of proof lies on that the two treatments are equivalent. Then, if the evidence in favor of equivalence is not strong enough, nonequivalence cannot be rejected. Essentially, the null and alternative (research) hypothesis in equivalent studies are reversed from those in the standard t-test.

The equivalence testing is widely used within the pharmaceutical researches to demonstrate the equivalence between two drug formulations, therapies or treatments. The most common approach for equivalence testing

under a two-arm parallel design [1] is the two one sided test (TOST), proposed by Westlake (1981) and Schuirmann (1981). TOST is not the uniformly most powerful (UMP) test, however it is the most common used method by regularity agencies like FDA, to demonstrate equivalence. The UMP test for equivalence was addressed comprehensively by Wellek (2002). More references on different methods for equivalence testing can be found in Hauschke et al. (2007), Meyners (2012) and Chu and Liu (2008).

Unfortunately, in some fields of research, it is common to use the traditional t-test to demonstrate the equivalence, see for example Allan and Cribbie (2013). According to Rogers et al. (1993) and Blackwelder (1982) the results from the equivalence test and the traditional t-test to show no difference between two items are not necessarily in agreement or opposed to each other. If both performed, three scenarios are possible: 1) equivalence of two items will be rejected by both methods, 2) or will not be rejected by either of them, 3) or will be rejected by one and will not be rejected by the other. As a result, failing to reject a no difference between two items (the null hypothesis in hypotheses testing framework) using the traditional t-test approach does not necessarily implies equivalence.

Like any other experiment, the statistical power calculations and sample size determinations has vital role in design and analysis of equivalence testing. Relevant literatures on sample size determination and the statistical power calculation for equivalence testing, can be found in Bristol (1993), Chow et al. (2002), Chow and Wang (2001), Diletti et al. (1991), Liu and Chow (1992), Muller-Cohrs (1990), Phillips (1990), Schuirmann (1987), Siqueira et al. (2005) and Wang and Chow (2002).

Finally, the concept of equivalent is not only limited to the pharmaceutical studies. It can be used likewise in other field of researches to demonstrate the resemblance of the means (proportions) for product measurements or process measurements.

[1] In clinical studies, parallel design is used to compare two or more treatments, like T_1, T_2,..., T_k. Participants are assigned to one and only one of these k groups randomly, treatments are administered, and then the results are compared. Note that the key element of the parallel design is randomization. Randomization guarantees that the results have a lower risk of being biased. Parallel design is also known as between patient design and non-crossover design. For more details see for example Ofori-Asenso and Adom Agyeman (2015).

5.2.1 Two One Sided Test (TOST) Procedure

Consider the common hypothesis test where we are comparing two populations by comparing their means.

$$H_0 : \mu_1 = \mu_2 \;\; vs. \;\; H_1 : \mu_1 \neq \mu_2, \tag{5.1}$$

Further, assume we are interested in the case where the null hypothesis is not rejected.

Now the important question is when we cannot reject the null hypothesis, what can be said about the population means?! We usually consider these populations similar. But this is under the idea that we are looking for the evidence to show that they are different. However, if we want to show that the means of two populations are similar, the correct approach is that we first assume they are different, and then try to gather evidence to the contrary. Therefore the equivalence test set up will be as follows:

$$H_0 : \mu_1 \neq \mu_2 \;\; vs. \;\; H_1 : \mu_1 = \mu_2, \tag{5.2}$$

We are now on the right track and thus the important question is how close is close enough to be considered the same?! So the equivalence test set up can be rewritten as follows:

$$H_0 : |\mu_1 - \mu_2| > d \;\; vs. \;\; H_1 : |\mu_1 - \mu_2| < d \tag{5.3}$$

In equivalence testing, the null hypothesis (H_0) is a difference of d or more, that can be restated as follows:

$$H_0 : \mu_1 - \mu_2 < -d \;\text{ or }\; \mu_1 - \mu_2 > d \tag{5.4}$$

This leads to the most famous form of equivalence testing approach named the two-one sided test (TOST). We assumes that the two populations are normally distributed and independent of each other. Moreover, the variances of both populations are known and equal to σ. Then the two one-sided test statistics can be constructed as follows:

$$\frac{\bar{Y_1} - \bar{Y_2} + d}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \quad , \quad \frac{\bar{Y_1} - \bar{Y_2} - d}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \tag{5.5}$$

where n_1 and n_2 are sample sizes and

$$\bar{Y_k} = \frac{\sum_{i=1}^{n_k} Y_{ki}}{n_k} \quad \text{for} \quad k = 1, 2 \tag{5.6}$$

Then, we conclude that the two population means are equivalent at the level of α, if and only if, both the following inequalities can be rejected.

$$\frac{\bar{y_1} - \bar{y_2} + d}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_{1-\alpha} \quad , \quad \frac{\bar{y_1} - \bar{y_2} - d}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -z_{1-\alpha} \tag{5.7}$$

where $\bar{y_1}$ and $\bar{y_2}$ are observed values for $\bar{Y_1}$ and $\bar{Y_2}$ from the samples, and $P(Z \leq z_{1-\alpha}) = 1 - \alpha$ where Z is a random variable from normal standard distribution.

Likewise the standard hypothesis testing for the mean difference where two populations are normally distributed, we can assume that 1) the variances are known and different in each population, 2) the variances are unknown and the same, 3) the variances are unknown and different from each other.

To deal with the first scenario, we only need to replace σ^2/n_1 and σ^2/n_2 by σ_1^2/n_1 and σ_2^2/n_2 in equation (5.7) respectively, and proceed the procedure.

In the second scenario, we reject the non-equivalence at the level of α, if and only if, the following inequalities can be rejected simultaneously.

$$\frac{\bar{y_1} - \bar{y_2} + d}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{1-\alpha,n_1+n_2-2} \quad , \quad \frac{\bar{y_1} - \bar{y_2} - d}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t_{1-\alpha,n_1+n_2-2} \tag{5.8}$$

where S_p^2 is the pooled sample variance and an unbiased estimator of the common variance σ^2:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{5.9}$$

and

$$S_k^2 = \frac{\sum_{i=1}^{n_k}(Y_{ki} - \bar{Y_k})^2}{n_k - 1} \quad \text{for} \quad k = 1, 2 \tag{5.10}$$

Further, $P(T \leq t_{1-\alpha,n_1+n_2-2}) = 1 - \alpha$, where T is a random variable from Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Finally, in the third scenario, we conclude equivalence if both the following inequalities can be rejected at the level of α:

$$\frac{\bar{y_1} - \bar{y_2} + d}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} > t_{1-\alpha,r} \quad , \quad \frac{\bar{y_1} - \bar{y_2} - d}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} < -t_{1-\alpha,r} \tag{5.11}$$

where

$$r = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} \tag{5.12}$$

and S_1^2 and S_2^2 can be calculated using equation (5.10). Moreover, $P(T \leq t_{1-\alpha,r}) = 1 - \alpha$, where T is a random variable from Student's t-distribution with degrees of freedom equal to r.

5.2.2 Confidence Interval (CI) Procedure

We first construct $(1 - 2\alpha)\%$ confidence interval for the mean difference as follows:

$$\bar{y_1} - \bar{y_2} \pm z_{1-\alpha}\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \tag{5.13}$$

Then, if both one-sided tests are rejected jointly, the $(1 - 2\alpha)\%$ confidence interval is completely included in the $(-d, +d)$ interval. Therefore, if $(-d, +d)$ fully covers the $(1 - 2\alpha)\%$ confidence interval, then the equivalence can be concluded. Thus, we can declare that the mean difference is very likely to be zero. Likewise the previous section, we could consider three more scenarios for the variances of two normal distributions, and construct the corresponding confidence intervals.

Lastly, note that testing equivalence between two items using the TOST and confidence interval procedures is not only limited to test the means. It can also be achieved through testing other parameters like proportions, odds ratios, hazard ratios and etc. We can also establish equivalence test for interested parameters when we deal with one sample or more than two samples.

Example 7. Consider two independent populations that are normally distributed. Further assume that the variances are unknown and not equal. Samples of size $n_1 = 95$ and $n_2 = 89$ were taken from the first and second population respectively. The mean and standard deviation of these two samples were recorded as $\bar{x_1} = 5.25$, $s_1 = 0.95$ and $\bar{x_2} = 5.22$, $s_2 = 0.83$. We want to know if we could declare that the two populations are equivalent in terms of their means. The equivalence limit was set to $d = 0.48$. More details on the example can be found in Lakens (2017a) and Lakens (2017b).

The following code lines perform the equivalent test using the TOSTER Lakens (2017b) package in R:

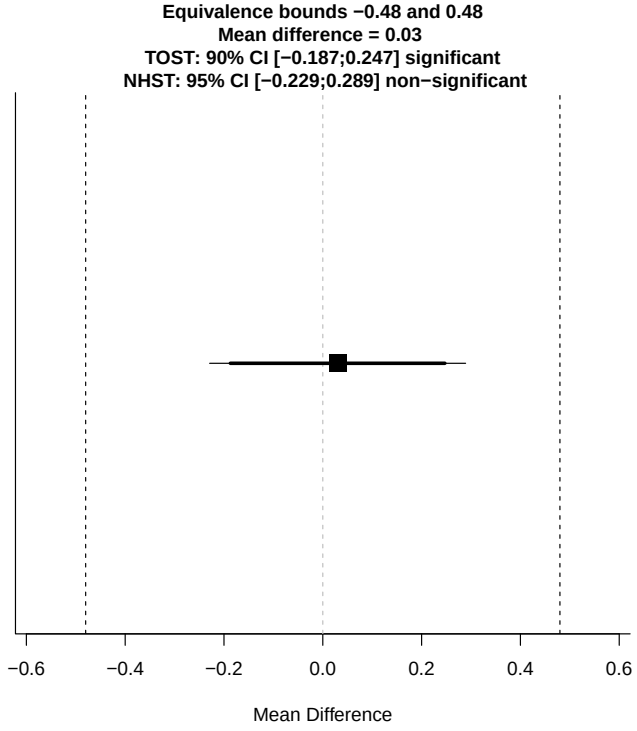
```
##
install.packages("TOSTER")
library("TOSTER")
TOSTtwo.raw(m1=5.25,m2=5.22,sd1=0.95,sd2=0.83,n1=95,
n2=89,low_eqbound=-0.48, high_eqbound=0.48,
alpha = 0.05)
##
```

that gives us the following output and plot:

```
##
Using alpha = 0.05 Welch's t-test was non-significant,
t(181.1344) = 0.2284794, p = 0.8195313

Using alpha = 0.05 the equivalence test based on
Welch's t-test  was significant,
t(181.1344) = -3.42719, p = 0.0003773807
TOST results:
t-value 1    p-value 1 t-value 2    p-value 2
df
1  3.884149 7.190772e-05  -3.42719 0.0003773807 181.1344

Equivalence bounds (raw scores):
low bound raw high bound raw
1         -0.48            0.48

TOST confidence interval:
Lower Limit 90% CI raw Upper Limit 90% CI raw
1              -0.1870843               0.2470843
##
```

The equivalence limits is shown with vertical dashed lines on points -0.48 and $+0.48$ in figure 5.1. As we can see, the 90% CI (the solid horizontal line) is included in the equivalence region (bounds) $(-0.48, +0.48)$, therefore we can declare the equivalence.

Figure 5.1: Equivalence region (bound) for mean difference



5.2.3 Equivalence Limit d

Results of an equivalence test is influenced by the choice of d. A small value of d could result in a tighter equivalence region $(-d, +d)$, that makes it much harder to demonstrate equivalence. Therefore, determining the equivalence limit d, before collecting any data, is an important task.

According to Walker and Nowacki (2011) "An equivalence study should be designed to minimize the possibility that a new therapy that is found to be equivalent to the current therapy can be nonsuperior to a placebo." To this end, we could choose d based on the limit of superiority of the current therapy against the placebo, using previous studies.

Determining the equivalence limit is not the aim of the current thesis. Therefore, the concentration of the rest of this chapter is on establishing

equivalence test and its application in surrogacy evaluation context. Thus, we will assume that d is given, and then construct TOST and confidence interval to explore equivalence.

Moreover note that, the equivalence region $(-d, +d)$ is not necessarily symmetric around zero. The more general form to the equivalence region is (d_1, d_2), where $|d_1|$ and $|d_2|$ could be different and d_1 and d_2 could be both negative or positive, or may have different signs.

5.3 The Surrogacy Region

As we mentioned earlier, the last Prentice's criterion plays an important role in the validation of surrogate endpoint. However, proving the validity of this condition is not straight forward. The condition holds if we could somehow prove that $f_{T|S,Z}(t|s,z) = f_{T|S}(t|s)$. A common approach is to perform a statistical hypothesis testing where the fourth Prentice's criterion is considered as a null hypothesis. However, proving no difference using a common statistical hypothesis testing is not an appropriate approach of dealing with such an issue. Therefore, testing such a null hypothesis is not a proper procedure to validate a good (valid or perfect) surrogate. However, it very well can be used to reject a poor surrogate.

Hence, we will get assist from the equivalence method in order to tackle this problem properly. We will also introduce the notion of the surrogacy region in which the candidate surrogate S is still a valid surrogate for the true endpoint T, in the presence of the treatment Z. Advantages of using equivalence procedure are as follows:

- It is the correct approach to demonstrate that the difference between $f_{T|S,Z}(t|s,z)$ and $f_{T|S}(t|s)$ is not significant.
- It determines the surrogacy region properly.

Now consider the fourth Prentice's criterion as follows:

$$f_{T|S,Z}(t|s,z) - f_{T|S}(t|s) = 0 \tag{5.14}$$

which can be reformulated as

$$d\big(f_{T|S,Z}(t|s,z) - f_{T|S}(t|s)\big) = 0 \tag{5.15}$$

where d represents any distance (divergence) function that measures the difference between two given items. for example $d(.,.)$ could be the Kullback Leibler measure of divergence. Following the equivalence framework to show that the difference between $f_{T|S,Z}(t|s,z)$ and $f_{T|S}(t|s)$ is not significant, we have:

$$H_0 : d\big(f_{T|S,Z}(t|s,z) - f_{T|S}(t|s)\big) > d \;\; vs. \;\; H_1 : d\big(f_{T|S,Z}(t|s,z) - f_{T|S}(t|s)\big) < d \tag{5.16}$$

In the following, will show the application of this concept for the validation of the fourth Prentice's criterion, $f_{T|S,Z}(t|s,z) = f_{T|S}(t|s)$.

5.3.1 Equivalence Test for the Parameters of Logistic Models (3.15) and (3.16)

In Chapter 3, to validate the fourth Prentice's criterion we showed that the null hypotheses of $\beta_S = 0$ and $\beta_S = \delta = 0$, the parameters of the models (3.15) and (3.16), shall not be rejected.

In the first scenario assume that model (3.15) is correct, and there exist no interaction term between S and Z based on some prior information. Then, consider β_S in model (3.15) and assume that the research hypothesis investigates whether β_S is close enough to zero, and therefore it forms the alternative hypothesis. Thus, in equivalence framework we want to test the following hypotheses:

$$H_0 : |\beta_S| > \delta \;\; vs. \;\; H_1 : |\beta_S| < d \tag{5.17}$$

Then, following the TOST procedure we have:

$$\frac{\hat{\beta}_S + d}{\sqrt{\hat{var}(\hat{\beta}_S)}} \;\; , \;\; \frac{\hat{\beta}_S - d}{\sqrt{\hat{var}(\hat{\beta}_S)}} \tag{5.18}$$

These statistics have approximately standard normal distributions in large samples.

Then, we conclude that β_S is close enough to zero, if and only if, we could reject the following inequalities simultaneously.

$$\frac{\hat{\beta}_S + d}{\sqrt{\hat{var}(\hat{\beta}_S)}} > z_{1-\alpha} \;\; , \;\; \frac{\hat{\beta}_S - d}{\sqrt{\hat{var}(\hat{\beta}_S)}} < -z_{1-\alpha} \tag{5.19}$$

Further, $(1-2\alpha)\%$ confidence interval for β_S can be constructed as follows:

$$\hat{\beta}_S \pm z_{1-\alpha}\sqrt{\hat{var}(\hat{\beta}_S)} \tag{5.20}$$

If the equivalence region $(-d, +d)$ contains the $(1-2\alpha)\%$ confidence interval, then we declare β_S is close enough to zero.

In the second scenario we consider model (3.16), and we aim to show that β_S and δ are close enough to zero. Therefore, using the equivalence set up we have:

$$H_0 : \begin{cases} \beta_S \neq 0 \\ \delta \neq 0 \end{cases} \quad vs. \qquad H_1 : \begin{cases} \beta_S = 0 \\ \delta = 0 \end{cases}$$

which can be represented as:

$$H_0 : \begin{cases} |\beta_S| > d_1 \\ |\delta| > d_2 \end{cases} \quad vs. \qquad H_1 : \begin{cases} |\beta_S| < d_1 \\ |\delta| < d_2 \end{cases}$$

To demonstrate that β_S and δ are not simultaneously different from zero, we construct the Bonferroni confidence intervals. Remember that, the Bonferroni correction allows us to do multiple testing at the overall level of interest α, by testing each individual hypothesis at a significance level of α/n, where n is the number of hypotheses. Therefore, $(1-2\alpha)\%$ Bonferroni confidence intervals for β_S and δ can be constructed as follows:

$$\hat{\beta}_S \pm z_{1-\alpha/2}\sqrt{\hat{var}(\hat{\beta}_S)} \ , \ \hat{\delta} \pm z_{1-\alpha/2}\sqrt{\hat{var}(\hat{\delta})} \tag{5.21}$$

If the equivalence regions $(-d_1, +d_1)$ and $(-d_2, +d_2)$ cover the Bonferroni confidence intervals for β_S and δ simultaneously, then we conclude that β_S and δ are not significantly different from zero.

In general, the length of the $(1-2\alpha)\%$ simultaneous confidence intervals are wider than the length of each $(1-2\alpha)\%$ individual confidence interval. Another joint confidence intervals is called Hotelling's T^2 confidence interval, which gives wider region than the Bonferroni confidence interval. Thus, if the Bonferroni confidence intervals are not fully cover by the the equivalence regions, neither the Hotelling's T^2 confidence interval are.

5.3.2 Application of The Equivalence Test in Asthma Trials

Consider now the asthma trials STEAM, STEP and STAY, that we introduced in Chapter 4. The $(1 - 2 \times 0.05)\%$ individual confidence intervals for β_S, based on model (3.15) without interaction term, and the $(1 - 2 \times 0.05)\%$ Bonferroni confidence intervals for β_S and δ based on model (3.16) with interaction term are represented in the tables 5.1, 5.2 and 5.3. Related scripts can be found in Chapter 9.

Given the equivalence region $(-d, +d)$ based on prior information, if $(-d, +d)$ contains the $(1 - 2 \times 0.05)\%$ individual confidence intervals for β_S, we will conclude that β_S is small enough that to be considered equivalent to zero.

Likewise, given any equivalence regions $(-d_1, +d_1)$ and $(-d_2, +d_2)$ for β_S and δ, if the $(1 - 2 \times 0.05)\%$ Bonferroni confidence intervals for β_S and δ are covered fully by the equivalence regions, we will declare that β_S and δ are not different from zero.

Table 5.1: The $(1 - 2 \times 0.05)\%$ individual confidence intervals for β_S, in STEAM, STEP, STAY asthma trials.

Trial	Estimate	C.I.
STEAM	-0.55	$(-1.01, -0.09)$
STEP	-0.38	$(-0.55, -0.21)$
STAY	-0.32	$(-0.51, -0.13)$

Table 5.2: the $(1 - 2 \times 0.05)\%$ Bonferroni confidence intervals for β_S, in STEAM, STEP, STAY asthma trials.

Trial	Estimate	C.I.
STEAM	-0.26	$(-1.70, 0.86)$
STEP	-0.40	$(-0.41, 0.47)$
STAY	-0.42	$(-0.67, 0.35)$

Table 5.3: the $(1 - 2 \times 0.05)\%$ Bonferroni confidence intervals for δ, in STEAM, STEP, STAY asthma trials.

Trial	Estimate	C.I.
STEAM	-0.42	$(-1.31, 0.79)$
STEP	0.03	$(-0.74, -0.05)$
STAY	-0.16	$(-0.81, -0.02)$

Note that in equivalence testing approach, if the $(1 - 2\alpha)$ confidence interval does not contain zero is not important. To show the equivalence, it is only important to demonstrate that the equivalence region contains the $(1 - 2\alpha)$ confidence interval.

CHAPTER 6

Discussion

Ever since Prentice (1989) introduced the Prentice practical criteria for the validation of surrogate endpoints, many researchers aimed to refine or reformulate the fourth Prentice's criterion. These attempts are based on various statistical frameworks such as information theoretic, parametric and non-parametric methods. However, all these methods were unable to treat the problem of proving no difference in Forth Prentice's criterion properly.

In the current dissertation, we came with the novel idea of applying equivalence method to validate candidate surrogate endpoints. In the parametric and non-parametric approaches presented in the earlier chapters to validate the fourth Prentice criterion, lacking to reject the null hypothesis, somehow result in accepting the null hypothesis. Eventually, the inferences we make are based on the unproven null hypotheses that might not be necessarily correct. The correct inference is that, if we could not find enough evidence to reject $f_{T|S,Z}(t|s,z) = f_{T|S}(t|s)$, it does not imply the equality of these two distributions. Therefore, to deal with the issue of proving no difference we proposed to take advantage of the equivalence test approach. Moreover, we introduced the notion of surrogacy region (bound) where the candidate surrogate is still valid on that region. Note that, the approach we introduced is generic and can incorporate any type of endpoints.

It is our hope that the method suggested in this dissertation can be extended for use in evaluation of surrogate endpoints, in the Bayesian frame work. According to Wellek (2002), to construct the equivalent test in the Bayesian setting, we first need to specify a joint prior distribution $\pi(.)$ of all unknown parameters presenting in the model underlying the data that we aim to analyze. Then, we reject the null hypothesis (non-equivalence) if the posterior probability of the region corresponding to the alternative hypothesis turns out to be larger than a suitably lower bound specified $1-\alpha$ as default.

We also have an interest in determining the possible values for d, and further explore the effect of different values of d on true positive rate (sensitivity) and eventually surrogacy evaluation.

Last but not least, it is always interesting to compare the old methods with the new proposed one. However, bear in mind that the underlying set up (in terms of null and alternative hypotheses and related inferences) in parametric and non-parametric approaches is different from the one in equivalence testing approach. Hence, caution should be taken while comparing them. We may vote that a candidate surrogate is valid after we could not reject the equivalence in $f_{T|S,Z}(t|s,z) = f_{T|S}(t|s)$, using parametric and

non-parametric methods. However, is such an inference correct in the first place?! This is an interesting question that worth further exploration in the future.

CHAPTER 7

Appendix I

7.1 Example 4.

```
#package contains permutation function
install.packages('gtools')
library(gtools)
#
#set.seed(859)
set.seed(9798)
group<-rep(c(0,1), c(200,300))
S1<-rnorm(500)
S2<-rnorm(500, mean=group/2)
#
#group variances are the same
t.test(S1~group, var.equal=TRUE)
#group variances are the same
t.test(S2~group, var.equal=TRUE)
#
S1.diff<-mean(S1[group==1])-mean(S1[group==0])
S2.diff<-mean(S2[group==1])-mean(S2[group==0])
#
L=1000 #number of permutations
S1.diff.perm=rep(NA,L)
S2.diff.perm=rep(NA,L)
#
for (i in 1:L) {
  S1.perm=permute(S1)
  S2.perm=permute(S2)
#
  S1.diff.perm[i]<-mean(S1.perm[group==1])-
mean(S1.perm[group==0])
  S2.diff.perm[i]<-mean(S2.perm[group==1])-
mean(S2.perm[group==0])
}
#
x11()
hist(S1.diff.perm,xlim =range(-0.4,+0.4),
xlab="dif.perm",main =
```

```
paste("Histogram of" , "difference in means for the first
scenario"))
abline(v=S1.diff, lwd=2, col="purple")
mean(abs(S1.diff.perm) > abs(S1.diff))
#
x11()
hist(S2.diff.perm,xlim =range(-0.4,+0.4),
xlab="dif.perm",main =
paste("Histogram of" , "difference in means for the second
secnario"))
abline(v=S2.diff, lwd=2, col="purple")
mean(abs(S2.diff.perm) > abs(S2.diff))
########################## THE END ################################
```

7.2 Example 7.

```
result=rep(NA,5)
n=seq(10,50,by=10)
#
for (j in 1:length(n)){
  sum=0
  for (i in 0:n[j]){
        sum0=abs(dpois(i, 5, log = FALSE)-
        dbinom(i, n[j], 5/n[j], log = FALSE))
        sum=sum+sum0/2
  }
  result[j]=sum
}
result


db <- rep(NA, n)
dp <- rep(NA, n)
for (i in 0:n){
        dp[i]=dpois(i, 5, log = FALSE)
      db[i]=dbinom(i, n, 5/n, log = FALSE)
```

```
}
max(abs(dp-db))
############################## THE END ###################################
```

CHAPTER 8

Appendix II

8.1 Scripts related to the results in sections 4.4 and 4.5.

```
rm(list=ls())
install.packages("glm2")
# vcd is the required package to apply odds-ratio on
#contingency tables
install.packages("vcd")
library(MASS)
library(glm2)
library(vcd)
library(grid)
#########################################################################
# Results here are related to the Criteria for
# the validation of surrogate endpoints.
# Function Prentice takes dataset and check
# for the first, second, third and fourth
# Prentice's criteria for a given dataset.
#########################################################################
Prentice=function(d)
{
dtable <- table(d$survind, d$pfsind, d$treat,
dnn = c("survind", "pfsind", "treat"))
#
#Cheking the first Prentice's criterion
C=margin.table(dtable, cbind(2,3))
#n(S,Z), xtable marginalized on T
oddsC=oddsratio(C, log=FALSE)
confintC=confint(oddsC)
#summary(C)
#or chisq.test(C),test of independence between S and Z is
#not significant
#
modC <- glm(pfsind ~  treat , data = d,
family = "binomial")
#summary(modC)
```

```
#modC$coefficients[2]
#exp(modC$coefficients[2])=oddsratio(C)
#beta coefficient for treatment in a glm model.
#######################################################################
#Cheking the 2nd Prentice's criterion
F=margin.table(dtable, cbind(1,3))
#n(T,Z), dtable marginalized on S
oddsF=oddsratio(F,log=FALSE)
confintF=confint(oddsF)
#summary(F)
#test of independence between T and Z is not significant
#
modF <- glm(survind ~  treat , data = d, family = "binomial")
#summary(modF)
#modF$coefficients[2]
#exp(modF$coefficients[2])=oddsratio(F)
#######################################################################
#Cheking the 3rd Prentice's criterion
D=margin.table(dtable, cbind(1,2))
#n(T,S), dtable marginalized on Z
oddsD=oddsratio(D, log=FALSE)
confintD=confint(oddsD)
#summary(D)
#test of independence between T and S is not significant
#
modD<- glm(survind ~  pfsind , data = d, family = "binomial")
#summary(modD)
#chisq.test(margin.table(dtable, cbind(1,2)))
#######################################################################
#Cheking the 4th Prentice's criterion
#First check to see if we could drop the interaction term
#(treat*pfsind) from the logistic model
mod1 <- glm(survind ~ pfsind + treat +treat*pfsind
, data = d, family = "binomial")
#summary(mod1)
mod2 <- glm(survind ~ pfsind + treat , data = d,
family = "binomial")
#summary(mod2)
```

```
#mod2$coefficients[3]
anov12=anova(mod2,mod1, test="Chisq")
#
#Check if we could drop the term involving
#treatment while we ajusted for pfsind
mod11 <- glm(survind ~ pfsind + treat  , data = d,
family = "binomial")
#summary(mod11)
mod22 <- glm(survind ~ pfsind  , data = d, family = "binomial")
#summary(mod22)
anov1122=anova(mod22,mod11, test="Chisq")
#
#p(T|S,Z)=p(T|S) is equivalent to show conditional
#independence of p(T,Z|S)=p(T|S)p(Z|S),
#to show it, we need to show that the odds ratio
#for a 2*2 table at each level of S is equal to one.
X1=ftable(dtable[,1,])                    #S=0
#summary(as.table(X1))
#Gives the odds ratio for a table of T and Z at S=0,
oddsX1=oddsratio(X1, log=FALSE)
summaryX1=summary(oddsX1)
confintX1=confint(oddsX1)
#
X2=ftable(dtable[,2,])                    #S=1
#summary(as.table(X2))
#Gives the odds ratio for a table of T and Z at S=1
oddsX2=oddsratio(X2, log=FALSE)
summaryX2=summary(oddsX2)
confintX2=confint(oddsX2)

return(list(dtable=ftable(dtable),tableC=ftable(C),
oddsC=oddsC,confintC=confintC,
statisticC=summary(C)$statistic,
p.valueC=summary(C)$p.value,
modC=summary(modC),tableF=ftable(F),
oddsF= oddsF, confintF=confintF,
statisticF=summary(F)$statistic,
p.valueF=summary(F)$p.value,modF=summary(modF),
```

```
tableD=ftable(D),
oddsD= oddsD ,confintD=confintD,
statisticD= summary(D)$statistic,
p.valueD= summary(D)$p.value, modD=summary(modD),
anov12=anov12, anov1122=anov1122,oddsX1=summaryX1,
confintX1=confintX1 ,oddsX2=summaryX2 ,
confintX2=confintX2))
}
Prentice(dfx)
Prentice(dfy)
Prentice(dfz)
########################## THE END ##############################
```

8.2 Scripts related to the results in section 4.7

```
#rm(list=ls())
install.packages('gtools')   #to run permutation
library(gtools)
#?permute
#########################################################################
# Results here are related to the An entropy-based
# nonparametric test for the validation of surrogate
# endpoints paper. # "nonparam" is a function that take
# dataset, number of permutations and degrees of freedom
#(df), the returns W observed ($\hat{d_{KL}}$),
# asymptotic test statistics, chi_square value from the
# chi-square table, P-value permutation and P-value
# asymptotic.
#########################################################################
nonparam=function(d,L=1000,df=2)
{
#
Wper=rep(0,L)     #Ingredient for W.per
#
sum=0
r1=0
```

```
r2=0
r3=0
#
dtable = table(d$survind,d$pfsind,d$treat)
A = dtable                                  #n(T,S,Z)
B=as.vector(margin.table(dtable,2))    #n(S)
#n(S,Z), dtable marginalized on T
C=ftable(margin.table(dtable, cbind(2,3)))
#n(T,S), dtable marginalized on Z
D=ftable(margin.table(dtable, cbind(1,2)))
E=as.vector(margin.table(dtable,3))          #n(Z)
#n(T,Z), dtable marginalized on S
F=ftable(margin.table(dtable, cbind(1,3)))
G=as.vector(margin.table(dtable,1))          #n(T)
#
#
for (i in 1:2)
 {
for (j in 1:2)
{
for (k in 1:2)
{
sum=sum+A[i,j,k]*log((A[i,j,k]*B[j])/
(C[j,k]*D[i,j]))#Ingredient for W
r1=r1+A[i,j,k]*log((A[i,j,k]*E[k])/(C[j,k]*F[i,k]))
#Ingredient for npLRF
r2=r2+A[i,j,k]*log((A[i,j,k]*dim(d)[1])/
(C[j,k]*G[i]))#Ingredient for npLRF
}
# Ingredient for npPIG
r3=r3+D[i,j]*log((D[i,j]*dim(d)[1])/(B[j]*G[i]))
 }
 }
Wobs=sum/dim(d)[1]      #Gives W.obs
Wstatobs=2*dim(d)[1]*Wobs  #W.stat.obs
chi2table=qchisq(0.95, 2, ncp = 0,
lower.tail = TRUE, log.p = FALSE)
#
```

```
npLRFobs=(1-exp(-2*r1/dim(d)[1]))/
(1-exp(-2*r2/dim(d)[1])) #Gives npLRF.obs
#
npPIGobs=r3/r2       #Gives npPIG.obs
#####################################################################
#Calculate P-value for W using permutation
#
for(m in 1:L)
{
sumper=0              #Ingredient for W.per
#####################################################################
#Permutation ingredients on W
#Splitting d to d.sub1 where S=1 and d.sub2 where S=0
dsub1 <- subset(d,d$pfsind==1)
dsub2 <- subset(d,d$pfsind==0)
dnew=rbind(dsub1,dsub2)
#
#Permute treatment column in d.sub1 and name it treat.per1
treatper1=permute(dsub1$treat)
#
#Permute treatment column in d.sub2 and name it treat.per2
treatper2=permute(dsub2$treat)
#
treatper=c(treatper1,treatper2)
#
#merge treat.per with dfxnew
dnew=cbind(dnew,treatper)
#
tableper = table(dnew$survind,dnew$pfsind,dnew$treatper)
Aper =tableper                    #n(T,S,Z)
Bper=as.vector(margin.table(tableper,2)) #n(S)
#n(S,Z), table.per marginalized on T
Cper=ftable(margin.table(tableper, cbind(2,3)))
#n(T,S), table.per marginalized on Z
Dper=ftable(margin.table(tableper, cbind(1,2)))
Eper=as.vector(margin.table(tableper,3))   #n(Z)
#n(T,Z), table.per marginalized on S
Fper=ftable(margin.table(tableper, cbind(1,3)))
```

```
Gper=as.vector(margin.table(tableper,1))      #n(T)
#########################################################################
for (i in 1:2)
 {
for (j in 1:2)
{
for (k in 1:2)
{
sumper=sumper+Aper[i,j,k]*log(((Aper[i,j,k]*Bper[j])
+0.000001)/(Cper[j,k]*Dper[i,j]))
 }
 }
 }
Wper[m]=sumper/dim(d)[1]      #W Permutation
}
P_val_perm=(1+sum(Wper >= Wobs))/(1+L) #P_value Permutation
#
#P-value Asymptotic
p_val_asympt=pchisq(Wstatobs, df, ncp = 0,
lower.tail = FALSE, log.p = FALSE)
#########################################################################
################ model based LRF and PIG ##########################
#########################################################################
# Computing Ajusted likelihood reduction factor (LRF)
# for the first trial considering model without
# interaction term
#########################################################################
#GLM model Z on T
mod1 <- glm(survind ~  treat , data = d, family = "binomial")
#
#GLM model Z and S on T
mod2 <- glm(survind ~ pfsind + treat , data = d,
family = "binomial")
#
#Gives LRT(Z+S:Z)=Residual deviance Z on T - Res. Dev. S,Z
# on T
AA=anova(mod1,mod2, test="Chisq")$Deviance[2]
#
```

```
#Gives LRT(Z+S:1)= Null deviance S,Z on T - Res. Dev. S,Z
# on T
BB=mod2$null.deviance-mod2$deviance
#
#Gives LRF based on the formula
LRF1=(1-exp(-AA/dim(d)[1]))/(1-exp(-BB/dim(d)[1]))
#####################################################################
# Computing proportion of information gain (PIG)
# for the first trial
# considering model with interaction effect.
#####################################################################
#GLM model S on T
mod3 <- glm(survind ~  pfsind , data = d, family = "binomial")
#
#Gives LRT(S:1)= Null deviance S on T - Res. Dev. S on T
CC=mod3$null.deviance-mod3$deviance
#
#Gives PIG based on the formula
PIG1=CC/BB
#####################################################################
# Computing Ajusted likelihood reduction factor (LRF)
# for the first trial
# considering model with interaction term
#####################################################################
#GLM model S, Z on T with interaction effect between S and Z
mod4 <- glm(survind ~ pfsind + treat +treat*pfsind  ,
 data = d, family = "binomial")
#
#Gives LRT(Z+S:Z)=Residual deviance Z on T - Res. Dev. S,Z
# on T
DD=anova(mod1,mod4, test="Chisq")$Deviance[2]
#
#Gives LRT(Z+S:1)= Null deviance S,Z on T - Res. Dev. S,Z
# on T
EE=mod4$null.deviance-mod4$deviance
#
#Gives LRF based on the formula
LRF2=(1-exp(-DD/dim(d)[1]))/(1-exp(-EE/dim(d)[1]))
```

```
#####################################################################
# Computing proportion of information gain (PIG) for the
# first trial considering model without interaction effect.
#####################################################################
#Gives PIG based on the formula
PIG2=CC/EE
#####################################################################
return(invisible(list(Wobs=Wobs, Wstatobs=Wstatobs,
chi2table=chi2table, P_val_perm=P_val_perm,
p_val_asympt=p_val_asympt, npLRFobs=npLRFobs,
npPIGobs=npPIGobs, LRF1=LRF1, LRF2=LRF2, PIG1=PIG1, PIG2=PIG2)))
}
#####################################################################
fdfx=nonparam(dfx)
fdfy=nonparam(dfy)
fdfz=nonparam(dfz)
#
c(W.obs=fdfx$Wobs, W.stat.obs=fdfx$Wstatobs,
chi2.table=fdfx$chi2table,
P_val_perm=fdfx$P_val_perm, p_val_asympt=fdfx$p_val_asympt,
npLRF.obs=fdfx$npLRFobs, npPIG.obs=fdfx$npPIGobs,
LRF1=fdfx$LRF1,
LRF2=fdfx$LRF2, PIG1=fdfx$PIG1, PIG2=fdfx$PIG2)
#
c(W.obs=fdfy$Wobs, W.stat.obs=fdfy$Wstatobs,
chi2.table=fdfy$chi2table,
P_val_perm=fdfy$P_val_perm, p_val_asympt=fdfy$p_val_asympt,
npLRF.obs=fdfy$npLRFobs, npPIG.obs=fdfy$npPIGobs,
LRF1=fdfy$LRF1,
LRF2=fdfy$LRF2, PIG1=fdfy$PIG1, PIG2=fdfy$PIG2)
#
c(W.obs=fdfz$Wobs, W.stat.obs=fdfz$Wstatobs,
chi2.table=fdfz$chi2table,
P_val_perm=fdfz$P_val_perm, p_val_asympt=fdfz$p_val_asympt,
npLRF.obs=fdfz$npLRFobs, npPIG.obs=fdfz$npPIGobs,
LRF1=fdfz$LRF1, LRF2=fdfz$LRF2, PIG1=fdfz$PIG1,
PIG2=fdfz$PIG2)
########################## THE END ##################################
```

CHAPTER 9

Appendix III

9.1 Scripts related to the results in sections 5.3.1.

```
#####################################################################
# Function CI.b takes the estimation of $\beta_S$ and its
# standard error, then returns the $(1-2\alpha)$
# individuals confidence intervals for $\beta_S$.
#####################################################################
CI.b=function(b,std.b)
{
z.90=qnorm(0.90, mean = 0, sd = 1, lower.tail = TRUE,
 log.p = FALSE)
U.b=b+z.90*std.b
L.b=b-z.90*std.b
return(c(L.b,U.b))
}
#####################################################################

#####################################################################
# Function bonf.CI takes the estimation of $\beta_S$, its
# standard error, the estimation of $delta$ and its standard
# error, then returns the $(1-2\alpha)$ Bonferroni
# confidence intervals for $\beta_S$ and $delta$.
#####################################################################
bonf.CI=function(b,std.b,d,std.d)
{
        z.95=qnorm(0.95, mean = 0, sd = 1, lower.tail = TRUE,
         log.p = FALSE)
        U.d=d+z.95*std.d
        L.d=d-z.95*std.d
        #
        U.b=b+z.95*std.b
        L.b=b-z.95*std.b
        #
        return(list(c(L.d,U.d),c(L.b,U.b)))
}
########################## THE END ###################################
```

Bibliography

Agresti, A. (2013). *Categorical Data Analysis, 3rd edition.* New York: Wiley.

AIDSinfo (2017). Hiv/aids glossary. `https://aidsinfo.nih.gov/understanding-hiv-aids/glossary/840/clinical-endpoint`.

Allan, T. and R. Cribbie (2013). Evaluating the equivalence of, or difference between, psychological treatments: An exploration of recent intervention studies. *Canadian Journal of Behavioral Science 45*, 320–328.

Alonso, A., G. Molenbergh, T. Burzykowski, D. Renard, H. Geys, Z. Shkedy, F. Tibaldi, J. Abrahantes, and M. Buyse (2004). Prentice's approach and the meta-analytic paradigm: a reflection on the role of statistics in the evaluation of surrogate endpoints. *Biometrics. 60*, 724–728.

Aronson, J. K. (2005). Biomarkers and surrogate endpoints. *Br J Clin Pharmacol*, 491–494.

Bender, R. (2001). Adjusting for multiple testingwhen and how? *Journal of Clinical Epidemiology. 54*, 343–1349.

Berry, K. J., J. E. Johnston2, and P. W. Mielke (2011). Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics 3*, 527–542.

Beyene, J. and R. Moineddin (2005). Methods for confidence interval estimation of a ratio parameter with application to location quotients. *BMC Medical Research Methodology 5*, 32.

Bickel, P. and W. Van Zwet (1987). Asymptotic expansion for the power of distribution-free tests in the two-sample problem. *annals of Statistics 6*, 987–1004.

Bishop, C. (2006). *Pattern Rrecognition and machine learning.* New York: Springer-Verlag.

Blackwelder, W. (1982). Proving the null hypothesis in clinical trails. *Psychological Bulletin 3*, 345–353.

Bristol, D. (1993). Probabilities and sample sizes for the two one-sided tests procedure. *Communications in Statistics-Theory & Methods 22*, 1953–1961.

Burzykowski, T., G. Molenberghs, and M. Buyse (2005). *The Evaluation of Surrogate Endpoints.* New York: Springer-Verlag.

Bushe, C. J., M. Taylor, and J. Haukka (2010). Review: Mortality in schizophrenia: a measurable clinical endpoint. *Journal of Psychopharmacology*, 17–25.

Buyse, M. and G. Molenberghs (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics 54*, 1014–1029.

Buyse, M., G. Molenberghs, T. Burzykowski, D. Renard, and H. Geys (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 49–67.

Buyse, M., G. Molenberghs, X. Paoletti, K. Oba, A. Alonso, W. Van der Elst, and T. Burzykowsk (2016). Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Statistics in Medicine 58*, 104–132.

Casella, G. and R. L. Berger (2002). *Statistical Inference (Second ed.)*. CA, USA: Thomson Learning.

Choi, S., S. W. Lagakos, R. T. Schooley, and P. A. Volberding (1993). Cd4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic hiv infection taking zidovudine. *Annals of Internal Medicine 118*, 674–680.

Chow, S., J. Shao, and H. Wang (2002). A note on sample size calculation for mean comparisons based on noncentral t-statistics. *Journal of Biopharmaceutical Statistics 12*, 441–456.

Chow, S. and H. Wang (2001). On sample size calculation in bioequivalence trials. *Journal of Pharmacokinetics and Pharmacodynamics 28*, 155–169.

Chu, S. and J. Liu (2008). *Design and Analysis of Bioavailability and Bioequivalence Studies,3rd edition.* Boca Raton: Chapman & Hall/CRC.

ClinicalTrials.gov (2017). Learn about clinical studies. `https://clinicaltrials.gov/ct2/about-studies/learn#ClinicalTrials`.

de Jong, P. and G. Z. Heller (2008). *Generalized linear models for insurance data.* Cambridge: Cambridge University Press.

Deza, M. and E. Deza (2016). *Encyclopedia of distances, 4th edition*. Berlin Heidelberg: Springer-Verlag.

Diletti, E., D. Hauschke, and V. Steinijans (1991). Sample size determination for bioequivalence assessment by means of confidence intervals. *International Journal of Clinical Pharmacology, Therapy and Toxicology 29*, 1–8.

Downing, G. J. (2000). Nih definitions working group. biomarkers and surrogate endpoints in clinical research: definitions and conceptual model. *Biomarkers and Surrogate Endpoints. Amsterdam: Elsevier*, 1–9.

Efron, B. and R. Tibshirani (1994). *An Introduction to the Bootstrap.* Boca Raton: Chapman & Hall/CRC.

Ellenberg, S. S. and J. M. Hamilton (1989). Surrogate endpoints in clinical trials: cancer. *Statistics in Medicine. 18*, 405–413.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters 27*, 861–874.

Fieller, E. (1940). The biological standardization of insulin. *Royal Statistical Society 7*, 1–64.

Fleming, T. and D. DeMets (1996). Surrogate end points in clinical trials: are we being misled? *Ann Intern Med. 125*, 605–613.

Fleming, T. R. (1992). Evaluating therapeutic interventions: Some issues and experiences (with discussion). *Statistical Science 7*, 428–456.

Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications, 2nd Edition.* NY, USA: John Wiley.

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

Fuhlbrigge, A., T. Bengtsson, S. Peterson, A. Jauhiainen, G. Eriksson, C. Da Silva, A. Johnson, T. Sethi, N. Locantore, R. Tal-Singer, and M. Fageras (2017). A novel endpoint for exacerbations in asthma to accelerate clinical development: A post-hoc analysis of randomised controlled trials. *Lancet Respir Med*. doi:10.1177/1948550617697177.

Games, P. (1977). An improved t table for simultaneous control on g contrasts. *Journal of the American Statistical Association 72*, 531–534.

Gellad, W. F. and A. S. Kesselheim (2003). Accelerated approval and expensive drugs: A challenging combination. *N Engl J Med. 376*, 2001–2004.

Global Asthma Network (2014). *Global Asthma Report, Auckland, New Zealand*.

Goshtasby, A. (2012). *Image Registration. Principles, tools and methods*. London: Springer-Verlag.

Hauschke, D., V. Steinijans, and I. Pigeot (2007). *Bioequivalence Studies in Drug Development: Applications*. Chichester: John Wiley & Sons, Ltd.

Herson, J. (1975). Feiller's theorem versus the delta method for significance intervals for ratios. *Journal of Statistical Computing and Simulation. 3*, 265–274.

Hill, A. (1965). The environment and disease: association or causation? *Proc R Soc Med*, 295–300.

Katz, R. (2004). Biomarkers and surrogate markers: An fda perspective. *NeuroRx*, 189–195.

Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika. 70*, 163–173.

Kullback, S. and R. Leibler (1951). Annals of mathematical statistics. *N Engl J Med 22*, 79–86.

Lakens, D. (2017a). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *PsyArXiv. March 4*. doi:10.1177/1948550617697177.

Lakens, D. (2017b). *TOSTER: Two One-Sided Tests (TOST) Equivalence Testing*. R package version 0.3.2 — For new features, see the 'Changelog' file (in the package source).

Legator, M. S. and D. L. Morris (2003). What did sir bradford hill really say? *Arch Environ Health.*, 718–720.

Lin, D., T. Fleming, and V. De Gruttola (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine. 16*, 1515–1527.

Liu, J. and S. Chow (1992). Sample size determination for the two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics 20*, 101–104.

MacKay, D. J. C. (2005). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

McCune, B., J. Grace, and D. Urban (2002). *Analysis of ecological communities.* Gleneden Beach, Oregon: MjM Software Design.

Meyners, M. (2012). Equivalence testsa review. *Food Quality and Preference 26*, 231–245.

Miao, X., Y. Wang, and A. Gangopadhyay (2012). An entropy-based nonparametric test for the validation of surrogate endpoints. *Statistics in Medicine 31*, 1517–1530.

Muller-Cohrs, J. (1990). The power of the anderson-haucks test and the double t-test. *Biometrical Journal 32*, 259–266.

Neaton, J. D., D. N. Wentworth, F. Rhame, C. Hogan, D. I. Abrams, and L. Deyton (1994). Considerations in choice of a clinical endpoint for aids clinical trials. *Statistics in Medicine*, 2107–2125.

O'Byrne, P., H. Bisgaard, P. Godard, M. Pistolesi, M. Palmqvist, Y. Zhu, T. Ekstrm, and E. Bateman (2005). Budesonide/formoterol combination therapy as both maintenance and reliever medication in asthma. *Am J Respir Crit Care Med 171*, 129–136.

Ofori-Asenso, R. and A. Adom Agyeman (2015). Understanding cross over and parallel group studies in drug research. *Precision Medicine 1*, 1–4.

Pesarin, F. (2001). *Multivariate Permutation Tests: With Applications in Biostatistics (1st Edition).* New York: John Wiley and Sons Ltd.

Phillips, K. (1990). Power of the two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics 18*, 137–144.

Piantadosi, S. (2005). *Clinical Trials: A Methodologic Perspective, Second Edition.* NJ, USA: John Wiley & Sons, Inc.

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

Qu, Y. and M. Case (2007). Quantifying the effect of the surrogate marker. *Biometrics. 63*, 958–963.

Rabe, K., E. Pizzichini, B. Stllberg, S. Romero, A. Balanzat, T. Atienza, P. Lier, and C. Jorup (2006). Budesonide/formoterol in a single inhaler for maintenance and relief in mild-to-moderate asthma: a randomized, double-blind trial. *Chest 129*, 246–256.

Reddel, H., D. Taylor, E. Bateman, L. Boulet, H. Boushey, W. Busse, T. Casale, P. Chanez, P. Enright, P. Gibson, J. de Jongste, H. Kerstjens, S. Lazarus, M. Levy, P. OByrne, M. Partridge, I. Pavord, M. Sears, P. Sterk, S. Stoloff, S. Sullivan, S. Szefler, M. Thomas, and S. Wenzel (2009). An official american thoracic society/european respiratory society statement: Asthma control and exacerbations. *American Journal of Respiratory and Critical Care Medicine 180*, 59–99.

Rogers, J., K. Howard, and J. Vessey (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin 113*, 553–565.

Schuirmann, D. (1981). On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics 37*, 617.

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

Scicchitano, R., R. Aalbers, D. Ukena, A. Manjra, L. Fouquert, S. Centanni, L. Boulet, I. Naya, and C. Hultquist (2004). Efficacy and safety of budesonide/formoterol single inhaler therapy versus a higher dose of budesonide in moderate to severe asthma. *CCurr Med Res Opin. 20*, 1403–1418.

Sidak, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association 62*, 626–644.

Siqueira, A., A. Whitehead, S. Todd, and M. Lucini (2005). Comparison of sample size formula for 2×2 cross-over designs applied to bioequivalence studies. *Pharmaceutical Statistics 4*, 233–243.

Spackman, K. A. (1989). Signal detection theory: Valuable tools for evaluating inductive learning. in: Proc. sixth internat. *Workshop on Machine Learning. Morgan Kaufman, San Mateo, CA*, 160–163.

Steffens, F. (1971). On confidence sets for the ratio of two normal means. *South African Statistical Journal 5*, 105–113.

Tilaki, K. H. (2013). Receiver operating characteristic (roc) curve analysis

for medical diagnostic test evaluation. *Caspian J Intern Med 4*, 627–635.

Troughton, R. W., C. M. Frampton, T. G. Yandle, E. A. Espine, M. G. Nicholls, and A. M. Richards (2000). Treatment of heart failure guided by plasma aminoterminal brain natriuretic peptide (n-bnp) concentrations. *The Lancet*, 1126–1130.

U.S. Department of Health & Human Services (2000). *Action Against Asthma: A Strategic Plan for the Department of Health and Human Services*.

U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER) (2014). Guidance for industry and fda staff qualification process for drug development tools. `https://www.fda.gov/downloads/drugs/guidances/ucm230597.pdf`.

Verrills, N., J. Irwin, X. He, L. Wood, H. Powell, J. Simpson, V. McDonald, A. Sim, and P. Gibson (2011). Identification of novel diagnostic biomarkers for asthma and chronic obstructive pulmonary disease. *Am J Respir Crit Care Med. 183*, 1633–43.

Volberding, P. A., S. W. Lagakos, M. A. Koch, C. Pettinelli, J. A. Bartlett, M. S. Hirsch, R. L. Murphy, W. D. Hardy, R. Soeiro, M. A. Fischl, J. G. Bartlett, T. C. Merigan, N. E. Hyslop, D. D. Richman, F. T. Valentine, L. Corey, the AIDS Clinical Trials Group of the National Institute of Allergy, and I. Diseases. (1990). Zidovudine in asymptomatic human immunodeficiency virus infection a controlled trial in persons with fewer than 500 cd4-positive cells per cubic millimeter. *N Engl J Med 322*, 941–949.

Walker, E. and A. S. Nowacki (2011). Understanding equivalence and noninferiority testing. *J Gen Intern Med 26*, 192–196.

Wang, H. and S. Chow (2002). On statistical power for average bioequivalence testing under replicated crossover designs. *Journal of Biopharmaceutical Statistics 12*, 295–309.

Wellek, S. (2002). *Testing Statistical Hypotheses of Equivalence*. Boca Raton: Chapman & Hall/CRC.

Westfall, P. H. and S. Young (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: John Wiley

and Sons Ltd.

Westlake, W. (1981). Response to t.b.l. kirkwood: Bioequivalence testinga need to rethink. *Biometrics 3*, 589–594.

Zerbe, G. (1978). On fiellers theorem and the general linear model. *The American Statistician 32*, 103–105.

Zissler, U., J. Esser-von Bieren, C. Jakwerth, A. Chaker, and C. Schmidt-Weber (2016). Current and future biomarkers in allergic asthma. *Allergy 171*, 475–94.