Developing Study Design and Analysis Methods for Evaluating Predictive Biomarker Clinical Utility and Reproducibility

by

Henok G. Woldu

(Under the Direction of Kevin K. Dobbin)

Abstract

Statistical methods for evaluating predictive biomarkers' clinical utility, reproducibility and sample size for specific study designs have been slow to develop in response to the surge of need. In this dissertation, we propose three statistical methodologies: one is develop a metric Ψ_B using Bayesian decision theoretic framework; the second is propose a sample size estimation method (SWIRL); and the third is develop a reproducibility metric Δ_r.

A metric Ψ_B which measures the decrease in the expected event rate as a result of predictive biomarker guided therapy is proposed using Bayesian decision theoretic framework for a count clinical end point. Since Phase II data are usually small, maximum likelihood based estimates are biased and inefficient. This new metric, which also incorporates clinician inputs in the form of a prior however, is informative in making a go-no-go decision and the study design to choose for Phase III studies. Using toy simulation and a simulation conducted to mimic asthma clinical trial study, we show the robustness of the method under different scenarios.

Sample size estimation methods that match the study design and the metric under con-

sideration are key in predictive biomarker clinical utility evaluation process. In this dissertation we propose a sample size estimation method, Squared Width Inversion Regression Line (SWIRL). The SWIRL method is used to estimate a sample size n such that the 95% confidence interval width of the metric under consideration is smaller than a user defined length (W_{targ}). This is the first sample size method developed for estimating this target predictive parameter.

During assay development and validation processes, an original clinically validated assay is required to be modified for a number of different reasons. However, such modification invalidates the previous biomarker-outcome association studies and would force researcher to re-run the previous studies under the modified biomarker. This is time consuming and expensive. Here, we propose a reproducibility metric Δ_r which measures the impact of assay modification on patient outcome. A combination of both novel equations and simulations were used to estimate Δ_r and the associated 95% confidence interval.

INDEX WORDS: Decision Theory, Bayesian statistics, Hamiltonian Monte Carlo, Sample size, Reproducibility, Asymptotic distribution, Predictive biomarkers

Developing Study Design and Analysis Methods for Evaluating Predictive Biomarker Clinical Utility and Reproducibility

by

Henok G. Woldu

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2018

Developing Study Design and Analysis Methods for Evaluating Predictive Biomarker Clinical Utility and Reproducibility

by

HENOK G. WOLDU

Major Professor:	Kevin K. Dobbin
Committee:	Stephen L. Rathbun
	Xiao Song
	David M. DeJoy

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2018

Developing Study Design and Analysis Methods for Evaluating Predictive Biomarker Clinical Utility and Reproducibility

Henok G. Woldu

July 25, 2018

DEDICATION

To my late parents whom I have not got a chance to know well.

Acknowledgments

Foremost, I want to offer this endeavor to the **Almighty God** for the wisdom, strength, peace of mind and good health he bestowed upon me to finish my thesis research.

Without the support and encouragement of several special people, finalizing this thesis would not have been possible. As such, I would like to take this opportunity to express my sincere gratitude to those who assisted me in many different ways.

I would like to express my heartfelt appreciation to my advisor **Dr. Kevin K. Dobbin**, for his insightful guidance, constant encouragement and expertise in this area which lead to the successful completion of this thesis. He was always prepared to sit, listen and steer me in the right direction. Because of his willingness to offer me so much of his time and intellect, I managed to complete my Ph.D thesis. Thank you so much Dr. Dobbin.

I would also like to thank all my thesis committee members, **Dr. Stephen L. Rathbun**, **Dr. Xiao Song** and **Dr. David M. Dejoy** for serving as my committee members on one hand and supporting me financially through your grant projects on the other hand. Also, I really appreciate that you let my defense be an enjoyable moment by providing your brilliant comments and suggestions. Once again thank you so much.

Thank you so much my dear sister **Tsigay G. Woldu**. No one but only you took the responsibility to raise me since my childhood. You have always provided me with unconditional love, and support. Without you I could not be where I am today. I also want to thank my grandparents, uncles, aunts and friends for believing in me and being there for practical support in all those things of life beyond doing a PhD.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Swift advancement in genome sequencing in the last few years is transitioning the "one size fits all" treatment model to the patient specific treatment model. In this transition stage biomarkers in general and predictive biomarkers in particular have been playing key roles. Predictive biomarker are biomarkers used to identify a subgroup of patients who are most likely to benefit from a given treatment. Treatment negative side effects and costs, however, are avoided for the rest. However, even though the discovery of such biomarkers has been of enormous interest in recent years, development of statistical methods to design studies and assess their clinical utility and reproducibility have not kept pace.

Biomarker by treatment interaction test is still a common statistical method used for evaluating predictive biomarker performance (Buyse, 2007, Taube et al., 2009, Freidlin et al., 2010, Tajik et al., 2013). However, it has been shown to be inadequate method for evaluating predictive biomarker clinical utility (Janes et al., 2011, Huang et al., 2012). Other frequentist graphical methods and metrics have been suggested as alternatives in the past decade or so (Song and Pepe, 2004, Gunter et al., 2007, Brinkley et al., 2010, Janes et al., 2011; 2014a). Detailed review of these methods is in Chapter 2. All these methods however, assume the drug was already approved and data are collected retrospectively to assess the biomarker

clinical utility. Ideally one often wants biomarker clinical utility evaluation to be done at the end of phase II clinical trial and make a go-no-go decision about whether the biomarker can be used in phase III study designs.

However, at the end of phase II, data collected are usually small and result in maximum likelihood based metrics which are biased and inefficient (Casella and Berger, 2002, Lehmann and Romano, 2006). On the other hand, it becomes customary to incorporate clinicians', biomarker scientists' and other experts' knowledge about the biomarker performance in the evaluation process. Further, drug negative side effects and monetary costs associated with the treatment need to be quantified since they affect patients' decision making. The first part of this dissertation is aimed at developing a metric Ψ_B used for assessing clinical utility of a predictive biomarker at the end of phase II study using Bayesian decision theoretic framework. This metric measures the decrease in the expected event rate as a result of predictive biomarker guided treatment. We focus here on a clinical trial with a count primary endpoint motivated by a phase II asthma clinical trial study.

During biomarker clinical utility evaluation process, developing a sample size estimation method that corresponds to the study design and the metric under consideration is key. Under a slightly different scenario and assumptions for a binary clinical endpoint, Janes et. al developed the metric Θ (Janes et al., 2014a). This metric measures the decrease in the proportion of unfavorable outcomes under biomarker guided treatment. However, there is no sample size estimation method developed for this metric. In Chapter 4 of this dissertation, we first develop novel alternative equations to estimate Θ and propose a sample size estimation method, Squared Width Inversion Regression Line (SWIRL). SWIRL methods are used to estimate a sample size n so that the 95% confidence interval mean width for Θ is less than a user defined target width (W_{targ}). R program is made available for its implementation.

A typical biomarker development and validation process follows three stages which are often interdependent (Ball et al., 2010). Stage one is about the technical development and

assay analytical validation process (Swanson, 2002). Stage two (middle stage) is qualification process to assess evidence of association between the biomarker and the outcome of interest (Williams et al., 2006, Koulman et al., 2009). The last stage focuses on clinical validation of the biomarker which is often a contextual analysis (Williams et al., 2006, Ball et al., 2010). For a number of different reasons initially promising biomarkers are required to be modified in the middle stage before moving to the final stage of clinical validation. However, such modification invalidates any previous studies of biomarker-outcome association and forces researchers to re-run the investigation under the modified biomarker. This process however is costly and time consuming and leaves many promising biomarkers in a dead end. In Chapter 5 of this dissertation, we develop a metric Δ_r which is used to directly estimate the impact of the modified biomarker on the metric of interest (Θ) using reproducibililty study.

In this dissertation, we aimed to propose three statistical methodologies: one is developing a metric Ψ_B using Bayesian decision theoretic framework; the second is proposing a sample size estimation method SWIRL; and the third is developing a reproducibility metric Δ_r.

1.1 Motivating examples

1.1.1 Phase II asthma clinical trial study

Asthma is a chronic inflammatory disease of the airways with marked heterogeneity in the clinical course and in response to treatment (Bel, 2004, Wenzel, 2006, Siddiqui and Brightling, 2007). Despite treatment with inhaled corticosteroids (ICSs), and other controller medications, a substantial proportion of patients continue to have uncontrolled asthma (Bateman et al., 2004, Corren et al., 2011). Consequently, part of the current high unmet medical need in asthma is uncontrolled disease that persists despite conventional treatment with guidelines-based standard-of-care therapy, which includes ICS therapy plus a second controller medication (Hanania et al., 2015). The Phase II clinical development plan to develop

drug AA was designed to test the efficacy and safety of AA in this patient population with uncontrolled asthma who have high unmet medical need. The primary end point of the study was the rate of asthma exacerabations over 52 weeks. Asthma exacerbation is defined as new or increased asthma symptoms that led to treatment with systematic corticosteroids or to hospitalization.

With small a data set collected from phase II study and prior information from the experts, the object is to assess the clinical utility of the predictive biomarker BMK and make a go-no-go decision on the study design for phase III. Existing statistical methods (Byar, 1985, Song and Pepe, 2004, Buyse, 2007, Janes et al., 2014a) have drawbacks: (1) they are designed for a binary clinical end point and (2) they are all frequentist methods and do not include prior information in the analysis. To overcome this gap we propose a statistical method for assessing the clinical utility of the predictive biomarker BMK using Bayesian decision theoretic framework where the clinical endpoint is a count. Even though demonstration of the method is done here using the Asthma clinical trial study, it is applicable for a more general context.

1.1.2 Oncotype DX predictive biomarker

Oncotype DX test is a genomic test that analyzes the activity of a group of 21 genes from a breast cancer tissue sample that can affect how a cancer is likely to behave and respond to treatment. Most early-stage, estrogen-receptor-positive, HER2-negative breast cancers that haven't spread to the lymph nodes are considered to be at low risk for recurrence. After surgery, hormonal therapies such as tamoxifen are prescribed to reduce the risk that the cancer will come back in the future. Whether or not chemotherapy also is necessary has been an area of uncertainty for patients and their doctors, especially for women with cancer that had spread to just one, two, or three lymph nodes. The Oncotype DX test was designed to offer more information to help women and their doctors make decisions about

chemotherapy. The Oncotype DX test results assign a Recurrence Score, a number between 0 and 100. A score less than 18 indicates that the cancer has a low risk of recurrence. That is the benefit of chemotherapy is likely to be smaller and will not outweigh the risks of side effects. If the score is ≥ 31 the cancer has a high risk of recurrence, and the benefits of chemotherapy are likely to be greater than the risks of side effects (Karapetis et al., 2008, Gluz et al., 2016).

A metric Θ which measures the decrease in an unfavorable event rate under marker guided treatment has been widely advocated as a global predictive biomarker clinical utility measure (Gunter et al., 2007, Song and Pepe, 2004, Janes et al., 2011, Brinkley et al., 2010, Janes et al., 2014a). However, there is no sample size calculation method that can be used to guide a study design for evaluating the biomarker clinical utility performance. To close this gap, we first developed novel equations and algorithms to get an estimate for Θ and proposed the SWIRL sample size estimation method.

1.1.3 Ki67 reproducibility study

An initially clinically validated biomarker X may be modified to biomarker W for a number of different reasons. This however, invalidates previous study results of the biomarker clinical utility performance. In breast cancer research, for example, the marker Ki67 has a potential use for prognosis, prediction and response monitoring (Dowsett et al., 2011, Goldhirsch et al., 2011, Yerushalmi et al., 2010, Viale et al., 2008). Ki67 is a nuclear proliferation marker used to determine the growth fraction of a given cell population. Despite the apparent utility of Ki67, it has been less used due to the lack of reproducibility in measuring it (Harris et al., 2007). To set a standard guide for Ki67 analysis, Polley et al. (2013) conducted a reproducibility study. A total of one hundred breast cancer cases where measured in eight different labs and each had a score for Ki67 recorded.

In the study published by Polley et al. (2013), intraclass correlation coefficient (ICC) was

used as a measure of reproducibility. However, if the objective is to assess the biomarker clinical utility, existing reproducibililty metrics alone can not be used for two main reasons. First, a high value of ICC between two biomarkers does not guarantee that the clinical utility of the two biomarkers will be the same when assessed using the metric Θ. This claim is shown to be true from our simulation studies as depicted in Figure 2.2. Second, assessing the clinical utility of the marker using Θ, we need to observe the outcome Y as well. So, when the biomarker is modified to W, to directly evaluate the clinical utility performance of W one has to wait to observe the outcome. However, waiting to observe the outcome Y under the modified biomarker W is costly and time consuming. This problem stymies the development of many initially encouraging biomarkers. To solve this, we propose a new reproducibility metric Δ_r which measures the difference in Θ when the modified biomarker W is observed instead of X without the need to wait to observe the outcome Y.

1.2 Research questions and objectives

The phase II asthma clinical trial is a typical example where with a small information about the biomarker one needs to make a decision before moving to phase III. The oncotype DX biomarker also exemplify a situation where the metric Θ can be used to evaluate the clinical utility of the biomarker but without the appropriate sample size needed for it. In this dissertation, we aimed to propose three statistical methodologies. Specifically we propose a metric Ψ_B using Bayesian decision theoretic framework to address the problems listed in the asthma phase II clinical trial study. To deal with the problems of sample size estimation that correspond the metric Θ we propose the SWILR sample size estimation method.

The rest of this dissertation is organized as follows. In chapter 2, we provide the literature review for predictive biomarker clinical utility, reproducibility and sample size estimation statistical methods. Chapter 3 introduce the proposed predictive biomarker utility evaluation

metric Ψ_B using Bayesian decision theoretic framework and its application using a simulation study done to mimic the phase II asthma clinical trial study. Our proposed sample size estimation method SWIRL is presented in Chapter 4. In Chapter 5 we study the proposed reproducibility metric Δ_r and provide the R program for its implementation. Summarizing the findings and outlining the future research work, this dissertation concludes in Chapter 6.

Chapter 2

Literature Review

The phase II asthma clinical trial poses an important question. With a small data set and prior information from expert, how do we proceed to evaluate the predictive biomarker's clinical utility performance and recommend a study design for phase III. On the other hand, when one wants to design a study to evaluate the clinical utility performance such as that of the Oncotype DX, what sample size should be used to guarantee us enough power? The Ki67 study, also poses an important research question regarding reproducibility study. A brief review of the literature regarding statistical methods for evaluating clinical utility and reproducibility of predicitive biomarkers and sample size estimation methods associated will be given in this chapter.

2.1 Log Linear Models and Bayesian Methods

2.1.1 Poisson Distribution

The number of occurrences of an event during a fixed time period is modeled using count data. Count data are encountered commonly in medical and public health research studies and frequentist statistical methods are typically used to make estimations and inferences

(Du et al., 2011, Lu et al., 2014). Poisson distribution is popular for modeling count data under three key assumptions: (a) the probability of an event is proportional to the length of the interval, (b) the number of events between non-overlapping intervals is independent and (c) for a given small sub-interval, either only one event occurs or no events occur at all. For a given outcome Y with a mean μ, the probability mass function of the Poisson(μ) distribution is:

$$P(y = y|\mu) = \frac{e^{-\mu}\mu^{y}}{y!} \qquad y = 0, 1, 2, ... \tag{2.1}$$

such that

$$\begin{aligned} \mathbb{E}(Y) &= \mu \\ \mathbb{V}(Y) &= \mu \end{aligned} \tag{2.2}$$

2.1.2 Regression Models for Count Outcome

The Poisson regression model for count outcome is a special case of the generalized linear models (GLM) as detailed in (Nelder and Baker, 1972, McCullagh, 1984). Let the n observations $y_1, y_2, ..., y_n$ be a realization from an independent Poisson random variable such that $Y_i \sim Pois(\mu_i)$. Further lets assume the mean (μ_i) depends on the covariates represented by a vector of $\boldsymbol{x_i}$ and is written as a simple linear model as, $\mu_i = \boldsymbol{x_i'\beta}$. Using generalized linear model with a log link function, the log linear model (Poisson regression model) is finally written as:

$$log(\mu_i) = \boldsymbol{x'\beta} \tag{2.3}$$

Estimation of the $\boldsymbol{\beta}$ coefficients is then done using principles of maximum likelihood by writing the likelihood function for the n independent Poisson observations as:

$$L(\boldsymbol{\beta}; \mathbf{Y}) = \prod_{i=1}^{n} \left\{ \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \right\} \tag{2.4}$$

Ignoring the constant involving $log(y_i!)$ and taking the logs , log-likelihood function is

$$log\ L(\boldsymbol{\beta}; \mathbf{Y}) = \sum_{i=1}^{n} \{y_i log(\mu_i) - \mu_i\} \tag{2.5}$$

Since μ_i is written as a function of $\boldsymbol{x_i}$ and parameters $\boldsymbol{\beta}$, to get the mle estimates $\hat{\boldsymbol{\beta}}$ we take the first derivatives of the log likelihood function of equation (2.5) and set them equal to zeros.

2.1.3 Overdispersion in Count Data

One common problem encountered when modeling count data is a phenomena called overdispersion. Overdispersion occurs when there are more zeros than expected from a an ordinary Poisson model (McCullagh, 1984, Jang et al., 2010). When this happens, the equality of mean and variance assumption of the Poisson regression model is violated and hence the inference that follows becomes invalid. Ideas for adjusting the probabilities of excess zeros in Poisson regression models date back to (Johnson and Kotz, 1969). The zero-inflated Poisson (ZIP) model (Lambert, 1992) and Poisson hurdle (PH) model (Mullahy, 1986) are commonly used as alternatives when there are excess zeros in the model. The observed excess zeros can be sampling zeros and/or structural zeros. Sampling zeros are zeros that are part of the Poisson distribution, and it is assumed these zeros are observed by chance. Whereas structural zeros arise due to a particular structure in the data set (Hu et al., 2011, Hua et al., 2014). The probability mass function of Y given π and μ can be written as:

$$Pr(Y = y|\pi, \mu) = \begin{cases} \pi + (1-\pi)e^{\mu} & \texttt{if y = 0} \\ (1-\pi)\frac{e^{-\mu}\ \mu^{y}}{y!} & \texttt{if y > 0} \end{cases} \tag{2.6}$$

where π represent the probability that the observed zero is from the zero-inflated stage and μ represent the mean for the Poisson count stage provided the observed value is not a zero inflated.

From equation (2.6) one can clearly see that, the ZIP regression model has two stages: the zero-inflation stage and Poisson count stage. Depending on the problem at hand, the covariates in a given data set can be used in both stages to estimate the parameters π and μ simultaneously. Commonly loglinear and logit models are used to relate the covariates with the parameters π and λ. To get the maximum likelihood estimators, similar to the standard Poisson regression model detailed above, one has to write the log-likelihood function of for equation (2.6), take the first derivatives with respect to the parameters and equate them to zero.

2.1.4 Bayesian Methods

Bayesian and frequentist methods overall fall within the same framework. Both approaches assume there is a population parameter θ that we want to make inference about and a likelihood distribution $f(y|\theta)$ that determines how likely is one to observe y given θ. The key difference is while θ is considered as fixed in frequentist methods, in Bayesian θ is treated as random with probability distribution $g(\theta)$.

Bayesian methods in additional to the ease of estimation allow us to incorporate prior information in the analysis by specifying a prior distribution to θ. Bayesian methods have increased in popularity since the past twenty years due to advances in the methodology,

notably the Markov Chain Monte Carlo and the computation power. In Bayesian methods, the focus is on the posterior distribution $\pi(\theta|y)$ which is the likelihood $\frac{f(y|\theta)}{f(y)}$ and the prior $g(\theta)$ product and is written as :

$$\pi(\theta|y) = \frac{f(y|\theta)g(\theta)}{f(y)} \tag{2.7}$$

where

$$f(y) = \int_{\theta} f(y|\theta)g(\theta)d\theta \tag{2.8}$$

and is called the marginal likelihood. The marginal likelihood does not depend on θ and is considered as a normalizing constant so that $\pi(\theta|y)$ is proper probability density (Winkelmann, 2008). Computing the denominator of equation (2.7) is difficult and becomes more intractable when θ is multivariate. However, when the joint posterior distribution and the prior distribution both are from the same family distribution, a closed form solution of equation (2.7) is possible. The is a situation which is commonly called conjugacy and such a prior is know as conjugate prior (Raiffa, 1974, DeGroot, 2005).

When the posterior distribution $\pi(\theta|y)$ of equation (2.7) is complex and a closed form solution does not exist, a stochastic simulations method such as Monte Carlo approaches is used. For a multivariate θ we get a joint posterior which is high dimension where generating independent samples becomes non trivial. But we can rather use Markov Chain Monte Carlo (MCMC) methods to draw dependent samples.

MCMC methods are a very powerful tool to approximate arbitrary probability distribution and their derivations without the need to know the normalization term. MCMC methods even though they have been around as long as the Monte Carlo techniques, their importance has been felt truly late (1990s) in the field of statistics (Gelfand and Smith, 1990, Andrieu et al., 2003, Robert and Casella, 2011, Geyer, 2011). In Bayesian analysis the

distribution of interest, i.e, the posterior distribution is often non-standard or so complex that we can not directly sample from it (Gilks et al., 1995). Therefore MCMC methods are used to draw samples from an alternative distribution and then accept certain samples while rejecting others in order to approximate the distribution of interest (Carlin and Louis, 2008, Klauenberg and Elster, 2016). This procedure usually produces samples that are dependent which are often called Markov chain. There are different MCMC methods of which the Metropolis-Hastings, Gibbs sampler and Hamiltonian Monte Carlo (HMC) are common. A brief history of the development of Meropolis-Hastings and Gibbs sampler algorithms can be found in (Tierney, 1994, Robert and Casella, 2011). The Metropolis-Hastings algorithm, for example, starts by first choosing a proposal distribution q. Typically this distribution is chosen in such a way that it is easy to sample from directly and is a great approximation for the distribution of interest f. In pseudocode, the Metropolis-Hastings algorithm can be written as (Robert, 2004):

Algorithm 1 Metropolis-Hasting algorithm

1: select the proposal distribution q

2: choose initial value X_0

3: for i=0,1,.....

- sample point Y from $q(./X_i)$
- Take $X_{i+1} = Y$ with probability $\alpha(X_i, Y)$
- $X_{i+1} = X_i$ otherwise

where $\alpha(X_i, Y) = \min(1, \frac{f(Y)q(X_i/Y)}{f(X_i)q(Y/X_i)})$

Another commonly used MCMC method is the Gibbs sampler which has gained a surge of popularity with Geman and Geman paper (1984) where they used it for image processing models (Geman and Geman, 1987, Casella and George, 1992, Molenaar, 1997). Gibbs sam-

pler generated a random variable from a marginal distribution indirectly without needing to calculate the density using full conditional distributions that are often from the known statistical distributions (Spiegelhalter et al., 1996, Molenaar, 1997, Lynch, 2007).

Metropolis-Hasting algorithm and Gibbs Sampler, though among the commonly used MCMC methods, have some difficulty when implementing them. One notable issue with Metropolis-Hastings algorithm is the choice of a proposal density as the success or failure of the algorithm depends on it. If the proposal is too narrow for example only the mode of the target distribution might be visited and on the other hand , choosing a wide proposal density would result in higher rate of rejection and hence high correlation (Andrieu et al., 2003, Carlin and Louis, 2008). As Gibbs samples needs a full conditional distribution specification, it is not always straightforward to obtain proper conditional densities and attain convergence(Casella and George, 1992). Instead one can use Hamiltonian Monte Carlo which is faster and can generate less correlated samples (Shahbaba et al., 2014, Brooks et al., 2011).

2.2 Predictive Biomarkers

Biomarkers that predict treatment efficacy hold great potential for improving clinical outcomes and decreasing medical costs. Treament selection biomarkers are sometimes called "predictive"(Sargent and Allegra, 2002, Simon and Maitournam, 2004, Simon, 2008) or "prescriptive" (Gunter et al., 2007) markers. If a predictive biomarker can identify which patients are likely to benefit from a treatment, assignment of the treatment can be limited to this subgroup of patients. Such an approach will prevent the remaining group of patients the needless and potentially toxic and costly therapy (Janes et al., 2015).

The Oncotype DX recurrence score, for example, is used to identify a subgroup of women who are unlikely to benefit from Chemotherapy following breast cancer surgery (Harris et al., 2016). Similarly, KRAS status is used to identify colorectal cancer patients likely to benefit

from Epidermal growth factor receptor (EGFR) inhibitor treatment (Amado et al., 2008, Mehta et al., 2010). From Figure 1, we can see that patients with a lower biomarker value will be better off if they avoid treatment assuming the outcome is measuring probability of unfavorable outcome. However, study designs and evaluation methods that assess the efficacy of the predictive biomarker need to be developed first, before, the biomarker is used in a clinical decision making.

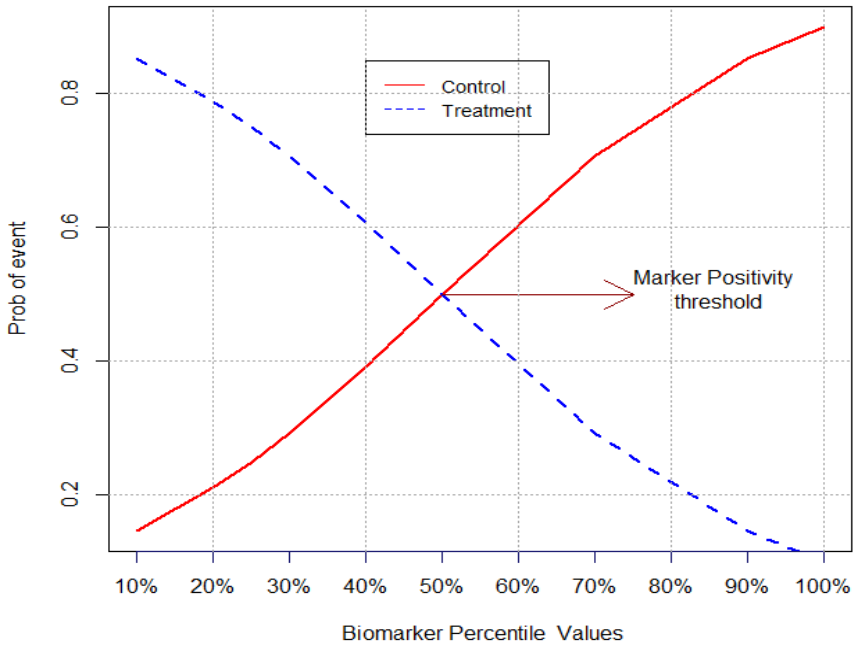

Figure 2.1: Shows the probability of a bad outcome under a given treatment. The marker positivity threshold is the point where treatment need to be switched. According this picture, subjects with smaller biomarker values should be advised not take treatment. This is a theoretical biomarker.

2.3 Basic Notations

Let A be a given treatment, $A = 1$ if a subject is assigned to a treatment group and $A = 0$ if assigned to a placebo or standard of care (soc) group. Further lets assume the primary clinical endpoint is binary and is denoted by Y, such that, $Y = 1$ represents unfavorable outcome

(death or recurrence of a disease) and $Y = 0$ otherwise. Additionally let, X represent a predictive biomarker which is measured at baseline for each subject. The biomarker X can be categorical or continuous. However, throughout this dissertation we will assume X is continuous and has a known probability density function given by $f(x)$.

The natural approach to represent the relationship between the outcome Y and the covariates (A and X) along the interaction term ($A * X$) is using multiple logistic regression model as:

$$Ln\left[\frac{Pr(Y=1|A,X)}{1-Pr(Y=1|A,X)}\right] = \beta_0 + \beta_1 X + \beta_2 A + \beta_3 AX. \tag{2.9}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ are the logistic model parameters.

Evaluation of predictive biomarker's clinical utility is often times done by testing the null hypothesis of no biomarker by treatment interaction (Byar, 1985, Buyse, 2007, Taube et al., 2009, Freidlin et al., 2010, Tajik et al., 2013). From equation (2.1), this will be to test :

$$H_o : \beta_3 = 0 \qquad H_a : \beta_3 \neq 0 \tag{2.10}$$

However, the interaction test though a necessary condition, fall short of being a sufficient condition for evaluating biomarker's clinical utility (Janes et al., 2011, Huang et al., 2012). Two predictive biomarkers X_1 and X_2 fitted using equation (2.1) can have the same β_3 but perform differently. The scale of β_3 also depends on the functional formal of the model under consideration and measurement unit of the biomarker. This makes biomarker comparison even more challenging (Huang et al., 2012). There are also settings where the biomarker by treatment interaction is not significant but the biomarker can be instructive (Song and Pepe, 2004). Additionally, interaction test being an indirect measure, is not easily understood by non-statistician and make conveying the message to broader audience difficult.

Song et. al (2004) proposed a graphical display, the selection impact(SI) curve which can

be estimated using parametric and nonparametric methods. The curve shows the proportion of people who respond to a given treatment as a function of a given predictive biomarker based treatment selection criteria. Lets assume the treatment policy is given in such a way that a patient is treated if $X > c$ and not treated if $X \leq c$, where X is the patient's biomarker value. The proportion of people who respond to a given treatment is then calculated as :

$$\begin{aligned} \theta &= P[Y = 1|(X > c \ \textit{and} \ A = 1) \ \texttt{or} \ (X < c \ \texttt{and} \ A = 0)] \\ &= P[Y^1 = 1|X > c]P[X > c] + P[Y^0 = 1|X < c]P[X < c] \end{aligned} \tag{2.11}$$

where Y^1 and Y^0 represent the case of observing the outcome when $A = 1$ or $A = 0$ respectively. The value of θ obtained from equation (2.11) above measures the proportion of patients who show the outcome of interest when the treatment policy is implemented as per the biomarker based criteria.

Brinkley et al. (2010) developed a generalized estimator of attributable benefit for an optimal treatment regime. First an algorithm was defined to assign treatment to patients according to their bimarker value in such a way that;

$$g_{opt}(x) = I\{P(Y = 1|A = 1, X) - P(Y = 1|A = 0, X) < 0\} \tag{2.12}$$

Based on equation (2.12), an individual with baseline biomarker value would be assigned to $A = 1$ if $P(Y = 1|A = 1, X) < P(Y = 1|A = 0, X)$, else assign the individual patient to $A = 0$. The I in equation (2.12) indicates a binary indicator for the treatment assignment. After determining the optimal treatment regime, the attributable benefit for a given $g_{opt}(x)$ is then obtained as:

$$AB_{opt} = 1 - \frac{P\{Y^*(g_{opt}(x)) = 1\}}{P(Y = 1)} \tag{2.13}$$

where $Y^*(g(x))$ represent the potential outcome and $P(Y = 1)$ denotes the current default treatment. AB_{opt} measures the the proportion events that could have been avoided had we used the optimal treatment as in equation (2.12) to assign treatment to individuals.

Using a similar potential outcome framework, Huang et al. (2012), developed an optimal rule for classifying individuals to the available treatment options and methods for evaluating continuous treatment selection markers. Let $D = Y(0) - Y(1)$, represent an individual patient's treatment benefit, where $Y(0)$ and $Y(1)$ are the potential outcomes associated with a subject being assigned to no treatment and treatment respectively. When being on treatment does not make a difference, $D = 0$, otherwise $D = 1$ when $Y(0) > Y(1)$. Using a Bayes' theorem, for a given biomarker value X, the ratio of risks in the no treatment over the treatment is shown to be:

$$\frac{P(X|D=0)}{P(X|D=1)} = \frac{P(D=0|X)P(D=1)}{P(D=1|X)P(D=0)} = \frac{P(D=1|X)}{1-P(D=1|X)}\frac{P(D=1)}{P(D=0)} \tag{2.14}$$

Under a monotone treatment effect assumption from equation (2.4),an optimal X can be found such that it maximizes the classification accuracy. After obtaining the optimal X, Huang et al. (2012) proposed a constrained maximum likelihood method to estimate the parameters in the risk model they defined which is similar to the one in equation (2.1). However, despite the progresses made in developing statistical methods for predictive biomarker evaluation, there was not a unified framework to achieve the intended purpose.

Janes et al. (2014a) proposed a comprehensive statistical framework for predictive biomarker evaluation. Their framework which included both descriptive summary measures and inferential methods is a comprehensive tool for individual biomarker evaluation and candidate biomarker comparisons. The descriptive summary measures of (Janes et al., 2014a) are related to the sub population treatment effect pattern plot of (Bonetti and Gelber, 2004) and others (Royston and Sauerbrei, 2004, Cai et al., 2010). However, Janes et al. (2014a) use

percentile scaling of the biomarker to make comparison of candidate markers easy.

The comprehenesive biomarker evaluation metric proposed by Janes et al. (2014a) measure the decrease in the expected event rate resulting from marker guided treatment and is represented as Θ. This metric is closely related to those previously proposed by (Song and Pepe, 2004, Gunter et al., 2007, Janes et al., 2011, Qian and Murphy, 2011, Brinkley et al., 2010, Janes et al., 2014a). Following similar procedures as those of Song and Pepe (2004), Brinkley et al. (2010), Huang et al. (2012), for a given biomarker X the absolute treatment effect is given as: $\Delta(X) = P(Y = 1|A = 0, X) - P(Y = 1|A = 1, X)$. The treatment rule is then set in such a way that if $\Delta(X) < 0$ the subject would be assigned to treatment and to placebo otherwise. Subject with $\Delta(X) < 0$ are referred as marker negative and those with $\Delta(X) >$ as marker positive. For a given specific treatment rule ($\Delta(X)$), average benefit among the marker negative is given as:

$$\begin{aligned} B_{neg} &= P(Y = 1|\ A = 1, \Delta(X) < 0) - P(Y = 1|\ A = 0, \Delta(X) < 0) \\ &= E(-\Delta(X)|\ \Delta(X) < 0) \end{aligned} \tag{2.15}$$

and proportion of subjects who can forego treatment is written as:

$$P_{neg} = P(A = 0) \tag{2.16}$$

The metric Θ which measures the decrease in the event rate that resulted from marker guided treatment is then calculated as:

$$\begin{aligned} \Theta &= P(Y = 1|A = 1) - [P(Y = 1|\ A = 1, \Delta(X) > 0)P(\Delta(X) > 0) + \\ &\quad P(Y = 1|\ A = 0, \Delta(X) < 0)P(\Delta(X) < 0)] \\ &= [P(Y = 1|\ A = 1, \Delta(X) < 0) - P(Y = 1|\ A = 0, \Delta(X) < 0)]P(\Delta(X) < 0) \end{aligned} \tag{2.17}$$

$$= \quad B_{neg} \ * \ P_{neg}$$

Janes et al. (2014a) then used empirical and model based methods to get an estimate of $\hat{\Theta}$ and which is given by:

$$\begin{aligned} \hat{\Theta}^e &= \hat{B}^e_{neg} \ * \ \hat{P}_{neg} \\ &= E(-\hat{\Delta}(X)|\ \hat{\Delta}(X) < 0) \ * \ E(-\hat{\Delta}(X)|\ \hat{\Delta}(X) > 0) \end{aligned} \tag{2.18}$$

and

$$\begin{aligned} \hat{\Theta}^m &= \hat{B}^m_{neg} \ * \ \hat{P}_{neg} \\ &= \int (-\hat{\Delta}(X)\ I[\hat{\Delta}(X) < 0)]\ d\hat{F}_{\Delta} \end{aligned} \tag{2.19}$$

such that $\hat{P}_{neg}$ is the proportion of marker negative subjects and $\hat{F}_{\Delta}$ is the CDF of $\Delta(X)$.

2.4 Sample size for predictive biomarker study design

Parallel to developing a metric for the purpose of evaluating a predictive biomarker the task of determining a sample size n is equally important. Even though the estimation of Θ starts with a logistic regression as given in equation (2.1), the functional form used to get the final estimate for Θ is different. Therefore, existing sample size estimation methods used for logistic regression (Whittemore, 1981, Hsieh, 1989, Demidenko, 2007; 2008) can not be used directly. For a multiple logistic regression with an interaction term, similar to equation (2.1), Demidenko (2008) suggested a sample size calculation formula for testing, $H_o : \beta_3 = 0$ vs $\quad H_1 : \ \beta_3 \neq 0$ as;

$$n = \frac{(Z_{1-\frac{\alpha}{2}} + Z_p)^2}{\beta_3^2}\ V \tag{2.20}$$

The major step in calculating sample size using equation (2.20) involves computing V in terms of the regression coefficients β_0 *to* β_3 given in equation (2.1). Let $A = e^{\beta_0}$, $B = e^{\beta_1}$, $G = e^{\beta_2}$ and $K = e^{\beta_3}$. Additionally let $p_x = Pr(X = 1|Z)$ and $p_z = Pr(z = 1)$. Assuming treatment assignment is dependent of the biomarker value, let

$$P(X = 1|Z) = \frac{e^{c+\delta z}}{1 + e^{c+\delta z}} \tag{2.21}$$

Given $D = e^{\delta}$ and $C = e^{c}$, we can further write

$$1 - p_x = \frac{p_Z}{1 + CD} \quad + \quad \frac{1 - p_z}{1 + C} \tag{2.22}$$

Letting $q = p_x(1 + D) \ + \ p_z(1 - D) \ - 1)$, we can write equation (2.16) in terms of C as

$$C = \frac{q + \sqrt{q^2 + 4p_x(1 - p_x)D}}{2(1 - p_x)D} \tag{2.23}$$

Finally an an estimate for V is obtained as; $V = \frac{1}{L} + \frac{1}{R} + \frac{1}{F} + \frac{1}{J}$. The quantities $L, R, FandJ$ are defined as follows:

$$\begin{aligned} L &= \frac{A(1 - p_z)}{(1 + A)^2(1 + C)} & F &= \frac{ABC(1 - p_z)}{(1 + AB)^2(1 + c)} \\ R &= \frac{ABCDGKp_z}{(1 + ABGK)^2(1 + CD)} & J &= \frac{AGp_z}{(1 + AG)^2(1 + CD)} \end{aligned} \tag{2.24}$$

This method of calculating the sample size needed to test the interaction coefficient needs nine parameters to be specified by the user. Additionally there are two main differences in the assumptions used from the logistic regression model of equation (2.1). One, this model assume a binary covariate for the biomarker not a continuous and second this model does not assume the independence of treatment assignment and biomarker value. Therefore, even if we want to calculate a sample size n that guarantee enough power to test $H_o : \beta_3 = 0$

vs $H_1 : \beta_3 \neq 0$, this will not serve the desired objective of calculating n for Θ. Sample size calculation for a quantitative variable and groups interaction for cox proportial hazard model was developed by Lachin (2013). This method however depends on weak asymptotic properties and does not work well often.

Janes et al. (2015) set four criteria choosing n so that biomarker clinical utility evaluation can be done from a given trial. These four criteria are:

Criterion 1 (Power for Interaction): for a given α-level of no treatment-biomarker interaction test the study to have a $1 - \beta_1$ power.

Criterion 2 (Detecting Improved Outcomes): This is to ensure that the lower bound (LB) of $(1 - \alpha_2)$ % CI for Θ lies about 0 with high probability,i.e,

$Pr(LB > 0|\Theta = \Theta_a, H_0\ rejected) \geq 1 - \beta_2$

Criterion 3 (Precision Estimation of Improved Outcomes: this is to ensure that we have enough power such that Θ is estimated with high precision.

$Pr(|\hat{\Theta} - \Theta_a| \leq \epsilon_1|\Theta = \Theta_a, H_0\ rejected) \geq 1 - \beta_4$ for a specified $\epsilon > 0$

Criterion 4 (Errors in Treatment Rule): this criteria is to make sure that the treatment effect among the marker negative is sufficiently small, i.e,

$X_t = arg\ max_{Y:\Delta(\hat{X})<0}\Delta(X), P(\Delta(X_t) < \epsilon_2|\Theta = \Theta_a, H_0\ rejected) \geq 1 - \beta_4$.

These criteria set by Janes et al. (2015) however, do not provide any direct sample size calculation formula or algorithm for the parameter of interest Θ. Additionally, these criteria are set under the assumption that biomarker evaluation is a secondary study objective rather than being primary and lead to a very large sample size. To fill this gap of sample size calculation for the metric Θ, in this dissertation we are proposing the Squared Width Inversion Regression Linear (SWIRL) method. The SWIRL method of sample size determination for Θ works in such a way that the 95% confidence interval width of of Θ is less than a user defined length (W_{targ}).

2.5 Reproducibility Metrics for Predictive Biomarkers

For a binary clinical endpoint, predictive biomarker clinical utility can be quantified strongly by using the metric Θ. For a given assay say X, Θ will define the reduction in the expected event rate that results from a biomarker guided treatment in comparison to the default (biomarker unguided) treatment. However, an initially validated assay like X is required to be modified say to an assay W for different reasons such as (1) reducing preparation cost, (2) migrating the assay platform, (3) simplify preparation methods and so on. In such an event, waiting to observe the outcome associated with the modified assay W is costly and time consuming as well. This hinders the discovery of many initially promising biomarker since clinical performance of the modified assay can not be achieved using existing reproducibility metrics.

One of the most widely used measure of reproducibility between two measurements is the product-moment correlation coefficient (ρ). Let the standard deviations associated with the original assay X and the modified assay W be σ_x and σ_w respectively. Then the correlation coefficient $\rho = \frac{cov(X,W)}{\sigma_x \sigma_w}$. However, ρ measures the strength of the association between two measurements rather than the agreement between them (Bland and Altman, 1986, Müller and Büttner, 1994). Two observations which seem to have poor agreement can produce a correlation coefficient which is high (Serfontein and Jaroszewicz, 1978).

Another method which is commonly misused to measure the reproducibility between two measurements is linear regression(Altman and Bland, 1983). Often testing the hypothesis for slope equal to one gives a misleading conclusion. This is similar to testing the correlation coefficient equal to zero since the $\hat{\beta}(slope) = r_{xw}\frac{s_x}{s_w}$, where r_{xw} is the correlation coefficient (X, W), s_x and s_w are the standard deviations for X and W respectively. A highly reproducible results could result in rejecting to the null (slope=1) due to small standard error and vice versa when the data is more scattered (Lawrence and Lin, 1989, Obuchowski et al.,

2015).

The Bland-Altman plot or sometimes called the difference plot, is another graphical method which is widely used to compare the agreement of two measurements (Bland and Altman, 1986; 1999). This method plots the difference between the two measurements against the average value of the measurements. In scenarios where one is a known "gold standard" method, the differences are plots against the gold standard (Krouwer, 2008). Using the Bland-Altman plot, lack of agreement is summarized by calculating the bias and standard deviation of the differences. Let $\bar{d}$ and s to be the mean and standard deviation of the differences between the two measurements. Assuming the differences follow a Gaussian distribution, 95% of the differences will lie within $\bar{d} \pm 1.96 * \frac{s}{\sqrt{n}}$. The upper and lower bound of the interval are referred as limits of agreement and a difference which falls within this interval would be considered clinically not significant.

In medical fields the concordance correlation coefficient (CCC) and intraclass correlation coefficient (ICC) are by far the most commonly used reproducibility metrics. The CCC was first developed by Lin (1989). For $i = 1, 2, ..., n$ pair of samples (X_{i1}, W_{i1}) which are sampled independently from a bivariate normal with means μ_x and μ_w and covariance matrix given by

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xw} \\ \sigma_{xw} & \sigma_w^2 \end{pmatrix} \tag{2.25}$$

The CCC between X and W is then calculated as the expected value of their squared difference as:

$$\begin{aligned} E[(X - W)^2] &= (\mu_x - \mu_w)^2 + (\sigma_x^2 + \sigma_w^2 - 2\sigma_{xw}) \\ &= (\mu_x - \mu_w)^2 + (\sigma_x - \sigma_w)^2 + 2(1 - \rho)\sigma_x\sigma_w \end{aligned} \tag{2.26}$$

One notable feature of of the CCC is it contains both measures of accuracy and precision which are two key characteristics of reproducibility. CCC does not only measure the strength of the association but also the degree of departure from the 45^o line. Even though the original CCC was developed to assess the repreducibility of two measurements only, later it was expanded by Chen and Barnhart (2008).

Assessment for reproducibility of measurements between labs, technicians or devices in bio medical research is also commonly done using intraclass correlation coefficient (ICC)(Bartko, 1966, DONNER, 1979, Gisev et al., 2013). The original ICC which began with the work of(Fisher, 1925) has been based on the one-way analysis of variance (ANOVA). From one-way anova study design, ICC can be calculated as

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2} \tag{2.27}$$

where σ_b^2 is the between subject variability and σ_e^2 is the within subject variability. This original design of ICC which is commonly referred as ICC_1 was further extended to ICC_2 and ICC_3 based on two-way ANOVA with and without interaction respectively (Bartko, 1966, Shrout and Fleiss, 1979, McGraw and Wong, 1996).

However, all the reproducibility metrics mentioned above; pearson correlation coefficient, regression line, Bland-Altman plot, CCC and ICC are not appropriate for the our purpose. The main point with all these being, though they can assess the reproducibility between X and W, none of them can evaluate how Θ value will change when the original assay is X is modified to W. High value of ICC between X and W can not be directly translated to mean W can replace X.

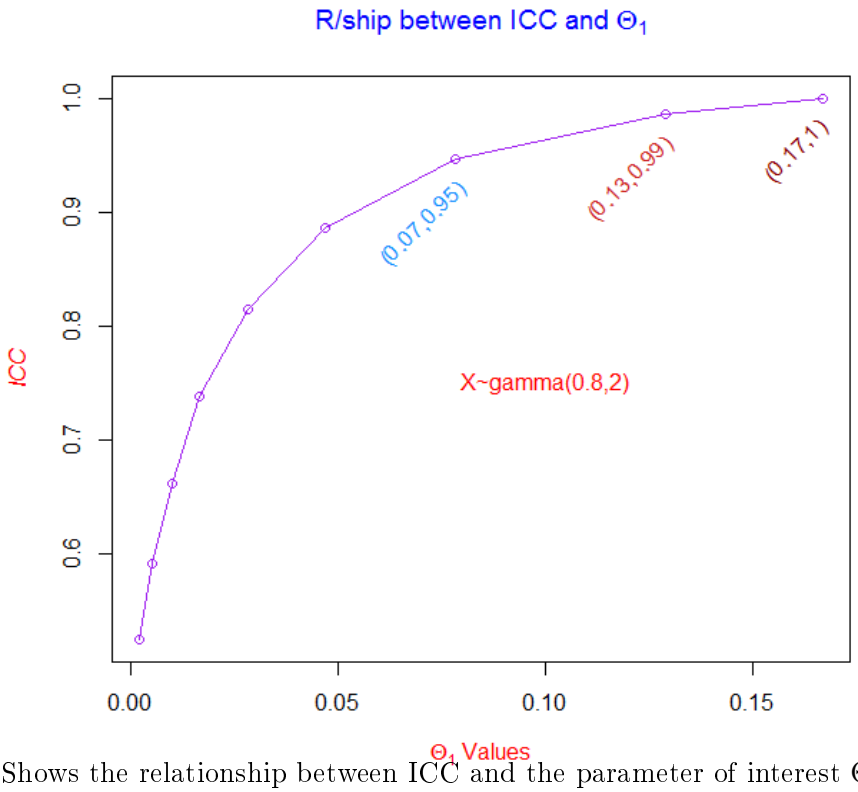

Figure 2.2: Shows the relationship between ICC and the parameter of interest Θ_1. X is assumed to be the gold standard biomarker and the modified assay is W, where $W = X + U$.The error term $U \sim \mathcal{N}(0, \sigma_e^2)$, where we considered different value of σ^2.

From Figure 2.2 above, it is clear that ICC alone can not capture the effect of the assay modification on Θ, which is our metric of biomarker clinical utility performance. Let Θ_x be a measure of the decrease in the proportion of event rate that resulted from biomarker guided treatment when the observed assay is X. Since the outcome Y associated with X is observed, Θ_x estimation can be done following Janes et al. (2014a) procedures or the modified equations we developed in this dissertation. However, since the outcome associated with the modified assay W is not observed, Θ_w can not be directly estimated. In this dissertation, we developed a reproducible metric Δ_r which captures the difference in Θ when the modified assay W is observed. Implementation of of our method is demostrated using the reproducibility data KI67. An R package (RMPB) which stands for Reproducibility Metric for Predictive Biomarerks is also made available.

Chapter 3

Bayesian decision theoretic framework for evaluating the clinical utility of a predictive biomarker with count endpoint

Henok G. Woldu and Kevin K. Dobbin

To be submitted to Journal of Statistics in Medicine.

Abstract

Maximum likelihood based estimators for evaluating the clinical utility of a predictive biomarker in early phase (I and II) clinical trials are biased and inefficient since they depend on large sample asymptotic properties. Further it is customary to include prior information about the biomarker performance, and costs associated with treatment as part of the analysis. This paper proposes a Bayesian decision theoretic framework for evaluating a predictive biomarker with a count end point. A metric Ψ_B which measures a decrease in the expected event rate as a result of marker guided treatment was developed and adjustment for zero inflated scenarios were studied. Bayesian credible interval was constructed to quantify uncertainty of the proposed metric Ψ_B. Toy simulation studies were first used to assess the robustness of this metric under different scenarios followed by a simulation done to mimic the phase II clinical trial conducted by Genentech to develop a drug for the treatment of Asthma.

Keywords : Decision theory; treatment selection biomarker; biomarker clinical utility; bayesian analysis

3.1 Introduction

Existing predictive biomarker evaluation methods are frequentist and retrospective in design (Song and Pepe, 2004, Brinkley et al., 2010, Janes et al., 2011). They assume the drug was already approved and data is collected thereafter to assess whether or not a particular biomarker could be utilized to guide treatment in the future. However, one would often want the biomarker evaluation to be done at the end of phase II and make a go-no-go decision whether or not to include the biomarker as part of the drug development plan. If the decision is in favor of including the biomarker, a subgroup of patients in Phase III would not receive the treatment because it does not benefit them (Janes et al., 2011, Baker and Kramer, 2015). This in return will result in better efficacy reports and higher approval chance for the drug under consideration.

Data available during phase II clinical trials are usually small and the resulting maximum likelihood based estimators are often biased and inefficient since they depend on large sample asymptotic properties (Casella and Berger, 2002, Lehmann and Romano, 2006). Further, the confidence intervals constructed from the maximum likelihood based methods are often misrepresented as probabilities that the unknown parameter of interest will be included in the interval, though this interpretation of the confidence interval aligns with the Bayesian view (Liu and Powers, 2012, Gelman et al., 2014). In this paper, a metric Ψ_B, used for quantifying clinical utility of a predictive biomarker during early phases of clinical trials is developed under Bayesian decision theoretic framework. Prior information about the biomarker performance and costs associated with a given treatment(negative side effect of the drug and monetary costs) are also included as part of the evaluation process.

The past few years has seen great hope and optimism in the shift from the one-size-fits-all treatment of a disease with drugs to a treatment which is restricted only to the subgroup of patients deemed to benefit from it. The task of identifying those likely to benefit from the

available treatment is mainly based on one or more biomarker measures obtained from each patient (Ru et al., 2011, Bossuyt and Parvin, 2015). These biomarkers commonly called treatment selection markers or predictive biomarkers (Janes et al., 2011, Baker and Kramer, 2015) in addition to helping patients get effective treatment, also help to minimize medical costs and improve the approval chance of a clinical drug in pipeline development. In an era where evidence-based decisions are predominant, a meaningful quantitative measure for the performance of a treatment selection marker is of paramount importance. Commonly, predictive biomarker evaluation involves testing for the relationship between the outcome and the marker by treatment interaction (Green, 1982, Yusuf et al., 1991, Buyse, 2007, Freidlin et al., 2010). However, statistical significance of an interaction test, is not a sufficient condition by itself to make a conclusion about the clinical utility of a biomarker (Janes et al., 2011, Huang et al., 2012).

Alternative assay performance metrics were developed in the last two decades such as the selective impact curve (Song and Pepe, 2004), Attributable Benefit (Brinkley et al., 2010), and the metric Θ (Janes et al., 2011). These metrics however, are derived retrospectively from phase III prospective Randomized Clinical Trials (RCTs) and don't take existing compelling evidence about the biomarker's clinical utility performance into consideration, nor does this paradigm fit as part of a drug development plan. Therefore, developing a metric that quantifies the clinical utility of an assay prior to phase III is of great importance. Having a metric to assess the assay performance prior to phase III, would help one to restrict treatment to patients who are likely to benefit from it during phase III randomization and produce better treatment efficacy results of the molecule under study.

Steve et al. (1978) introduced a decision analysis framework in clinical settings in the 1970s. Decision theory allows one to make use of both quantitative and qualitative inputs to make logically reproducible decisions (Ball et al., 2010). Decision analysis requires the person in charge of making the decision to break down the components of the decision into parts

and clearly specify the assumptions involved. The list of all available choices is then charted in to a decision tree. Decision tree is graphical representation for all the possible decisions made and their possible consequences. Finally, mathematical models are used to evaluate the likely outcomes of each choice. When using biomarker information to assign treatment, the decision one has to make is binary in nature, i.e, either to treat or not treat a subject. Using an appropriate mathematical model, one then can assess the risk associated with each decision and make the appropriate choice. The aforementioned predictive biomarker performance metrics all involve parameter estimation which is a decision theory problem since one has to select a particular value from a set of possible values. However, these methods do not take the prior belief or information one has about the performance of the assay under consideration.

A Bayesian decision theoretic framework, can help address many of the common concerns that arise in the process of evaluating the clinical utility of a predictive biomarker. Clinicians, assay developers and others who have expertise working with the molecule under consideration may have more information to provide about the assay performance than what is available in the data at hand. A key feature of the Bayesian approach is its ability to incorporate the subjective belief one has as part of the analysis in the form of a prior. Choosing a prior is daunting and there is not a straightforward procedure for this. However, if the existing information is used appropriately to construct the prior, the parameter estimates obtained from the Bayesian method are contained within a reasonable range of values and help the stability of MCMC algorithm used for Bayesian inference (Korner-Nievergelt et al., 2015, Brooks et al., 2011). In our setting, we bring into play the best scientific guess of the experts to construct our priors. A set of equations is developed first to change the clinician inputs to model parameters which were used as centers of the prior distributions considered. These beliefs are later updated using data collected during Phase II to develop the posterior metric used to assess the clinical utility of the predictive biomarker.

In this paper, we propose a Bayesian decision theoretic framework design for evaluating the clinical utility of a predictive biomarker prior to phase III and develop a metric used to quantify the net benefit of this biomarker. This metric (Ψ) which we termed Bayesian Biomarker Net Benefit (BBNB) measures the decrease in the expected unfavorable event rate as a result of biomarker guided treatment. This metric is closely related to the one developed by Janes et al. (2011) but permits wider range of loss functions and incorporates existing belief as prior during estimation. Even though, the development in this paper is done assuming a count endpoint, extensions to scenarios with continuous, binary or time to event could follow directly. In dealing with count data, overdispersion is a typical problem, overdispersion adjustment using the zero-inflated Poisson and negative binomial models were further considered.

3.2 Motivational Context

Asthma is a chronic inflammatory disease of the airways with marked heterogeneity in the clinical course and in response to treatment (Bel, 2004, Wenzel, 2006, Siddiqui and Brightling, 2007). Despite treatment with inhaled corticosteroids (ICSs), and other controller medications, a substantial proportion of patients continue to have uncontrolled asthma (Bateman et al., 2004, Corren et al., 2011). Consequently, part of the current high unmet medical need in asthma is uncontrolled disease that persists despite conventional treatment with guidelines-based standard-of-care therapy, which includes ICS therapy plus a second controller medication (Hanania et al., 2015). The Phase II clinical development plan to develop drug AA was designed to test the efficacy and safety of AA in this patient population with uncontrolled asthma who have high unmet medical needs. The primary end point of the study was the rate of asthma exacerabations over 52 weeks. Asthma exacerbation is defined as a new or increased asthma symptoms that led to treatment with systematic corticosteroids

or to hospitalization. Our objective here is to propose a statistical method for assessing the clinical utility of the predictive biomarker BMK using a Bayesian decision theoretic framework. Even though demonstration of the method is done here using the Athma clinical trial study, it is applicable in a more general context.

3.3 Settings and Notations

Let the outcome of interest be $\mathbf{Y}$, asthma exacerbation rate (average number of asthma exacerbations over a time period t) which is a count, such that, $\mathbf{Y} \in \{0, 1, 2, ...\}$. Further let the input variables be represented by a vector $\mathbf{x}$ such that $\mathbf{x} = (\mathbf{x_1}, \mathbf{A})$ where $\mathbf{x_1}$ denote the biomarker BMK level measure taken at baseline from each subject and $\mathbf{A}$ denote the treatment assignment such that $A = 1$ if the subject is assigned to active treatment and $A = 0$ otherwise. Additionally, $\theta \in \Theta$ will represent the parameter subspace that relates the outcome Y with the inputs $\mathbf{x}$. The natural approach to represent the relationship between the outcome Y and the vector of input $\mathbf{x}$ is through a log linear model, which can be written as,

$$ln\{E(Y_i|X_i, A_i)\} \quad = \quad \beta_0 + \ \beta_1 X_i + \beta_2 A_i + \beta_3 X_i A_i \tag{3.1}$$

such that $i = 1, 2, ...n$ where n represent the number of subjects in the study. To address the problem at hand through the principles of decision theory, one first needs to thoroughly define the three spaces which are core to decision theory (Robert, 2007, Berger, 2013). The true state of the world, the decision space and the consequence of a particular action. Let Θ represent the space for the true state of the world, $\mathcal{D}$ for the decision space and $\mathcal{R}$ for the consequence of the action. These three spaces together are linked by a loss $\mathcal{L}$ function which

is defined as:

$$\mathcal{L} : \Theta \times \mathcal{D} \mapsto \mathcal{R} \tag{3.2}$$

Clearly, a loss function $\mathcal{L}$ is a benchmark used for assessing a possible act $\delta \in \mathcal{D}$, for a given true state of the world $\theta \in \Theta$. The loss function then takes values in the space of consequences $\mathcal{R}$. The objective of the problem therefore is to minimize the loss $\mathcal{L}$, by choosing an optimal decision δ^*, which is defined as,

$$\delta^* := \underset{\delta \in \mathcal{D}}{\operatorname{argmin}}\, \mathbb{E}[L(\theta, \delta)] \tag{3.3}$$

In Bayesian context we pre-specify a prior distribution for $\theta \in \Theta$. Applying the principles of decision theory, let $\mathcal{L}(\theta, \delta)$ be the loss only due to the increase in the expected number of exacerbations in the asthma condition. If we let δ_T represent the decision made to assign all subjects to treatment, then the loss associated with this decision can be written as,

$$\begin{aligned} \mathcal{L}(\theta, \delta_T) &= E[Y|X_1, A = 1] \\ &= \mu_T(\mathbf{x}) \end{aligned} \tag{3.4}$$

$\mu_T(\mathbf{x})$ is the mean exacerbation rate to be calculated from the log linear model of equation (3.1) when subjects are assigned to treatment group and $\theta = (\beta_0, \beta_1, \beta_2, \beta_3)$. Similarly, loss function associated with the decision to assign all subjects to the placebo (standard of care), can be written as,

$$\begin{aligned} \mathcal{L}(\theta, \delta_P) &= E[Y|X_1, A = 0] \\ &= \mu_P(\mathbf{x}) \end{aligned} \tag{3.5}$$

where $\mu_P(\mathbf{x})$ is the mean exacerbation rate to be calculated from the log linear model of equation (1) when subjects are assigned to placebo. After clearly specifying the loss and decision functions, the Bayesian risk functions is finally defined as:

$$R_B(\pi(\theta), \delta) \quad = \quad \int_{\Theta} \int_{\mathcal{Y}} \mathcal{L}(\theta, \delta) p(\mathbf{y}|\theta) \pi(\theta) dy d\theta \tag{3.6}$$

where $\mathcal{L}(\theta, \delta)$ will be the expected loss, $p(\mathbf{y}|\theta)$ is the likelihood function of the long linear model in equation (3.1) with mean given by $\mu(\mathbf{x})$ and $\pi(\theta)$ is the prior distribution for the unknown states of the world.

3.4 Non-crossing Risk Curves

In many clinical trial settings, a simple endpoint like exacerbation rate does not fully capture the consequences associated with the treatment. Treatment related toxicity and cost incurred from the treatment play a key role in making a final decision when evaluating the clinical utility of a predictive biomarker. The aforementioned two factors and others have an important influence in making an optimal treatment decision and optimal decision about how to design the phase III trial. In the simplest case, all the treatment related costs can be assumed to be constant, say c, regardless of individual subject's biomarker values. However, this can further be extended in an event where there is enough evidence to suggest that treatment related toxicity is dependent on subject's biomarker value. Figure 3.1 demonstrates a scenario where treatment related costs was set to constant (c =0.25). The plot in the left shows risk curve of each treatment arm without taking the treatment related costs into account, but plot on the right show the same risk curves with treatment linked costs taking in to consideration. Looking at the left plot, one can conclude that no matter the biomarker value of the subject, one would decide to treat all the subjects. However, if we look at the

plot in the right which takes the drug negative side effect and cost in to consideration, the optimal decision would be to recommend treatment only for subject with a biomarker value above the 50th percentile value.

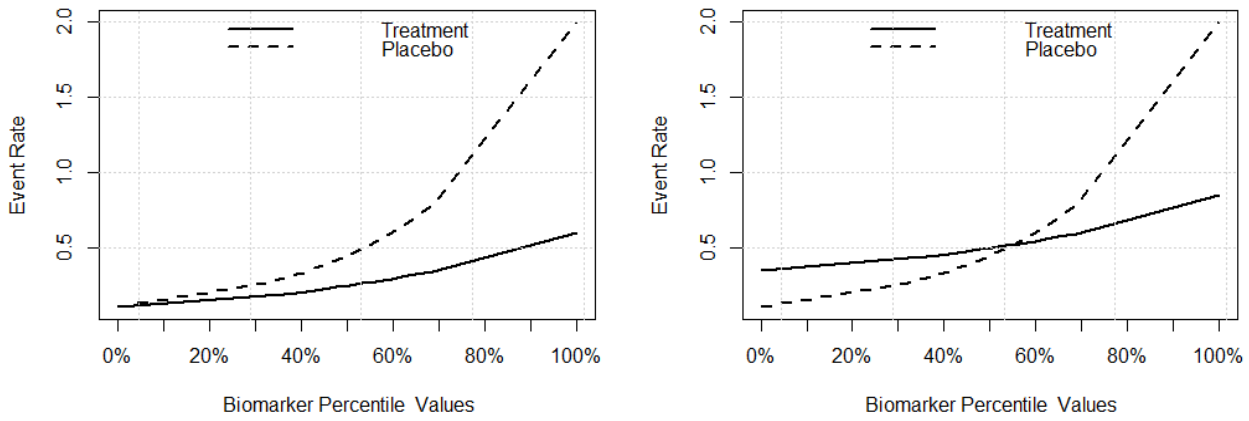

Figure 3.1: Risk curves for each treatment arm as a function of biomarker percentile value. The left-hand plot shows risk curves estimated without taking the treatment related cost into consideration while the right-hand plots takes the treated related cost into account. Treatment related cost was assumed to be constant regardless of the subjects biomarker value and was set to c=0.25. This cost was added only for those subjects assigned to the treatment group.

3.5 Optimal Treatment Decision Rule

Let the loss function $\mathcal{L}$ associated with the decision δ_T and the slope parameters of the log linear model β be written as,

$$\begin{aligned} \mathcal{L}(\beta, \delta_T) &= E[y] \\ &= E_x E[Y|x, A=1] \\ &= \int \{exp\{\beta_0 + \ \beta_1 x + \beta_2 A + \beta_3 x A\}\} f(x) dx \end{aligned} \tag{3.7}$$

where βs are obtained either using frequentis or Bayesian approach, $f(x)$ is the probability distribution of the biomarker X and the constant c is used to quantify the cost and negative side effect of the drug. Similarly the loss function associated with the decision δ_P can be written as,

$$\begin{aligned} \mathcal{L}(\beta, \delta_P) &= E[Y] \\ &= E_x E[Y|x, A=0] \\ &= \int \{exp\{\beta_0 + \beta_1 x\}\} f(x) dx \end{aligned} \tag{3.8}$$

From equations 3.8 and 3.9 above, one can see that, the loss associated with the decision to assign subjects to treatment group is equal to the loss associated with the decision to assign subjects to placebo group if and only if,

$$\{\beta_2 + \beta_3 x\} = 0$$

From equation(3.10), assuming $\beta_3 > 0$, a subject will be better off if he/she is assigned to treatment provided that $X < \frac{-\beta_2}{\beta_3}$ and to placebo if $X \geq \frac{-\beta_2}{\beta_3}$. when $\beta_3 < 0$ treatment assignment will be the reverse. Therefore, the optimal decision δ_{opt} which minimizes the expected loss is written as:

$$\begin{aligned} \mathcal{L}(\beta, \delta_{opt}) &= \int_{\mathcal{X}_1} \{exp\{\beta_0 + \beta_1 x + \beta_2 A + \beta_3 x A\}\} f(x) dx \\ &\quad + \int_{\mathcal{X}_0} \{exp\{\beta_0 + \beta_1 x\}\} f(x) dx \end{aligned} \tag{3.9}$$

where δ_{opt} is the optimal decision, $\mathcal{X}_1 \in \{X : X < \frac{-\beta_2}{\beta_3}\}$ and $\mathcal{X}_0 \in \{X : X \geq \frac{-\beta_2}{\hat{\beta}_3}\}$.

3.6 Biomarker Net Benefit (BNB)

Here we define a new metric, Biomarker Net Benefit (BNB), which will be represented by Ψ. The metric Ψ measures the decrease in expected event rate as a result of using an optimal decision rule δ_{opt} which has a value in the positive real line, i.e, $\Psi \in [0, \infty]$. If the expected loss function $L(\hat{\beta}, \delta)$ is estimated using a frequentist approach, we will call it Frequentist Biomarker Net Benefit (FBNB) and represent it by Ψ_F. However, when Bayesian approach is used to estimate the expected loss function, we will call it Bayesian Biomarker Net Benefit (BBNB) and represent it as Ψ_B. In this section we will lay out the steps for estimating Ψ_B. Depending on the default treatment, BBNB can be estimated as,

$$\hat{\Psi}_{B_P} = \mathcal{L}(\hat{\beta}_B, \delta_P) - \mathcal{L}(\hat{\beta}_B, \delta_{opt}) \tag{3.10}$$

when the default treatment is to assign all subjects to placebo (standard of care), and

$$\hat{\Psi}_{B_T} = \mathcal{L}(\hat{\beta}_B, \delta_T) - \mathcal{L}(\hat{\beta}_B, \delta_{opt}) \tag{3.11}$$

when the default treatment is to assign all subjects the current active treatment. The subscripts B, P and T represents for *Bayesian, Placebo* & *Treatment*. Here we will first demonstrate how to get the Bayesian estimates of the slope parameters $\hat{\beta}$s and then proceed to show the derivations for $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ respectively.

3.6.1 Bayesian For Count Data

Maximum likelihood (MLE) estimators are appealing and in common use as they have the necessary asymptotic properties. Such estimators are assumed to be consistent, efficient, and have a normal distribution as the sample size $n \to \infty$ (Casella and Berger, 2002, Lehmann and Romano, 2006). However, it not uncommon to see such assumptions being violated

because the sample size is small. Early phase (I and II) clinical trials, for example, are usually small. Further, maximum likelihood methods naturally quantify uncertainty of the MLE by constructing confidence intervals which are often misinterpreted as probabilities that the unknown parameter of interest will be contained within the interval.

However, this confidence interval interpretation is rather inline with the Bayesian view (Carlin and Louis, 2010, Liu and Powers, 2012, Gelman et al., 2014). Bayesian analysis methods, in additional to incorporating prior information in the analysis, address the aforementioned drawbacks of the maximum likelihood methods by allowing us to write the full joint probability distribution of all the parameters of interest that takes into consideration the various sources of uncertainty which provides a commonsense interpretation of the Bayesian credible intervals (Gelman et al., 2014).

3.6.1.1 Bayesian For Standard Poisson Regression Model

Here we first layout the Bayesian framework for the Standard Poisson Regression(SPR) model (without overdispersion). Using a SPR model, the relationship between our outcome variable Y and the input variables X &A is expressed as in equation (3.1). Let $\lambda(x) = \mathbf{x}'\boldsymbol{\beta}$, $\mu(x) = e^{\lambda(x)}$, $y \sim Pois(\mu(x))$, where, $\mathbf{x} = \{\mathbf{x_1}, \mathbf{A}\}$, such that $\mathbf{x_1} = \{x_{i1}, ..., x_{n1}\}$ represents the predictive biomarker measured at baseline for each subject and $\mathbf{A}$ the treatment assignment, such that, $A = 1$ if a subject is assigned to treatment and $A = 0$ if assigned to the standard of care.

The likelihood function associated with the SPR model is then written as:

$$L(\boldsymbol{\beta}; \mathbf{Y}) \quad = \quad \prod_{i=1}^{n} \left\{ \frac{e^{-e^{\mathbf{x}_i'\boldsymbol{\beta}}} \quad e^{(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}}}{y_i!} \right\}$$

Introducing independent Gaussian priors on the slope parameters of the SPR model, such that, $\beta_j \sim \mathcal{N}(\mu_{\beta_j}, \sigma^2_{\beta_j})$, where $j = 0, ..., p$, the joint probability distribution function of $\boldsymbol{\beta}$ is

given as:

$$\pi(\boldsymbol{\beta}) \;=\; \prod_{j=0}^{p}\left\{\frac{1}{\sqrt{2\pi\sigma^2_{\beta_j}}}\, e^{\left\{\frac{-(\beta_j-\mu_{\beta_j})^2}{2\sigma^2_{\beta_j}}\right\}}\right\} \tag{3.12}$$

where $\boldsymbol{\beta}$ is a vector of length p such that p represents the number of parameters the SPR model of equation(1). Finally given data $\{\mathbf{x}, \mathbf{y}\}$, using Bayes' rule, the unstandardized joint posterior distribution of $\boldsymbol{\beta}$ is proportional to:

$$\begin{aligned}\pi(\boldsymbol{\beta}|\mathbf{x}, y) &\propto f(y|\mathbf{x}, \beta)\pi(\beta) \\ &\propto \prod_{i=1}^{n}\left\{\frac{e^{-e^{\mathbf{x}_i'\boldsymbol{\beta}}}\; e^{(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}}}{y_i!}\right\}\prod_{j=0}^{p}\left\{\frac{1}{\sqrt{2\pi\sigma^2_{\beta_j}}}\, e^{\left\{\frac{-(\beta_j-\mu_{\beta_j})^2}{2\sigma^2_{\beta_j}}\right\}}\right\}\end{aligned} \tag{3.13}$$

3.6.1.2 Bayesian For Zero Inflated Poisson Regression Model

When modeling a count data, it is not uncommon to encounter data with an excess of zeros. These data are usually called zero-inflated (ZI) outcome data. With ZI outcome data the number of observed zeros is greater than one would expect from a standard Poisson model (Jang et al., 2010). This in turn leads to violation of the common Poisson model, where the variance is equal to the mean. As a result, the zero-inflated Poisson (ZIP) model (Lambert, 1992) and Poisson hurdle (PH) model (Mullahy, 1986) have been developed to overcome the difficulties that arise from the ZI outcome data. When the outcome data contains both excess sampling zeros and structural zeros, ZIP models are typically used. Sampling zeros are zeros that are part of the Poisson distribution, and it is assumed these zeros are observed by chance. Whereas structural zeros arise due to a particular structure in the data set (Hu et al., 2011, Hua et al., 2014) such that asking person the number of cigarettes he/she smoked in the past week though that person is not a smoker. In this subsection we lay out the steps to obtain a Bayesian estimates of the slope parameters for a ZIP regression model. When

dealing with ZI count data, one can think of the observed counts as two different outcomes: (1) those which are inflated (more than those expected) zeros and (2) those which are in agreement with the underlying Poisson distribution. The probability mass function of Y given π and λ can be written as:

$$Pr(Y = y|\pi, \lambda) = \begin{cases} \pi + (1-\pi)e^{-\lambda} & \texttt{if y = 0} \\ (1-\pi)\frac{e^{-\lambda}\ \lambda^y}{y!} & \texttt{if y > 0} \end{cases} \tag{3.14}$$

where π represent the probability that the observed zero is from the zero-inflated stage and λ represents the mean for the Poisson count stage provided the observed value is not zero inflated. From equation (3.14) one can clearly see that, the ZIP regression model has two stages: the zero-inflation stage and Poisson count stage. Depending on the problem at hand, the covariates in a given data set can be used in both stages to estimate the parameters π and λ simultaneously. Commonly loglinear and logit models are used to relate the covariates with the parameters π and λ. In this particular case, assume all the covariates are used to in estimating π and λ, such that,

$$\begin{aligned} log(\lambda) &= \mathbf{x_1'\beta_1} = \beta_{10} + \beta_{11}X + \beta_{12}A + \beta_{13}XA \\ log(\frac{\pi}{1-\pi}) &= \mathbf{x_2'\beta_2} = \beta_{20} + \beta_{21}X + \beta_{22}A + \beta_{23}XA \end{aligned}$$

With a little algebric simplification, we can express the equations above and write $\lambda = e^{\mathbf{x_1'\beta_1}}$ and $\pi = \frac{e^{\mathbf{x_2'\beta_2}}}{1+\ e^{\mathbf{x_2'\beta_2}}}$. The likelihood function for the random variable $Y \sim ZIP(\lambda, \pi)$ is finally expressed as

$$f(Y|\boldsymbol{\beta_1}, \boldsymbol{\beta_2}) \ = \ \prod_{i=1}^{m} \left\{ \frac{e^{\mathbf{x_2'\beta_2}}}{1+\ e^{\mathbf{x_2'\beta_2}}} \ + \ \left(1 \ - \ \frac{e^{\mathbf{x_2'\beta_2}}}{1+\ e^{\mathbf{x_2'\beta_2}}}\right) \ e^{-e^{\mathbf{x_1'\beta_1}}} \right\} \tag{3.15}$$

$$\times \prod_{i=k+1}^{n} \left\{ \left(1 - \frac{e^{\mathbf{x_2'}\boldsymbol{\beta_2}}}{1 + e^{\mathbf{x_2'}\boldsymbol{\beta_2}}} \right) e^{-e^{\mathbf{x_1'}\boldsymbol{\beta_1}}} \frac{(e^{\mathbf{x_1'}\boldsymbol{\beta_1}})^{y_i}}{y_i!} \right\}$$

where from the n total number of observations, the first m subjects who responded have zero value and the rest have nonzero values. To proceed with Bayesian analysis, one first needs to elicit a prior distribution for the $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ coefficients. If an independent Gaussian prior is assumed for each coefficient of $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ such that, $\beta_{1j} \sim \mathcal{N}(\mu_{\beta_{1j}}, \sigma^2_{\beta_{2j}})$ and $\beta_{2j} \sim \mathcal{N}(\mu_{\beta_{2j}}, \sigma^2_{\beta_{2j}})$, then we can write the joint prior distribution as:

$$h(\boldsymbol{\beta_1}, \boldsymbol{\beta_1}) = \prod_{j=0}^{3} \left\{ \frac{1}{\sqrt{2\pi\sigma^2_{\beta_{1j}}}} e^{\left\{ \frac{-(\beta_{1j} - \mu_{\beta_{1j}})^2}{2\sigma^2_{\beta_{1j}}} \right\}} \right\} \prod_{j=0}^{3} \left\{ \frac{1}{\sqrt{2\pi\sigma^2_{\beta_{2j}}}} e^{\left\{ \frac{-(\beta_{2j} - \mu_{\beta_{2j}})^2}{2\sigma^2_{\beta_{2j}}} \right\}} \right\} \tag{3.16}$$

Using the likelihood function in equation (3.15) and the joint prior distribution functions given equation (3.16), applying Bayes rule, the unstandardized joint posterior distribution for a ZIP regression model written as:

$$\begin{aligned} g(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}|\mathbf{Y}) \propto & \prod_{i=1}^{k} \left\{ \frac{e^{\mathbf{x_2'}\boldsymbol{\beta_2}}}{1 + e^{\mathbf{x_2'}\boldsymbol{\beta_2}}} + \left(1 - \frac{e^{\mathbf{x_2'}\boldsymbol{\beta_2}}}{1 + e^{\mathbf{x_2'}\boldsymbol{\beta_2}}} \right) e^{-e^{\mathbf{x_1'}\boldsymbol{\beta_1}}} \right\} \\ & \times \prod_{i=k+1}^{n} \left\{ \left(1 - \frac{e^{\mathbf{x_2'}\boldsymbol{\beta_2}}}{1 + e^{\mathbf{x_2'}\boldsymbol{\beta_2}}} \right) e^{-e^{\mathbf{x_1'}\boldsymbol{\beta_1}}} \frac{(e^{\mathbf{x_1'}\boldsymbol{\beta_1}})^{y_i}}{y_i!} \right\} \times \\ & \prod_{j=0}^{3} \left\{ \frac{1}{\sqrt{2\pi\sigma^2_{\beta_{1j}}}} e^{\left\{ \frac{-(\beta_{1j} - \mu_{\beta_{1j}})^2}{2\sigma^2_{\beta_{1j}}} \right\}} \right\} \prod_{j=0}^{3} \left\{ \frac{1}{\sqrt{2\pi\sigma^2_{\beta_{2j}}}} e^{\left\{ \frac{-(\beta_{2j} - \mu_{\beta_{2j}})^2}{2\sigma^2_{\beta_{2j}}} \right\}} \right\} \end{aligned} \tag{3.17}$$

A closed form solution for equation (3.13) and equation (3.17) is analytically unobtainable because of the lack of conjugacy between the standard Poisson likelihood function and the Gaussian priors as well as between the ZI Poisson likelihood function and the Gaussian prior. Instead we will use a Markov Chain Monte Carlo (MCMC) methods called Hamiltonian Monte Carlo (HMC) to get the posterior mean estimates and the 95% credible intervals

(CIs) of the $\beta's$ which in turn are used to estimate our metric of interest Ψ.

3.6.2 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is a Markov chain Monte Carlo (MCMC) algorithm that avoids the random walk behavior and sensitivity to correlated parameters that plague many MCMC methods by taking a series of steps informed by first-order gradient information (Hoffman and Gelman, 2014, Brooks et al., 2011). To implement Hamiltonian Monte Carlo, one first needs to write the Hamiltonian which is an energy function for the joint state of "position", q, and "momentum", p as:

$$H(q,p) \quad = \quad U(q) + K(p) \tag{3.18}$$

where $U(q)$ is the potential energy and $K(p)$ is the kinetic energy. Further q and p are assumed to be independent and each has a canonical distribution. In what follows we will use q to represent the variable of interest, and p will be introduced to implement the Hamiltonian principle. In the Bayesian context, the posterior distribution of the model parameters is the target of the analysis. These parameters take the position of q and using the potential energy concept we can express the posterior distribution as a canonical distribution $U(q) = -log[\pi(q)L(q/D)]$, where $\pi(q)$ is the prior density, and $L(q/D)$ is the likelihood function given data D. The kinetic energy $K(p)$, is mainly taken to be the negative log probability density of a Gaussian distribution with mean zero and convariance matrix, M and is written as

$$K(p) = \frac{p^T M^{-1} p}{2} \tag{3.19}$$

where M is a symmetric, positive-definite "mass matrix". The partial derivatives of $H(q,p)$ determine how q and p change over time, according to Hamiltonian equations:

$$\begin{aligned} \frac{dq}{dt} &= [M^{-1}p_i] \\ \frac{dp_i}{dt} &= -\frac{\partial U}{\partial q_i} \end{aligned} \tag{3.20}$$

To implement HMC on the computer, we first need to approximate the Hamiltonian equations by discretizing time into increments of ϵ, where ϵ is small. Let the Hamiltonian be $H(\beta, w) = U(\beta) + K(w)$ where we replace q by β and p by w to reflect the parameters of interest and the moment variables respectively. Further assume that M is a diagonal matrix with diagonal elements given as $m_1, ...m_d$, which leads to $K(w) = \sum_{i=1}^{d} \frac{w_i^2}{2m_i}$. Approximate solutions for systems of differential equations are better obtained using the leap-frog method:

$$\begin{aligned} w_i(t+\epsilon/2) &= w_i(t) + (\epsilon/2)\frac{\partial U}{\partial \beta_i}\beta(t) \\ \beta_i(t_\epsilon) &= \beta_i(t) + \epsilon\frac{w_i(t+e/2)}{m_i} \\ w_i(t+\epsilon) &= w_i(t+\epsilon/2) - (\epsilon/2)\frac{\partial U}{\partial \beta_i}\beta(t+\epsilon) \end{aligned} \tag{3.21}$$

where the derivatives with respect to time are obtained from equation(3.20). The basic idea here is, if we start with $\beta_i(0)$ and $w_i(0)$ at $t = 0$, we can use equation (3.21) above iteratively to get the trajectory values of position and momentum at times $\epsilon, 2\epsilon, ...$, and the final values for $\beta(\tau)$ and $w(\tau)$. The total number of steps then will be $\frac{\tau}{\epsilon}$. Let $\beta^{(i)}$ be the current value of the parameter β. The leap-frog Hamiltonian Monte Carlo algorithm is:

Algorithm 2 Hamiltonian Monte Carlo algorithm

1: Sample $w \sim \mathcal{N}_d(0, D)$ ▷ where D is the covariance Matrix

2: Using leapfrog method, simulate Hamiltonian dynamics on location $\beta^{(i)}$ and momentum w for L steps with stepsize ϵ. Let these updated value be β^* and w^*.

3: set $\beta^{(i+1)} = \beta^*$ with probability $min\{1, r(\beta^{(i)}, \beta^*)\}$ such that

$$\begin{aligned} r(\beta^{(i)}, \beta^*) &= \frac{p(\beta^*|y}{p(\beta^{(i)}|y} \frac{p(w^*}{p(w^i)} \\ &= \frac{p(y|\beta^*)p(\beta^*)}{p(y|\beta^{(i)})p(\beta^{(i)})} \frac{\mathcal{N}_d(w^*; 0, D)}{\mathcal{N}_d(w^{(i)}; 0, D)} \end{aligned}$$

Otherwise set $\beta^{(i+1)} = \beta^i$.

3.6.3 Prior Elicitation From Clinician Inputs

One of the major concerns when fitting a Bayesian model is prior elicitation. Here we outline first how the best clinician guesses or inputs can be converted into hyperparameters which are incorporated during prior elicitation. Lets say K_1, K_2, K_3 and K_4 are expected asthma exacerbation rates given the 25^{th} and 75^{th} percentile value of the biomarker for the standard of care and active treatment groups respectively. The K values reflect the best judgment of the experts about the likely outcome of the disease under consideration give the predictive biomarker value. Further, letting Z_1 and Z_2 representing the 25^{th} and 75^{th} percentile value of the biomarker, with a little algebra manipulation, we can use the following equations to get the Poisson model parameters (Details of the derivation is given in Appendix A.0.3).

$$\begin{aligned} \beta_0 &= \frac{K_1 * Z_1 - K_1 * Z_2}{Z_1 - Z_2} \\ \beta_1 &= \frac{K_1 - K_2}{Z_1 - Z_2} \\ \beta_2 &= \frac{K_1 * Z_2 - K_2 * Z_1 - K_3 * Z_2 + K_4 * Z_1}{Z_1 - Z_2} \end{aligned} \tag{3.22}$$

$$\beta_3 = \frac{K_2 - K_1 + K_3 - K_4}{Z_1 - Z_2}$$

where $ln(K_1) = \beta_0 + \beta_1 X_{z_1}$, $ln(K_2) = \beta_0 + \beta_1 X_{z_2}$, $ln(K_3) = \beta_0 + \beta_2 + (\beta_1 + \beta_3) X_{z_1}$ and $ln(K_4) = \beta_0 + \beta_2 + (\beta_1 + \beta_3) X_{z_2}$. Priors of the normal distribution centered around the $\beta' s$ obtained from the above equations were used. To assess the effect of variance specification on the main parameter of interest Ψ_{B_P}, priors were elicited from being vague to more informative. Vague priors are priors with a high variance used to express the probability mass is spread out over a large plausible values instead of concentrating in specific values. On the other hand, informative priors are priors with a smaller variance express a strong belief one has about the parameters of interest before the data collection.

3.6.4 Estimation of Ψ_B (BBNB)

Assume the default treatment is "treat none (standard of care)". After getting the estimated $\hat{\beta}$ values as posterior means applying the Bayesian framework method mentioned in detail in the previous subsections (5.1 and 5.2), we can use equation (3.10) to get an estimate of the BBNB (Ψ_{B_P}) as:

$$\begin{aligned}
\hat{\Psi}_{B_P} &= L(\hat{\beta}_B, \delta_P) - L(\hat{\beta}_B, \delta_{opt}) \qquad (3.23)\\
&= \int \{exp\{\hat{\beta}_0 + \hat{\beta}_1 x\}\} f(x) dx - \left\{ \int_{\mathcal{X}_1} \{exp\{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 A + \hat{\beta}_3 x A\}\} f(x) dx \right\}\\
&\quad - \int_{\mathcal{X}_0} \{exp\{\hat{\beta}_0 + \hat{\beta}_1 x\}\} f(x) dx\\
&= \int_{\mathcal{X}_1} \{exp\{\hat{\beta}_0 + \hat{\beta}_1 x\}\} f(x) dx - \left\{ \int_{\mathcal{X}_1} \{exp\{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 A + \hat{\beta}_3 x A\}\} f(x) dx \right\}\\
&= \int_{\mathcal{X}_1} \left\{ exp\{\hat{\beta}_0 + \hat{\beta}_1 x\} - \left\{ exp\{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 A + \hat{\beta}_3 x A\} + c \right\} \right\} f(x) dx
\end{aligned}$$

where L refers to the loss incurred for a given $\hat{\beta}_B$, δ a treatment decision, $\hat{\beta}_B$ posterior mean estimates of βs' , $f(x)$ the probability density function of the biomarker under study and

$\mathcal{X}_1 \in \{X : X < \frac{-\beta_2}{\beta_3}\}$ provided $\beta_3 > 0$. If we assume, for example, the biomarker under study has a $\mathcal{U}[a, b]$, then equation (3.23) can be simplified further to have a closed form as:

$$\begin{aligned}\hat{\Psi}_{B_P} &= \int_a^d \Big\{exp\{\beta_0 + \ \beta_1 x\} - exp\{\hat{\beta}_0 + \ \hat{\beta}_1 x + \hat{\beta}_2 A + \hat{\beta}_3 x A\}\Big\} \frac{1}{b-a} dx \\ &= \frac{1}{b-a}\left\{\left[\frac{e^{\hat{\beta}_0 + d\hat{\beta}_1}}{\hat{\beta}_1} - \frac{e^{\hat{\beta}_0 + a\hat{\beta}_1}}{\hat{\beta}_1}\right] - \left[\frac{e^{\hat{\beta}_0 + \hat{\beta}_2 + d(\hat{\beta}_1 + \hat{\beta}_3)}}{\hat{\beta}_1 + \hat{\beta}_3} - \frac{e^{\hat{\beta}_0 + \hat{\beta}_2 + a(\hat{\beta}_1 + \hat{\beta}_3)}}{\hat{\beta}_1 + \hat{\beta}_3}\right]\right\}\end{aligned}$$

such that $\beta_1 \neq 0$, $(\beta_1 + \beta_3) \neq 0$ and $d = \frac{-\beta_2}{\beta_3}$. d is the cut-off value used under the optimal treatment rule as obtained in equation (3.9). Similarly, when the biomarker under consideration has a $\mathcal{N}(\mu, \sigma^2)$ distribution, $\hat{\Psi}_{B_P}$ is obtained as:

$$\begin{aligned}\hat{\Psi}_{B_P} &= \int_{\mathcal{X}_1} \Big\{exp\{\hat{\beta}_0 + \ \hat{\beta}_1 x\} - exp\{\hat{\beta}_0 + \ \hat{\beta}_1 x + \hat{\beta}_2 A + \hat{\beta}_3 x A\}\Big\} f(x) dx \\ &= \int_{\mathcal{X}_1} \Big\{exp\{\hat{\beta}_0 + \ \hat{\beta}_1 x\} - exp\{\hat{\beta}_0 + \ \hat{\beta}_1 x + \hat{\beta}_2 A + \hat{\beta}_3 x A\}\Big\} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx\end{aligned}$$

When the default treatment is "treat all", an estimates of the $\hat{\Psi}_{B_T}$ which can be obtained following similar steps is provided in Appendix.

3.7 Simulation Study

To demonstrate our method, we used two different simulation studies. The first is a toy-simulation while the second a simulation that was done to mimic the phase II clinical trial conducted by Genentech to develop drug AA for the treatment of asthma. For the toy-simulation, we generated data from the standard Poisson regression model of equation (3.1). This simulation was done under the following scenarios:

- Randomization was 1 : 1;
- Biomarker was assumed to have a know probability distribution: Uniform or Normal;

and

- Different biomarker performance scenarios (see Figure 3.2): Strong , moderate and weak.

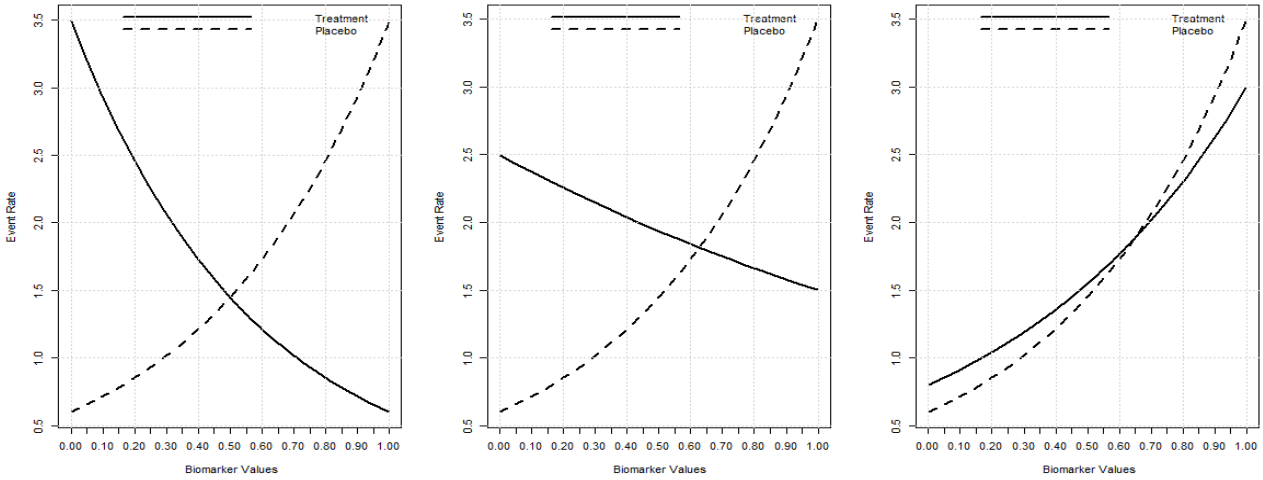

Figure 3.2: Plots showing the expected event rate of each treatment arm for a given biomarker and different combination of K values. For the plot in the left (strong biomarker), $K_1 = 0.6, K_2 = 3.5, K_3 = 3.5$ and $K_4 = 0.6$. The plot in the middle (moderate biomarker) has $K_1 = 0.6, K_2 = 3.5, K_3 = 2.5$ and $K_4 = 1.5$ while the plot in the left (weak biomarker) has $K_1 = 0.6, K_2 = 3.5, K_3 = 0.8$ and $K_4 = 3.0$. The biomarker is assumed to have a standard uniform distribution.

For the purpose of specifying the priors, we considered different sets of the K values and converted them into the $\beta's$ using equation (3.22). These $\beta's$ in turn were used as mean values when we specified a normal prior for each parameter of the model. The variances of the priors were set in such a way that they reflect a weak and strong prior belief of the clinicians about the performance of the biomarker under consideration before data collection. A weak prior belief was reflected in our simulation studies by assigning a large variance to each of the normal priors we considered and vise versa to show a strong prior belief.

Our second simulation was conducted to mimic the phase II clinical trial study conducted by Genentech to develop drug AA used for the treatment of asthma. As per the protocol of the study we used 1:1 randomization to assign half of the subject to placebo and the

rest half to treatment (taking any of the the three doses of drug AA). A biomarker was generated from a normal distribution to mirror the log of the BMK biomarker measured at baseline for each patient. BMK was the predictive biomarker under consideration to assign treatment for patients with uncontrolled Asthma. It was previously claimed that drug AA works more effectively for patients with high pretreatment BMK values (Corren et al., 2011). The high-BMK subgroup was defined as patients with baseline BMK level greater or equal to the median value. By specifying the $\beta's$ that mimic results of the phase II clinical trial, we generated data from a standard Poisson regression model. To set the normal priors we took into consideration the results of the previous study when setting up the K values and used different variances to reflect the clinician belief on the performance of the BMK biomarker as predictive biomarker.

Regardless of the prior picked, conducting a sensitivity analysis to assess the effect of elicitation of different priors and other features of the model on the posterior inferences is a customary practice when fitting models using Bayesian methods to assess robustness. In the context of our method, we investigated the sensitivity of the metrics $\hat{\Psi}_{B_T}$ and $\hat{\Psi}_{B_P}$ using three different approaches. In the first case, we considered a range of values for the hyperparameter $\sigma^2_{\beta_j}$ associated with each β_j, such that, $j = 1, ..., p$, where p is the number of coefficients in the model and computed the posterior means and the 95% credible interval (CI) for $\hat{\Psi}_{B_T}$ and $\hat{\Psi}_{B_P}$. In case two, we centered the priors at different values and computed the posterior means and 95% CIs for $\hat{\Psi}_{B_T}$ and $\hat{\Psi}_{B_P}$. Finally, as over-dispersion is a common phenomenon when fitting a Poisson regression models, we assessed how robust the estimated mean posterior values of $\hat{\Psi}_{B_T}$ and $\hat{\Psi}_{B_P}$ can be during over/under dispersion. For this purpose, we compared the Bayesian estimates of $\hat{\Psi}_{B_T}$ and $\hat{\Psi}_{B_P}$ under standard Poisson, zero-inflated Poisson and negative binomial regression models.

3.8 Results

Results of the first simulation are presented in Table 3.1 where the biomarker was assumed to have a $\mathcal{U}(0,1)$ distribution. In all the toy simulations conducted, a sample size of $n = 350$, 5000 iterations with 4 chains was used. Three different scenarios were taken to reflect the prior belief of the experts by specifying different combination of K-values. These K-values were chosen in such a way that they show a strong, moderate and weak biomarker performance. As stated previously, we used a Gaussian prior for each of the coefficients with a mean specified by converting the Ks to βs. To assess the effect of the variance assigned to each coefficient, we investigated three different cases: in case 1 $\sigma_\beta^2 = 0.5$; case 2 $\sigma_\beta^2 = 10$ and case 3 $\sigma_\beta^2 = 100$. Under a strong biomarker performance assumption (scenario 1) we set $K_1 = 0.6$, $K_2 = 3.5$, $K_3 = 3.5$ and $K_4 = 0.6$. Further setting $\sigma_\beta^2 = 0.5$ in order to reflect a more informative prior, the posterior mean and standard error of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ were estimated to be $0.603(se = 0.079)$ and $0.549(se = 0.072)$ with their respective 95% credible intervals given by $(0.460, 0.766)$ and $(0.419, 0.679)$. In Table 3.1, in addition to the poster mean estimates, standard error and credible intervals for $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$, three other estimated metrics are presented. The soc (standard of care) refers to the posterior estimated mean even rate when all the subjects are assigned to the control arm regardless of their biomarker values. The act (active treatment) when all subjects are assigned to treatment arm and the opt (optimal treatment) when subjects are assigned treatment based on their biomarker value. Further to make sure that the posterior estimated means are not far from what one would expect, we first simulated a large data set of sample size 10,000 and estimated $\hat{\Psi_P}$ and $\hat{\Psi_T}$ (0.584 and 0.506 respectively). Similarly under the prior belief only $\hat{\Psi_P}$ and $\hat{\Psi_T}$ were calculated to be (0.687 and 0.691 respectively). If we look at scenario 1 of Table 1, when $\sigma_\beta^2 = 0.5$ for example, as one would expect $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ have values that fall between what one can get using only the data and the prior belief. As it can be seen from

Table 3.1, as the prior becomes less informative (σ_β^2 gets larger) the posterior estimates of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ get closer to the estimated values one could obtain using only the data.

The sensitivity of the $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ estimates to misspecification of the prior means was assessed and results are presented in Table 3.1. The means of the β priors were set to different values in scenario 1, scenario 2 and scenario 3. Generally the posterior estimates of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ showed a slight variation when a very informative priors were used (i.e) setting $\sigma_\beta^2 = 0.5$. Looking at $\hat{\Psi}_{B_P}$ for example, its values changed from 0.603 under scenario 1 to 0.549 under scenario 2 and to 0.468 under scenario 3 while σ_β^2 was fixed to 0.5. Nevertheless, under less informative priors (when $\sigma_\beta^2 = 10$ and $\sigma_\beta^2 = 100$) $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ did not display a noticeable change under the three scenarios. Density plots for the posterior estimates of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ under a standard Poisson model for the three different priors is shown in Figure 3.3.

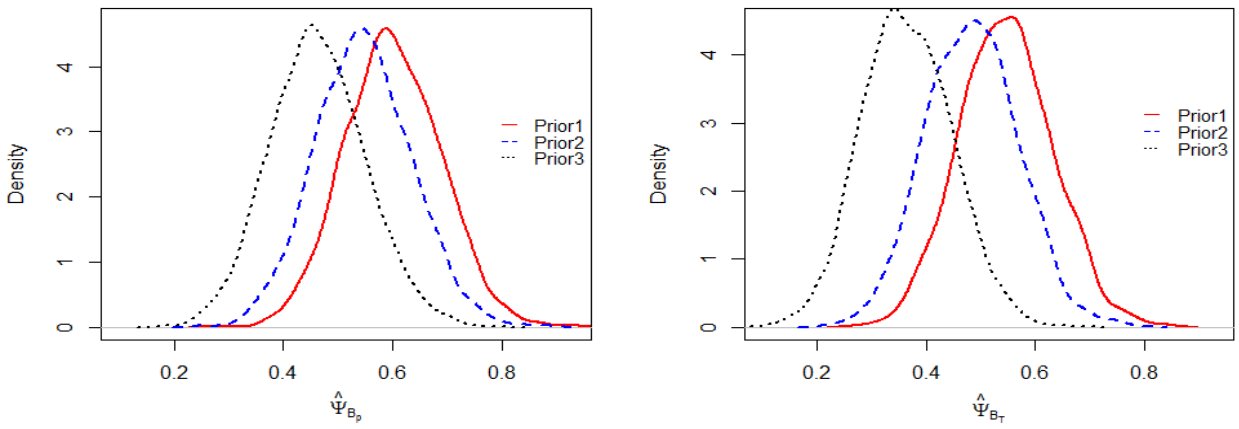

Figure 3.3: Shows the posterior density plots of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ under a standard Poisson regression model for three different priors.

From Table 3.1, we can further see how robust the estimators $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ are to misspecifying a wrong variance for each prior. For this purpose we set $\sigma_{\beta_j}^2$ to $0.5, 10$ and 100. The posterior estimate of $\hat{\Psi}_{B_P}$ for example changed from 0.603 when $\sigma_\beta^2 = 0.5$ to

0.576 when $\sigma_\beta^2 = 10$ but remain unchanged when σ_β^2 was further increased to 100 under scenario 1 of Table 3.1. This trend was consistent in all the three scenarios considered. A slight decrease in the posterior means was observed as σ_β^2 was increased from 0.5 to 10 but remained constant when σ_β^2 was further increased to 100. The posterior density plots for $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ under scenario 1 are shown in Figure 3.4. An important point from this simulation is that, generally the posterior estimates of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ reasonably seems to be robust to the miss specifying a wrong variance to the priors.

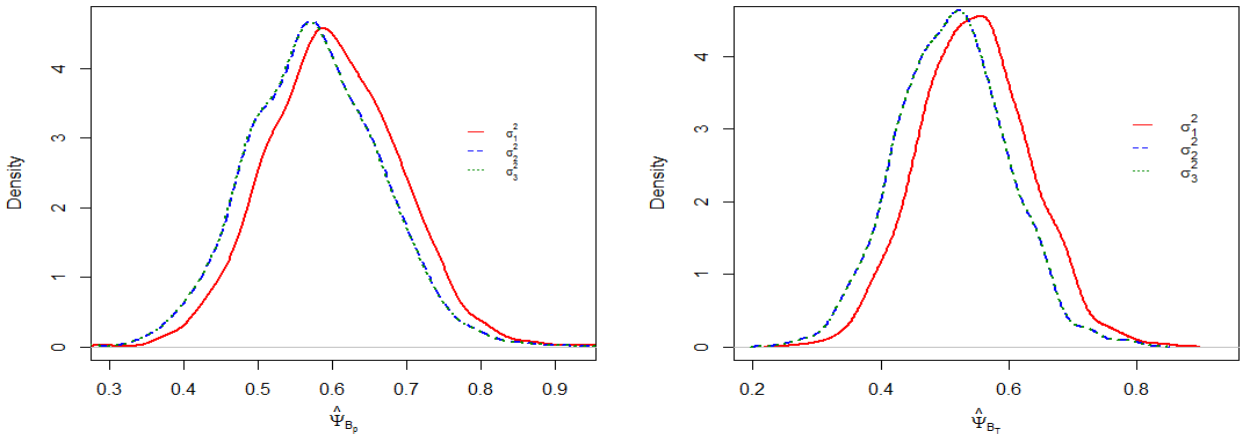

Figure 3.4: Shows the posterior density plots of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ under scenario 1, setting σ_β^2 to different values. The density with red color corresponds to case where σ_β^2=0.5, the blue density for σ_β^2=10 and the green density for σ_β^2=100.

As overdispersion is a common phenomena when fitting a log linear model, further robustness check of the $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ estimators was conducted by estimating these metrics using a Bayesian zero-inflated Poisson and negative binomial regression models. The results are provide in Table 3.2 (for ZIP model) and Table 3.3 (for negative binomial modes). Taking scenario 1 and the setting where $\sigma_\beta^2 = 0.5$ for example, the estimated value of $\hat{\Psi}_{B_P}$ was 1.07 under the standard Poisson regression model, 1.09 under negative binomial model, and 1.08 under ZIP model. Similar results are observed under the other two scenarios and

other setting where σ_β^2 was set to 10 and 100. A similar simulation was further conducted assuming a $\mathcal{N}(0,1)$ and results are presented as supplementary in Table S1, Table S2 and Table S3. The key point is, both $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ were reasonably robust to misspecification of the prior means, variances and models (standard Poisson, ZIP or negative binomial).

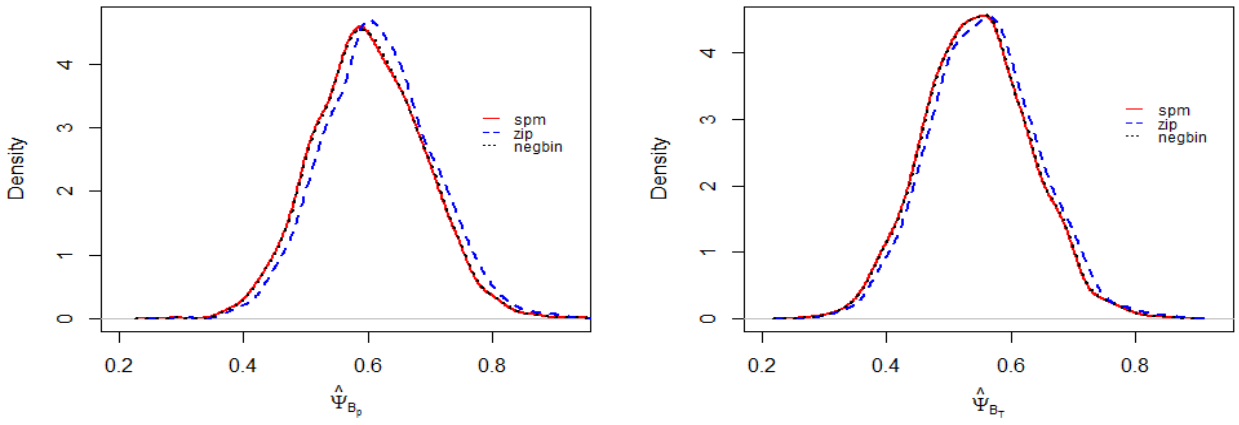

Figure 3.5: Shows the posterior density plots of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ for standard Poisson regression model (spm), zero-inflated regression Poisson model (zip) and negative binomial regression model (negbin).

The second simulation was done to mimic the phase II clinical trial study conducted by Genetech to develop drug AA for asthma treatment. A logarithmic transformation was first used to normalize the BMK biomarker assumed to be measured at baseline for each subject, such that, it has a $\mathcal{N}(3.89, 0.21)$ distribution. Taking the coefficients associated with a SPR model fit from previous studies, a data of sample size 460 was simulated to mirror the phase II clinical trial study. This is somewhat large for a randomized phase II study. Figure 3.6, shows the exacerbation rate per year as a function of the percentile biomarker (log of BMK) value for the active treatment and standard of care groups. The left plot shows the expertise prior belief about the relationship between exacerbation rate and and log of the BMK values while the right plot shows the posterior relationship after updating the priory belief using

data collected from the phase II clinical trial study.

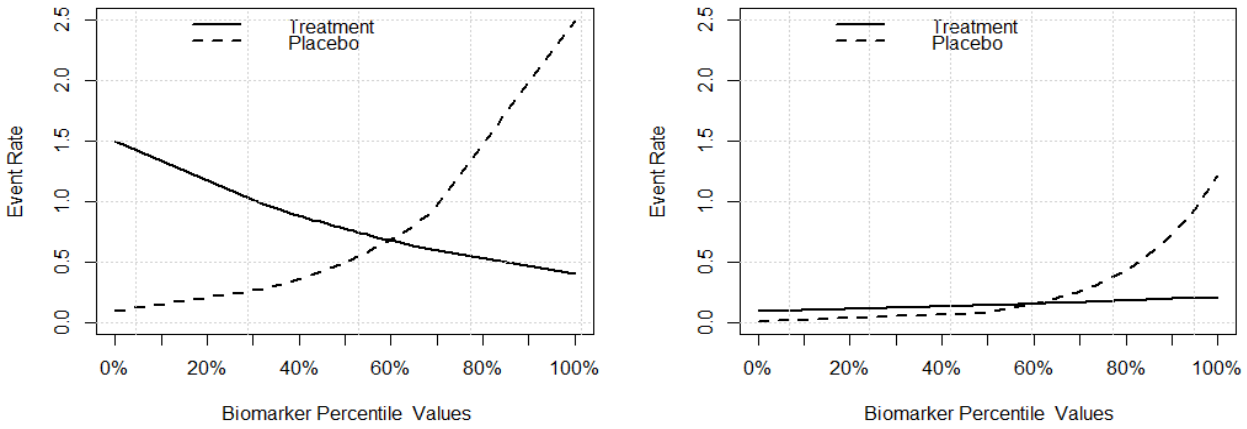

Figure 3.6: Plot showing the relationship between exacerbation rate per year and the percentile values of the biomarker for each arm, the treatment group and the standard of care group. For the plot in the left only prior information was used while for plot in the right the prior information was updated using phase II data.

The posterior mean estimates of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ were found to be 0.276(se=0.044) and 0.011 (se=0.005) with the prior for each of the β's assumed to have a Gaussian distribution such that $\beta_0 \sim \mathcal{N}(-3.72, 0.5)$, $\beta_1 \sim \mathcal{N}(0.78, 0.5)$, $\beta_2 \sim \mathcal{N}(5.01, 0.5)$ and $\beta_3 \sim \mathcal{N}(-1.27, 0.5)$ under scenario 1 of Table 3.3 where a standard Bayesian Poisson regression model was fitted. From Table 3.3, Table S5 and Table S6, one can look at scenario 1, scenario 2 and scenario 3 fixing σ^2_β to 0.5 or 10 or 100 to assess the sensitivity of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ to a change in the mean value of the priors. To evaluate robustness of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ to model misspecification (standard Poisson, ZIP or negative binomial) one can look at scenario 1 and a setting where σ^2_β=0.5, for example, and compare the values from Table 3.3, Table S5 and Table S6. Generally misspecification of prior means and models have less impact on the estimated values of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$. However, we observed a deviation in the estimated value of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ when the prior variance was changed from 0.5 to 10(or 100). This difference is not far from expected. Since setting σ^2_β=0.5 (very informative prior) puts a lot

of weight on the prior belief, it is customary to see the values of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ deviating a little from what one could have expected from the information available in the data only.

3.9 Discussion

Our method, the Bayesian decision theoretic framework for evaluating clinical utility of predictive biomarker, is used to estimate the expected reduction in an event rate under a biomarker guided treatment with a count endpoint. This design enable assessing a treatment selection biomarker at the end of phase II by incorporating experts (clinicians, assay developers or/and biomarker scientists) believe about the biomarker performance as part of the evaluation method under a more general circumstance. Our approach, in addition to integrating prior belief about the predictive biomarker performance, take the biomarker distribution in to consideration. From a biomarker development design point of view, one would expect to have a clear picture about the biomarker distribution at the end of phase II. Further data from the phase II studies can be used to get an understanding about the biomarker distribution.

Frequentist designed metrics for evaluation treatment selection biomarker have been previously proposed (Song and Pepe, 2004, Brinkley et al., 2010, Janes et al., 2011). The metrics $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ are closely related to the metrics Θ_0 and Θ_1 proposed by (Janes et al., 2011) respectively. Our Bayesian decision theoretic framework however, extends the existing methods in a number of ways. One notable difference is the ability of our method to incorporate expertise belief about the biomarker performance in the analysis in the form of a prior and integrating the biomarker distribution in the analysis are other additions in our proposed method.

Prior elicitation is a major hurdle when implementing a Bayesian framework. To ease this problem, we developed easy to use equations to convert experts belief into β coefficients which

are in turn used as hyperparameters. Further, we conducted extensive sensitivity analysis to assess on how misspecification of the priors' means, variances and the model in general could affect the estimated values of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$. Results from our simulation indicate that our proposed metrics are generally less sensitive to the aforementioned misspecifications. However, a slight difference was observed when using very informative and less informative priors. But this is in line with the Bayesian analysis methodology.

To make conclusions regarding the biomarker performance based on $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ estimates could be challenging. A high value of $\hat{\Psi}_{B_P}$ or $\hat{\Psi}_{B_T}$ is shows a better biomarker performance. However, the question of how large the values of $\hat{\Psi}_{B_P}$ or $\hat{\Psi}_{B_T}$ need to be to consider the biomarker as clinically valid for the purpose of guiding treatment for patients, depend on many factors. The type of disease under consideration, the time in which the event rate is being estimated (per week/month/year etc) and others. Once a clinically meaningful value of $\hat{\Psi}_{B_P}$ or $\hat{\Psi}_{B_T}$ is defined consulting expert clinicians, making a valid inference would follow straight forward based on the 95% credible intervals associated with $\hat{\Psi}_{B_P}$ or $\hat{\Psi}_{B_T}$.

The Bayesian decision theoretic framework described here, even though the setup was done for a count end point, extension should be straightforward if one want to consider a binary, continuous or time-to event endpoints. In the event of binary and time-to event endpoint, the boundaries of $\hat{\Psi}_{B_P}$ or $\hat{\Psi}_{B_T}$ would be [0,1] and ease the challenge in the interpretation of the results. Our method considered only a continuous uniform and normal biomarker, but cases with a discrete biomarker and biomarkers with other distributions could be handled with a slight modification in the equations developed to estimate $\hat{\Psi}_{B_P}$ or $\hat{\Psi}_{B_T}$. Further, this method, though it was set up with an intention of making a treatment selection biomarker in phase II for a 1:1 randomized clinical trial design, can be generalized to other study designs as needed.

Table 3.1: Posterior mean, standard error and 95% credible intervals of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ fitting a standard Poisson regression model assuming a $\mathcal{U}(0,1)$ biomarker. Data was generated from a standard Poisson model with sample size of 350. Coefficients used for data simulation are: $\beta_0 = -0.40$, $\beta_1 = 2.75$,$\beta_2 = 1.45$ and $\beta_3 = -3.00$.

Scenario 1: K-values $K_1 = 0.6$	$K_2 = 3.5$	$K_3 = 3.5$	$K_4 = 0.6$		
μ_{β_j}	$\beta_0 \sim N(-0.50, \sigma_j^2)$	$\beta_1 \sim N(1.76, \sigma_j^2)$	$\beta_2 \sim N(1.76, \sigma_j^2)$	$\beta_3 \sim N(-3.5, \sigma_j^2)$	
σ_j^2	0.5		10		100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.535(0.084)	(1.385,1.680)	1.548(0.086)	(1.394,1.693)	1.549(0.086)	(1.393,1.694)
Act	1.481(0.094)	(1.313,1.677)	1.486(0.096)	(1.315,1.684)	1.486(0.096)	(1.315,1.686)
Opt	0.932(0.053)	(0.822,1.045)	0.969(0.068)	(0.815,1.117)	0.970(0.068)	(0.814,1.119)
$\hat{\Psi}_{B_P}$	0.603(0.079)	(0.460,0.766)	0.576(0.086)	(0.420,0.767)	0.578(0.086)	(0.419,0.768)
$\hat{\Psi}_{B_T}$	0.549(0.072)	(0.419,0.679)	0.516(0.085)	(0.363,0.652)	0.516(0.086)	(0.363,0.651)

Scenario 2: K-values $K_1 = 0.6$	$K_2 = 3.5$	$K_3 = 2.5$	$K_4 = 1.5$		
μ_{β_j}	$\beta_0 \sim N(-0.50, \sigma_j^2)$	$\beta_1 \sim N(1.76, \sigma_j^2)$	$\beta_2 \sim N(1.42, \sigma_j^2)$	$\beta_3 \sim N(-2.27, \sigma_j^2)$	
σ_j^2	0.5		10		100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.538(0.084)	(1.388,1.681)	1.548(0.086)	(1.395,1.694)	1.549(0.086)	(1.394,1.694)
Act	1.477(0.094)	(1.311,1.675)	1.486(0.096)	(1.315,1.685)	1.486(0.096)	(1.314,1.685)
Opt	0.989(0.054)	(0.877,1.107)	0.970(0.068)	(0.815,1.119)	0.970(0.068)	(0.814,1.119)
$\hat{\Psi}_{B_P}$	0.549(0.079)	(0.401,0.713)	0.578(0.089)	(0.419,0.768)	0.578(0.089)	(0.422,0.768)
$\hat{\Psi}_{B_T}$	0.489(0.072)	(0.359,0.621)	0.515(0.086)	(0.361,0.651)	0.515(0.086)	(0.363,0.651)

Scenario 3: K-values $K_1 = 0.6$	$K_2 = 3.5$	$K_3 = 0.8$	$K_4 = 3.0$		
μ_{β_j}	$\beta_0 \sim N(-0.50, \sigma_j^2)$	$\beta_1 \sim N(1.76, \sigma_j^2)$	$\beta_2 \sim N(0.28, \sigma_j^2)$	$\beta_3 \sim N(-0.44, \sigma_j^2)$	
σ_j^2	0.5		10		100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.556(0.085)	(1.406,1.700)	1.548(0.085)	(1.394,1.693)	1.549(0.085)	(1.394,1.694)
Act	1.459(0.095)	(1.292,1.659)	1.486(0.096)	(1.315,1.685)	1.486(0.095)	(1.314,1.685)
Opt	1.088(0.056)	(0.971,1.211)	0.971(0.068)	(0.815,1.119)	0.970(0.068)	(0.814,1.118)
$\hat{\Psi}_{B_P}$	0.468(0.080)	(0.313,0.637)	0.578(0.089)	(0.419,0.768)	0.578(0.089)	(0.419,0.769)
$\hat{\Psi}_{B_T}$	0.371(0.071)	(0.246,0.499)	0.515(0.086)	(0.362,0.651)	0.516(0.086)	(0.363,0.653)

Table 3.2: Posterior mean, standard error and 95% credible intervals of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ fitting a standard Poisson regression model assuming a $\mathcal{N}(4.8, 3.24)$ biomarker. Data was generated from a standard Poisson model with sample size of 350. Coefficients used for data simulation are: $\beta_0 = -0.10$, $\beta_1 = 0.08$,$\beta_2 = 0.65$ and $\beta_3 = -0.15$.

Scenario 1: K-values	$K_1 = 0.6$	$K_2 = 3.5$	$K_3 = 3.5$	$K_4 = 0.6$	
μ_{β_j}	$\beta_0 \sim N(-0.15, \sigma_j^2)$	$\beta_1 \sim N(0.11, \sigma_j^2)$	$\beta_2 \sim N(1.04, \sigma_j^2)$	$\beta_3 \sim N(-0.22, \sigma_j^2)$	
σ_j^2	0.2		5	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.292(0.071)	(1.156,1.426)	1.334(0.076)	(1.189,1.472)	1.334(0.076)	(1.189,1.471)
Act	1.254(0.071)	(1.122,1.379)	1.263(0.074)	(1.126,1.409)	1.263(0.074)	(1.126,1.410)
Opt	1.081(0.048)	(0.993,1.168)	1.132(0.066)	(1.009,1.244)	1.133(0.066)	(1.009,1.246)
$\hat{\Psi}_{B_P}$	0.211(0.073)	(0.101,0.342)	0.202(0.081)	(0.076,0.338)	0.201(0.081)	(0.078,0.339)
$\hat{\Psi}_{B_T}$	0.173(0.043)	(0.102,0.269)	0.131(0.065)	(0.028,0.273)	0.131(0.065)	(0.028,0.273)

Scenario 2: K-values	$K_1 = 0.6$	$K_2 = 3.0$	$K_3 = 2.5$	$K_4 = 0.6$	
μ_{β_j}	$\beta_0 \sim N(-0.18, \sigma_j^2)$	$\beta_1 \sim N(0.10, \sigma_j^2)$	$\beta_2 \sim N(0.80, \sigma_j^2)$	$\beta_3 \sim N(-0.18, \sigma_j^2)$	
σ_j^2	0.2		5	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.301(0.072)	(1.161,1.436)	1.334(0.076)	(1.191,1.471)	1.334(0.076)	(1.189,1.470)
Act	1.239(0.071)	(1.108,1.365)	1.263(0.075)	(1.126,1.409)	1.263(0.075)	(1.126,1.410)
Opt	1.102(0.049)	(1.011,1.189)	1.132(0.066)	(1.009,1.244)	1.133(0.066)	(1.009,1.244)
$\hat{\Psi}_{B_P}$	0.199(0.076)	(0.087,0.336)	0.202(0.081)	(0.078,0.340)	0.201(0.081)	(0.076,0.339)
$\hat{\Psi}_{B_T}$	0.138(0.041)	(0.072,0.234)	0.131(0.064)	(0.025,0.272)	0.131(0.065)	(0.025,0.273)

Scenario 3: K-values	$K_1 = 0.6$	$K_2 = 3.0$	$K_3 = 0.8$	$K_4 = 2.5$	
μ_{β_j}	$\beta_0 \sim N(-0.18, \sigma_j^2)$	$\beta_1 \sim N(0.10, \sigma_j^2)$	$\beta_2 \sim N(0.19, \sigma_j^2)$	$\beta_3 \sim N(-0.03, \sigma_j^2)$	
σ_j^2	0.2		5	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.329(0.073)	(1.185,1.459)	1.334(0.076)	(1.190,1.472)	1.334(0.076)	(1.190,1.472)
Act	1.208(0.072)	(1.083,1.333)	1.263(0.074)	(1.126,1.409)	1.263(0.075)	(1.126,1.410)
Opt	1.149(0.054)	(1.045,1.241)	1.133(0.065)	(1.010,1.243)	1.133(0.066)	(1.009,1.248)
$\hat{\Psi}_{B_P}$	0.179(0.085)	(0.056,0.334)	0.201(0.081)	(0.076,0.339)	0.201(0.081)	(0.076,0.338)
$\hat{\Psi}_{B_T}$	0.058(0.033)	(0.015,0.147)	0.130(0.065)	(0.029,0.272)	0.131(0.065)	(0.028,0.273)

Table 3.3: Posterior mean, standard error and 95% credible intervals of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ fitting a standard Poisson regression model to mirror the AA clinical trial study. Data was generated from a standard Poisson model with sample size of 460. Coefficients used for data simulation are: $\beta_0 = -9.85$, $\beta_1 = 2.20$,$\beta_2 = 5.33$ and $\beta_3 = -1.52$.

Scenario 1: K-values $K_1 = 0.10$	$K_2 = 2.50$	$K_3 = 1.50$	$K_4 = 0.40$		
μ_{β_j}	$\beta_0 \sim N(-3.72, \sigma_j^2)$	$\beta_1 \sim N(0.78, \sigma_j^2)$	$\beta_2 \sim N(5.01, \sigma_j^2)$	$\beta_3 \sim N(-1.27, \sigma_j^2)$	
σ_j^2	0.5		10	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	0.445(0.043)	(0.359,0.524)	0.462(0.040)	(0.385,0.540)	0.462(0.040)	(0.386,0.541)
Act	0.179(0.029)	(0.122,0.248)	0.205(0.031)	(0.150,0.276)	0.205(0.031)	(0.151,0.276)
Opt	0.169(0.025)	(0.119,0.224)	0.155(0.026)	(0.110,0.219)	0.154(0.027)	(0.109,0.219)
$\hat{\Psi}_{B_P}$	0.276(0.044)	(0.199,0.353)	0.307(0.037)	(0.231,0.377)	0.308(0.037)	(0.231,0.378)
$\hat{\Psi}_{B_T}$	0.011(0.005)	(0.003,0.025)	0.050(0.018)	(0.020,0.092)	0.051(0.019)	(0.021,0.094)

Scenario 2: K-values $K_1 = 0.10$	$K_2 = 2.50$	$K_3 = 1.00$	$K_4 = 0.30$		
μ_{β_j}	$\beta_0 \sim N(-3.72, \sigma_j^2)$	$\beta_1 \sim N(0.78, \sigma_j^2)$	$\beta_2 \sim N(4.24, \sigma_j^2)$	$\beta_3 \sim N(-1.07, \sigma_j^2)$	
σ_j^2	0.5		10	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	0.446(0.043)	(0.359,0.524)	0.462(0.040)	(0.385,0.541)	0.462(0.040)	(0.386,0.540)
Act	0.179(0.029)	(0.121,0.247)	0.205(0.031)	(0.151,0.276)	0.205(0.031)	(0.151,0.277)
Opt	0.172(0.027)	(0.119,0.229)	0.155(0.026)	(0.110,0.219)	0.154(0.027)	(0.109,0.219)
$\hat{\Psi}_{B_P}$	0.274(0.046)	(0.196,0.353)	0.307(0.037)	(0.231,0.377)	0.308(0.037)	(0.231,0.378)
$\hat{\Psi}_{B_T}$	0.007(0.004)	(0.002,0.018)	0.049(0.018)	(0.020,0.092)	0.051(0.019)	(0.020,0.093)

Scenario 3: K-values $K_1 = 0.1$	$K_2 = 2.50$	$K_3 = 0.40$	$K_4 = 1.25$		
μ_{β_j}	$\beta_0 \sim N(-3.72, \sigma_j^2)$	$\beta_1 \sim N(0.78, \sigma_j^2)$	$\beta_2 \sim N(2.30, \sigma_j^2)$	$\beta_3 \sim N(-0.51, \sigma_j^2)$	
σ_j^2	0.5		10	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	0.448(0.045)	(0.356,0.527)	0.462(0.040)	(0.385,0.541)	0.462(0.040)	(0.386,0.540)
Act	0.176(0.029)	(0.118,0.243)	0.205(0.031)	(0.150,0.276)	0.205(0.031)	(0.151,0.276)
Opt	0.175(0.029)	(0.116,0.239)	0.156(0.026)	(0.111,0.220)	0.154(0.027)	(0.109,0.219)
$\hat{\Psi}_{B_P}$	0.273(0.048)	(0.188,0.357)	0.307(0.037)	(0.229,0.376)	0.308(0.037)	(0.231,0.378)
$\hat{\Psi}_{B_T}$	0.001(0.001)	(0.000,0.005)	0.049(0.018)	(0.019,0.091)	0.051(0.019)	(0.021,0.092)

Chapter 4

Sample size estimation for evaluating biomarker clinical utility: a predictive biomarker and binary endpoint

Henok G. Woldu and Kevin K. Dobbin

Submitted to Journal of Computational Statistics and Data Analysis, 02/17/2018.

Abstract

Sample size calculations that supplement the study design for biomarker evaluation are key part of the process. With the recent surge of new biomarker discoveries, statistical methods for assessing the clinical utility of these biomarkers have also been advancing. In the past few years, a metric that measures the decrease in the population event rate under biomarker guided therapy has been advocated as a global predictive biomarker clinical utility measure. However, there has not been a sample size estimation method developed that compliment this metric. In this paper we (1) developed alternative mathematical equations for estimating this metric, compare these to existing estimators, and present the asymptotic properties. (2) Propose a sample size estimation method, Squared Width Inversion Regression Linear (SWIRL) for this metric. Our SWIRL method is used to estimate a sample size n such that the 95% CI mean width of this metric is smaller than a user defined length (W_{targ}). An R program for the sample size calculation is made available.

Keywords : SWIRL ; sample size ; predictive biomarker ; clinical utility

4.1 Introduction

Sample size calculation is key when we develop a study design for biomarker evaluation. For a randomized control trial study design with a binary clinical end point, a metric Θ which measures the decrease in the expected proportion of events under biomarker guided therapy has been been advocated as a measure of biomarker utility performance. However, none of the existing sample size calculation methods can be used to calculate a sample size n that guarantees enough power for the metric Θ in order to make a decision about the biomarker performance. Therefore, a sample size calculation method that supplements Θ and the intended study design is required to enhance biomarker evaluation process.

Swift advancement in genome sequencing is rendering the reality that a patient specific care will be our future treatment model. As such, currently, biomarker discovery, validation, regulatory acceptance and qualification are areas of enormous interest and need (Amur et al., 2015). Predictive biomarkers, for example, are used to enhance drug and biologics development and to guide treatment for patients (Fine and Amler, 2009, Janes et al., 2015, La Thangue and Kerr, 2011, Lavezzari and Womack, 2016).

A number of predictive biomarkers are already guiding therapy for cancer patients. K-RAS mutation status, is being used to pinpoint colorectal cancer patients likely to benefit from Epidermal Growth Factor Receptor (EGFR) inhibitor treatment (Amado et al., 2008, Mehta et al., 2010, Karapetis et al., 2008). Oncotype DX recurrence score also helps to guide whether a patient takes adjuvant chemotherapy or not after breast cancer surgery (Albain et al., 2010, Gluz et al., 2016, Harris et al., 2016). However, parallel to the innovations in biomarkers discovery, statistical methods for evaluating their clinical utility have not kept pace.

Biomarker evaluation has three critical components according to the framework set by the Institute of Medicine (IOM) per the US Food and Drug Administration (FDA) request:

analytical validity, evidentiary qualification and utilization analysis (Ball, et. al 2010). Utilization is a concept of use (COU), where a specific proposed intended use for the biomarker needs to be prespecified. In the drug development plan, a predictive biomarker can be used in selecting patients for phase III studies. Also, in medical settings, a predictive biomarker can be utilized to advise for or against a given treatment. These biomarkers, however, need to be evaluated for whether their use produces a positive net health impact, i.e, quantifying their utilization according to COU.

Often times, biomarker utility qualification, is done by testing a null hypothesis of no biomarker by treatment interaction. (Buyse, 2007, Taube et al., 2009, Freidlin et al., 2010, Tajik et al., 2013). However, though a necessary condition, the interaction test is not sufficient to evaluate biomarker's utility working (Janes et al., 2011, Huang et al., 2012). Two biomarkers (X_1 & X_2), can have the same interaction coefficient but, behave differently in guiding treatment. Besides this, the scale of an interaction coefficient depends on functional form of the model and biomarker measurement unit. This adds another challenge and makes comparing different biomarkers difficult. (Huang et al., 2012). Interaction test, being an indirect measure, is also hard to comprehend by non-statisticians.

Graphical biomarker utility assessment tools, marker-by-treatment predictiveness curves (Janes et al., 2011) and selective impact curve (Song and Pepe, 2004) have been proposed as alternatives to an interaction test. More recently however, Janes et al. (2014a) presented a comprehensive summary of the previous work and proposed a metric (Θ) as a measure of biomarker utility performance. Θ measures the decrease in an unfavorable event rate under marker guided treatment. This measure (Θ) is widely advocated as a global predictive biomarker clinical utility measure (Gunter et al., 2007, Song and Pepe, 2004, Janes et al., 2011, Brinkley et al., 2010, Janes et al., 2014a). With a binary end point, for example, Θ measures the reduction in population event rate under biomarker guided treatment in comparison to the standard (biomarker unguided) treatment. These procedures, however,

assume we already have a data set collected. In pharmaceutical industries, for example, a decision about inclusion of a biomarker as part of a drug development plan has to be made prior to phase III most of the time. With little data at hand, prospectively evaluating clinical utility of a biomarker at that stage becomes impractical.

In this paper, we first develop simple tools to convert clinician inputs to model parameters that would help to evaluate a biomarker's clinical utility. In addition to the clinician inputs, we further assume a distribution of the biomarker under study. As utility evaluation is done at a later biomarker stage, it suffices to make an assumption about its distribution at this stage. Adding the assumption about the biomarker, we develop alternative equations for estimating the metric Θ. Sample size estimation methods that correspond to the study design and metric under consideration are key to the biomarker utility evaluation process. There is no previously developed sample size estimation method for the metric Θ. In this paper we propose a sample size estimation method, Squared Width Inversion Regression Line (SWIRL). Our SWIRL method is used to estimate a sample size n such that the 95% CI mean width of Θ is smaller than the user defined length (W_{targ}).

The rest of this paper is organized as follows: In Section 2 we first set the scenario and briefly introduce the notations. Then mathematical derivation of the marker positivity criteria and development of the equations used to estimate the parameter Θ follow. In Section 3, we present the approach used to change clinician defined inputs to slope parameters. Sample size estimation using our proposed SWIRL approach for two study designs (biomarker stratified design and biomarker strategy design) are briefly discussed in Section 4. Various bootstrap methods we used to estimate the 95% CI width of Θ are presented in Section 5. The paper then concludes in Section 6 with simulation results.

4.2 Methods

4.2.1 Settings and Notations

Let the clinical endpoint or accepted surrogate endpoint of interest be Y, with a known probability distribution and values that can be continuous (survival/relapse time) or binary (cure/death). The focus of this paper is on a clinical endpoint which is binary. In many cancer studies for example, $Y = 1$ would represent death or relapse before time t and $Y = 0$ for a cure or relapse beyond time t. Additionally, let T represent the available treatment: $T = 1$ if a subject is in the active treatment arm and $T = 0$ if in the standard of care (SOC) or placebo arm. Further, let the biomarker of interest be X which is continuous and measured for each subject at baseline. For now, we will assume the biomarker X has a known probability density function $f(x)$.

Given a binary response, it is customary to assume Y has a binomial distribution with success probability π. With this, the natural approach to represent the relationship between the binary clinical endpoint, treatment, biomarker and interaction of treatment and biomarker would be a multiple linear logistic regression. In addition to the common logistic model assumptions, we will further assume in our case: (1) the outcome explains all the impact of the assigned treatment and no other factor has any additional influence on the outcome (Janes et al., 2011). However, if we assume other clinical variables have a potential to influence the outcome, this assumption can be loosened and the method is expanded to accommodate this situation. (2) The Stable Unit Treatment Value Assumption (SUTVA) of Rubin (1986) holds (Brinkley et al., 2010, Rubin, 1974). This assumption states that the value of the potential outcome for a patient does not depend on the treatment assignment of other patients.

4.2.2 Biomarker Guided Treatment Strategy

As stated above, the relationship between the outcome $Y \in \{0, 1\}$ and the covariates (T and X) along with the interaction term ($T * X$) is commonly represented using multiple logistic regression as:

$$Ln\left[\frac{Pr(Y=1|T,X)}{1-Pr(Y=1|T,X)}\right] = \beta_0 + \beta_1 X + \beta_2 T + \beta_3 TX. \tag{4.1}$$

where β_1, β_2 and β_3 in the model represent the biomarker, treatment and biomarker by treatment interaction effects respectively. Now let the biomarker X have a known probability density function given by $f(x)$ where $X \in (-\infty, \infty)$ or narrower depending on the minimum or maximum values of the biomarker under study. By the time we want to evaluate the clinical utility of the biomarker, we will have some information about the distribution of the biomarker, assuming the biomarker has already gone through the initial two stages: validation and qualification. Therefore, making an assumption about the biomarker distribution at this stage is more realistic. Incorporating the biomarker's distributional information, now we can modify the model in equation (4.1) as:

$$Pr(Y=1|T=0) = \int \left\{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right\} f(x)dx \tag{4.2}$$

$$Pr(Y=1|T=1) = \int \left\{\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}}\right\} f(x)dx \tag{4.3}$$

Both equations (4.2) and (4.3) calculate the probability of unfavorable outcome, assuming all subjects are assigned to SOC or active arm respectively. However, our objective is to evaluate the clinical utility of the biomarker. With this, we would want to use the biomarker information to assign available treatment to patients or use that information to recruit patients for phase III clinical trial studies. However, determining a biomarker cut-off point

to make the intended decision is never a straightforward decision. Many factors like cost of the medication, side effects of the medication, etc., have to be taken into consideration. For now, lets assume that the outcome of interest captures all factors which are likely to affect it. Using equation (4.1), it is straightforward to show that:

$$\begin{aligned} \frac{Odds(Y=1|T=1,X=x)}{Odds(Y=1|T=0,X=x)} &= \frac{exp\{\beta_0+\beta_2+(\beta_1+\beta_3)x\}}{exp\{\beta_0+\beta_1 x\}} \\ &= exp\{\beta_2+\beta_3 x\} \end{aligned} \tag{4.4}$$

From equation (4.4) it is clear that, the odds of an unfavorable outcome are greater among subjects in the active arm than the SOC arm if $\beta_2+\beta_3 x > 0$. Based on this, the biomarker guided treatment decision (T_{opt}) can be set in such a way that :

$$T_{opt}(X=x) = \begin{cases} T=1 & : \beta_3 x < (-\beta_2) \\ T=0 & : \beta_3 x \geq (-\beta_2) \end{cases} \tag{4.5}$$

Individuals who respond to a given treatment are generally referred as marker positive and assigned to the active arm ($T=1$) while those who do not are called as marker negative and assigned to SOC ($T=0$). Depending on the sign of the interaction (β_3) coefficient a threshold for the marker guided therapy can be written as:

$$T_{opt}(X=x) \Rightarrow \texttt{if} : \beta_3 < 0 \begin{cases} T=1: & x > \frac{-\beta_2}{\beta_3} \\ T=0: & x \leq \frac{-\beta_2}{\beta_3} \end{cases}$$

$$T_{opt}(X=x) \Rightarrow \texttt{if} : \beta_3 > 0 \begin{cases} T=1: & x < \frac{-\beta_2}{\beta_3} \\ T=0: & x \geq \frac{-\beta_2}{\beta_3} \end{cases}$$

This way of specifying the biomarker threshold value used for guiding treatment matches the one used by Song and Pepe (2004), Brinkley et al. (2010), and Janes et al (2014a). Let

$\mathcal{A}_1 = \{x : \beta_3 x < (-\beta_2)\}$, and $\mathcal{A}_0 = \Re^1 \backslash \mathcal{A}_1$ (where "\" is the set difference symbol). The limit values for $\mathcal{A}_1$ & $\mathcal{A}_0$ will vary depending on the support of the probability density function for X. For $X \sim \mathcal{U}(0,1)$ a table of all possible combinations is presented in Appendix A.0.2. Once the biomarker cut-off point is determined, the probability of unfavorable outcome under biomarker guided treatment is shown to be:

$$\begin{aligned} Pr(Y=1|T_{opt}) &= \int_{\mathcal{A}_1} \left\{ \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}} \right\} f(x)dx + \\ & \int_{\mathcal{A}_0} \left\{ \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \right\} f(x)dx. \end{aligned} \tag{4.6}$$

Under the expression in equation (4.6), a subject will be treated if he/she is marker positive and will skip treatment if marker negative.

4.2.3 Development of Equations for Estimating Θ

Biomarker ulitility evaluation is a COU. With intended purpose of using the biomarker for guiding treatment selection, let the parameter of interest be Θ as proposed by Janes et. al (2014). With a binary clinical endpoint, Θ measures the average decrease in the population unfavorable event rate under biomarker guided treatment. The biomarker unguided treatment will be either "Treat All" or "Treat None" depending on the current default treatment. We will use Θ_1 if current default treatment is "Treat All" i.e $T = 1$ and Θ_0 when default treatment is "Treat None" i.e $T = 0$. Then,

$$\begin{aligned} \Theta_1 &= Pr(Y=1|T=1) - Pr(Y=1|T_{opt}) \\ &= \int_{\mathcal{A}_0} \left\{ \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}} - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \right\} f(x)dx \end{aligned} \tag{4.7}$$

and

$$\begin{aligned}\Theta_0 &= Pr(Y=1|T=0) - Pr(Y=1|T_{opt}) \\ &= \int_{\mathcal{A}_1} \left\{ \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 y}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}} \right\} f(x)dx \end{aligned} \tag{4.8}$$

Now let's assume the biomarker X has a uniform distribution on the unit interval $[0,1]$. If the original biomarker value is not uniformly distributed, we can transform X to $F_{x^*}(x^*)$ where F_{x^*} is the cumulative distribution function (CDF) of the transformed variable X. Further assume the current default treatment (standard care) is "Treat None", and $\mathcal{A}_1 \in [a,b]$, then Θ_0 will be:

$$\begin{aligned}\Theta_0 &= \int_a^b \left\{ \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}} \right\} f(x)dx \\ &= \begin{cases} (b-a)\frac{e^{\beta_0}(1-e^{\beta_2})}{(1+e^{\beta_0+\beta_2})(1+e^{\beta_0})} & : \beta_1 = 0, \beta_3 = 0 \\ (b-a)\frac{e^{\beta_0}}{1+e^{\beta_0}} - \frac{1}{\beta_3} ln[\frac{1+e^{\beta_0+\beta_2+\beta_3 b}}{1+e^{\beta_0+\beta_2+\beta_3 a}}] & : \beta_1 = 0, \beta_3 \neq 0 \\ \frac{1}{\beta_1} ln[\frac{1+e^{\beta_0+\beta_1 b}}{1+e^{\beta_0+\beta_1 a}}] - (b-a)\frac{e^{\beta_0+\beta_2}}{1+e^{\beta_0+\beta_2}} & : \beta_1 + \beta_3 = 0 \\ \frac{1}{\beta_1} ln[\frac{1+e^{\beta_0+\beta_1 b}}{1+e^{\beta_0+\beta_1 a}}] - \frac{1}{\beta_1+\beta_3} ln[\frac{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)b}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)a}}] & : otherwise \end{cases} \end{aligned} \tag{4.9}$$

If the current default treatment is "Treat All", and $\mathcal{A}_0 \in [c,d]$,then

$$\begin{aligned}\Theta_1 &= Pr(Y=1|T=1) - Pr(Y=1|T_{opt}) \\ &= \int_{\mathcal{A}_0} \left\{ \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}} - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \right\} f(x)dx \end{aligned} \tag{4.10}$$

$$= \begin{cases} (d-c)\frac{e^{\beta_0}(e^{\beta_2}-1)}{(1+e^{\beta_0+\beta_2})(1+e^{\beta_0})} & : \beta_1 = 0, \beta_3 = 0 \\ \frac{1}{\beta_3} ln(\frac{1+e^{\beta_0+\beta_2+\beta_3 d}}{1+e^{\beta_0+\beta_2+\beta_3 c}}) & : \beta_1 = 0, \beta_3 \neq 0 \\ (d-c)\frac{e^{\beta_0+\beta_2}}{1+e^{\beta_0+\beta_2}} - \frac{1}{\beta_1} ln(\frac{1+e^{\beta_0+\beta_1 d}}{1+e^{\beta_0+\beta_1 c}}) & : \beta_1 + \beta_3 \neq 0 \\ \frac{1}{\beta_1+\beta_3} ln(\frac{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)d}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)c}}) - \frac{1}{\beta_1} Ln(\frac{1+e^{\beta_0+\beta_1 d}}{1+e^{\beta_0+\beta_1 c}}) & : otherwise \end{cases}$$

In case we have a normally distributed biomarker with mean μ and variance σ^2 and the limit integrals defined as $\mathcal{A}_1 \in [a, b]$ and $\mathcal{A}_0 \in [c, d]$, the equations for estimating Θ_0 and Θ_1 are respectively given as:

$$\Theta_0 = \int_a^b \left\{ \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}} \right\} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2} dx$$
$$\Theta_1 = \int_c^d \left\{ \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}} - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \right\} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

There is no analytical solution when the biomarker distribution is normal and numerical integration is used to get the estimated value of $\Theta_0(\Theta_1)$.

4.2.4 Monte Carlo Evaluation of the Equations for Θ

To ensure the exactness of the formulas we used in section 4.2.3 to calculate $\Theta_0(\Theta_1)$, a large Monte Carlo simulation was used. The estimates for the probabilities of unfavorable outcomes under a default treatment ("Treat All" or "Treat None") and biomarker guided treatment were obtained as: $Pr(Y = 1|T) = \frac{\#(Y=1,T)}{\#(T)}$, where # represents for the number of patients.

4.3 From Clinician Inputs to Model Parameters

The two stage procedure of estimating Θ we outlined above, assume that we already have data. In that case, estimating Θ will be straightforward using the equations we developed. First fit the multiple logistic model stated, take the estimated coefficients, plug them in equations (9) and (10), assuming the biomarker has a uniform distribution and get an estimated value of Θ (Song and Pepe, 2004, Brinkley et al., 2010, Janes et al., 2011). However, when we don't have any data or have only a little data at hand, setting the slope parameters is not intuitive. In drug development for example, a decision about a biomarker has to be made prior to phase III or sometimes phase II. During these stages, the data we have is mainly from preclinical and early phase studies. However, such data is barely enough to fit a model and help us to make a go-no-go decision about the biomarker.

Instead of being stranded by the absence of data, we can use clinician's best guess about the likely outcome of the disease under study and biomarker's distributional assumption information to move forward. Let the clinician give us the expected proportion of unfavorable outcomes K_1, K_2, K_3 and K_4 given the 25^{th} and 75^{th} percentiles value of the biomarker for the SOC and active arms respectively. Specifying the $K's$, requires clinicians' best judgment taking the disease and study drug in to consideration. Setting K_1, K_2 will be intuitive as they are related to the existing treatment. In drug development, K_3 and K_4 will be the clinicians' best guess about the performance of the new drug in pipeline as compared to existing drug. Given this information, we developed equations to convert the clinician inputs to model parameters. These model parameters are in turn used to generate the data and proceed to estimate Θ. Detailed derivation of this is given in Appendix A.0.3. The final closed form expression for the parameter values are given below:

$$\beta_0 = \frac{K_2 * z_1 - K_1 * z_2}{z_1 - z_2}$$

$$\beta_1 = \frac{K_1 - K_2}{z_1 - z_2}$$
$$\beta_2 = \frac{K_1 * z_2 - K_2 * z_2 - K_3 * z_2 + K_4 * z_1}{z_1 - z_2}$$
$$\beta_3 = \frac{K_2 - K_1 + K_3 - K_4}{z_1 - z_2}$$

where $K_1 = Ln\left[\frac{P(Y=1|A=0,x=F^{-1}(0.25))}{1-P(Y=1|A=0,x=F^{-1}(0.25))}\right]$ $K_2 = Ln\left[\frac{P(Y=1|A=0,x=F^{-1}(0.75))}{1-P(Y=1|A=0,x=F^{-1}(0.75))}\right]$

$K_3 = Ln\left[\frac{P(Y=1|A=1,x=F^{-1}(0.25))}{1-P(Y=1|A=1,x=F^{-1}(0.25))}\right]$ $K_4 = Ln\left[\frac{P(Y=1|A=1,x=F^{-1}(0.75))}{1-P(Y=1|A=1,x=F^{-1}(0.75))}\right]$

$z_1 = F^{-1}(0.25)$ and $z_2 = F^{-1}(0.75)$

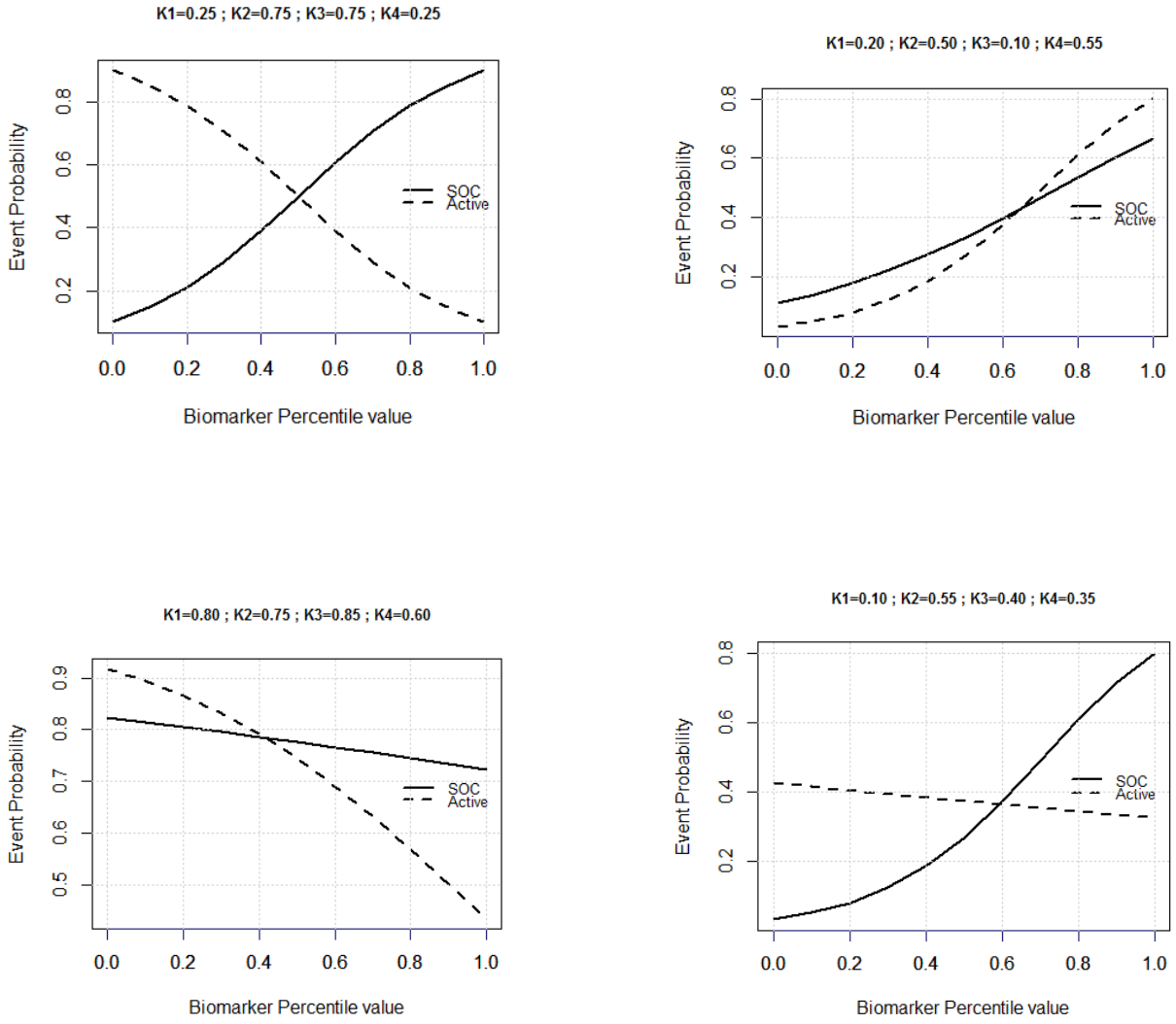

Figure 4.1: Risk curves that correspond each treatment arm for a given clinician input values.

4.4 SWIRL Sample Size Estimation Method

Existing logistic sample size estimation methods (Whittemore, 1981, Hsieh, 1989, Demidenko, 2007; 2008) are not applicable to estimate sample size needed to ensure adequate power for the metric Θ. Primarily, the functional form of Θ is different from the logistic

regression model of equation (4.1) with which we start. Secondly, the existing methods could estimate sample size needed to guarantee the test for $\beta_3 \neq 0$ with enough power. However, a test of $\beta_3 \neq 0$ does not insure a test for $\Theta \neq 0$ (Janes et al., 2014a). In this paper we propose the squared width inversion regression linear (SWIRL) sample size estimation method for the metric Θ.

With SWIRL method, a sample size n is chosen, such that, the 95% CI mean width of Θ is smaller than a user defined target length (w_{targ}). The SWIRL method is easy to implement and depends on the assumption that, there is a very strong linear relationship between sample size n and inverse of the 95% CI mean width squared for the parameter Θ. We have developed the asymptotic properties of Θ, from which this assumption can be justified. Additional simulation studies were conducted to investigate this relationship for three distributions (uniform, normal and gamma) and did not encounter any violation. Details of the asymptotic development of Θ are provided in Appendix B.0.1. Plots showing the linear relationship between n and inverse of the 95% CI mean width squared for Θ are also provided in the supplementary material.

Our proposed method requires a Monte Carlo simulation of m trials with sample size ranging from $n_1, n_2, ...n_m$ in increments of c to get a 95% CI mean width (W) that correspond with each sample size. Then an ordinary least square regression is fitted as:

$$n = \alpha_0 + \alpha_1 W^{-2} \tag{4.11}$$

Finally the sample size that guarantees the 95% CI mean width of the parameter $\Theta_0(\Theta_1)$ to be less than a user defined target (W_{targ}) is estimated as:

$$\hat{n} = \hat{\alpha_0} + \hat{\alpha_1} W_{targ}^{-2} \tag{4.12}$$

where $\hat{\alpha_0}$ and $\hat{\alpha_1}$ are the OLS estimates of equation (4.11). The steps required for imple-

menting the SWIRL method are summarized in Algorithm 3 below.

Algorithm 3 SWIRL Sample Size Estimation Method

1: For $K_1, ..., K_4$ calculate the values of $\beta_0, ..., \beta_3$.
2: For $i = 1, ..., m$ generate $X_{i,1}, ..., X_{i,n_m}$ iid r.vs from a distribution with pdf $f(x)$.
3: For $i = 1, ..., m$, sample at random $n/2$ integers from $1, ..., n$, calling this set T_1 and assign all a value (=1) and the rest T_0 assign a value (=0).
4: Simulate **m** trials using Monte Carlo Simulation with sample size ranging from $n_1, n_2, ...n_m$ in increment of **c** using the model

$$Ln\left[\frac{Pr(Y=1|T,X)}{1-Pr(Y=1|T,X)}\right] = \beta_0 + \ \beta_1 X + \beta_2 T + \beta_3 TX$$

5: For $i = 1, ..., m$ Calculate $\hat{\Theta}_i$ and the 95% CI mean width (W_i) associated with $\hat{\Theta}_i$
6: Using Ordinary Least Square Regression fit :

$$n = \alpha_0 + \alpha_1 W^{-2}$$

7: For user defined length (W_{targ}), the sample size n is then estimated as :

$$\hat{n} = \hat{\alpha_0} + \hat{\alpha_1} W_{targ}^{-2}$$

8: If $\hat{n}$ falls outside the range of $n_1, ..., n_m$, then go to step 1 and add another simulation instead by increasing the sample size such that $n_m = 2 * \hat{n}$.

4.5 Bootstrap for Estimating the 95% CI Width of Θ

Our SWIRL sample size estimation method requires estimating the 95% CI mean width of Θ as first step. For this purpose and assessing the coverage probability of our method of estimating Θ, we used the normal and percentile bootstrap confidence interval methods. Here we assume the current default treatment is “treat all", but this procedure can be extended in a similar manner if current default treatment is “treat none". Let Θ_1^* be the bootstrap estimates. The first bootstrap we used is based on the assumption that our estimated parameter $\widehat{\Theta}_1$ has a normal distribution. Then the $(1-\alpha)100\%$ confidence interval

is evaluated as:

$$CI = \left(\widehat{\Theta}_1 \pm Z_{\frac{\alpha}{2}} * \widehat{se}_{\theta_1^*}\right)$$

Percentile bootrap, which is very similar to the the basic bootstrap but uses percentiles of the bootstrap distribution and a different formula, is the second bootstrap method considered. Davison and Hinkley (1997, equ. 5.18 p. 203) and Efron and Tibshirani (1993, equation 13.5 p. 171) could be referred for further explanation of this procedure. The confidence interval formula is:

$$CI = \left(\theta^*_{1\,\frac{\alpha}{2}}, \theta^*_{1(1-\frac{\alpha}{2})}\right)$$

where θ^* is the p^{th} percentile of the bootstrap estimates.

Results of these confidence interval for $\widehat{\Theta}$ are compared with results from treatment selection R package of Janes (2014) . In terms of coverage probability both methods work almost equally. However the mean width of the confidence interval in our method is slightly narrower. This can be attributed mainly due the fact that our method of estimating Θ assumes a distribution about the biomarker.

4.6 Simulation Results

Monte Carlo simulation and methods from Janes et al. (2014) were used to ensure our proposed formulas to calculate Θ_0 and Θ_1 are valid. In Table 4.1, simulation results with sample size of 100,000 are presented for $\mathcal{U}(0,1)$ and $\mathcal{N}(0,1)$ distributed biomarkers. We considered different combinations of clinician input K values. Results obtained using our developed formulas agree with those of the simulation and Jane's regardless of the biomarker distribution. Further investigation was done assuming a gamma(2,2) distribution for the

biomarker and all results match each other (results not presented). The simplicity of our approach is that, depending on the clinician inputs $K_1 - K_4$ values, one can proceed to calculate the clinical utility of the biomarker even in the absence of data.

Tables 4.2 and Table S7 (supplementary) depict the 95% confidence interval (CI), coverage probability and width of the CI for the estimator Θ_1 assuming $\mathcal{U}(0,1)$ and $\mathcal{N}(0,1)$ distributed biomarkers respectively. We used the two bootstrap techniques explained in section 5 for our method and compared these with the corresponding results of Janes empirical and model based percentile bootstrap methods. In Table 4.2, results for uniformly distributed biomarker and five different K combination values are presented. Output from the two bootstrap estimate of Θ_1 using our methods and Janes empirical and model based match in their coverage probability, almost all being $\geq 94\%$, except in one where the coverage is 93%. Generally, the Janes method CI mean width is slightly larger than ours. Similar results are shown in Table S8 (supplementary) for normally distributed biomarker and for Θ_0 (result not shown here).

Table 4.3 shows the sample size estimates for $\mathcal{U}(0,1)$ and $\mathcal{N}(0,1)$ distributed biomarkers, respectively, using our proposed SWIRL method for two biomarker study designs: biomarker stratified and biomarker strategy designs. Details about these two study designs are in the paper published by (Freidlin et al., 2010). As expected the estimated sample size is larger for biomarker strategy design than for biomarker stratified design. Monte Carlo simulations were further performed to check how closely the estimated $\hat{n}$ was able to achieve the desired W_{targ}. As the results show, W_{targ} was achieved with high precision in almost all cases regardless of the study design and biomarker distribution.

To assess robustness, we looked at how a change in the assumed link function of the model affects the sample size results. For given clinician input values K_1 to K_4, estimated sample size results under probit and logit link functions were similar. Specifically, when the estimated sample size $\hat{n}$ is within the range of the sample sizes used in the simulation

to estimate the 95% CI widths results are adequate. However, when the estimated sample size falls out of the initial range ($> n_m$), the results deviate a little. This could be mainly due to extrapolation. In such a scenario, as per Algorithm 3, the program should be rerun by increasing the range of the initial sample size. The effect on biomarker distributional assumption can be assessed by looking at the results in Table 4.3. The sample size estimates obtained using SWIRL under a standard uniform and standard normal is relatively similar except in one of our simulations. Further we investigate how deviation from the normality assumption for the biomarker distribution affects the sample size estimate . A moderate skewness overall has a small effect on the estimated sample size. However, for a highly skewed distribution, it is recommended that one use an appropriate transformation before using the SWIRL method. The Janes method of estimating Θ on the other hand does not depend on the marginal distribution of the biomarker.

4.7 Discussion

Development of statistical methods used for evaluating the clinical utility of predictive biomarkers is of great interest. A metric (Θ) which measures the decrease in the population event rate under biomarker guided therapy has been advocated in recent years. In this paper we developed alternative mathematical equations for estimating Θ, compare these to existing estimators, proposed a sample size estimation method for Θ and provided the accompanying computer program to perform the sample size estimation. The sample size estimation method, Squared Width Inversion Regression Linear (SWIRL), is used to estimate a sample size n such that the 95% CI mean width of Θ is less than a user defined length(W_{targ}). An R code used to implement the SWIRL method is made available with this publication. Additionally, the asymptotic results of Θ were provided which in turn are used to guarantee the linear relationship between sample size and the 95% CI mean width

assumed under the SWIRL method. Through simulation, we found this method to work well and provide a proper sample size for Θ for a given 95% CI target width (W_{targ}).

Previous predictive biomarker clinical utility performance evaluation methods assume that data needed for fitting the proposed model is available (Song and Pepe, 2004, Janes et al., 2011, Brinkley et al., 2010, Janes et al., 2014a). However, in many circumstances (like drug development plan and sample size planning), biomarker evaluation has to be performed prior to a stage where data were made available. Additionally these methods do not take the biomarker's probability distribution into consideration and fail to incorporate this additional information in the process. From a design perspective, it suffices to make an assumption about the biomarker distribution and use this information in designing the study. In this paper, we developed simple tools to convert clinician inputs to model parameters, and used to estimate Θ and the sample size. Our focus in this paper was on sample size estimation for the parameter Θ that guarantee the 95% CI mean width. Such criterion has been previously used for sample size determination (Zou, 2012, Dobbin and Ionan, 2015).

In this paper, we have focused on two biomarker study designs: biomarker stratified and biomarker strategy designs (Freidlin et al., 2010) and two biomarker distributions, uniform and normal. However, all the methods developed here could easily be extended to accommodate other study designs and biomarker distributions. Getting closed form equations for estimating Θ might not be always possible so we have to use numerical integration with integral limits subject of the support of the biomarker distribution under consideration.

The procedures used for estimating the sample size using SWIRL in this paper takes the marginal distribution of the biomarker into consideration. Therefore, when the underlying assumptions about the biomarker distribution are violated, the 95% CI mean width might not be the correct mean width. In such a scenario, we recommend (1)appropriate transformation to be performed or (2) use the Janes method to get the 95% CI mean width first and then use the SWIRL method.

Acknowledgment

Research reported in this publication was supported by the National Institutes of Health grant 1R21CA201207.

Table 4.1: Comparison of Θ_0 and Θ_1 using Our formula , Janes (2014) method and Monte Carlo simulation. Sample size used, $n = 100,000$

Biomarker has standard Uniform distribution								
K_1	K_2	K_3	K_4	**Estim.**	**Our Formula**	**Janes Emp.**	**Janes Mod.**	**MC Sim**
0.25	0.75	0.75	0.25	Θ_0	0.232	0.221	0.221	0.221
				Θ_1	0.233	0.227	0.231	0.228
0.10	0.90	0.90	0.10	Θ_0	0.345	0.341	0.341	0.341
				Θ_1	0.346	0.342	0.344	0.342
0.10	0.55	0.90	0.45	Θ_0	0.095	0.097	0.097	0.099
				Θ_1	0.433	0.434	0.431	0.436
0.25	0.75	0.50	0.50	Θ_0	0.116	0.109	0.109	0.109
				Θ_1	0.116	0.111	0.113	0.112
0.90	0.45	0.10	0.55	Θ_0	0.433	0.433	0.433	0.436
				Θ_1	0.095	0.096	0.094	0.099
Biomarker has standard Normal distribution								
K_1	K_2	K_3	K_4	**Estim.**	**Our Formula**	**Janes Emp**	**Janes Mod**	**MC Sim**
0.25	0.75	0.75	0.25	Θ_0	0.245	0.25	0.25	0.249
				Θ_1	0.248	0.250	0.248	0.248
0.10	0.90	0.90	0.10	Θ_0	0.345	0.355	0.355	0.353
				Θ_1	0.348	0.352	0.352	0.349
0.10	0.55	0.90	0.45	Θ_0	0.125	0.130	0.130	0.127
				Θ_1	0.441	0.439	0.438	0.437
0.25	0.75	0.50	0.50	Θ_0	0.123	0.129	0.129	0.128
				Θ_1	0.124	0.124	0.119	0.124
0.90	0.45	0.10	0.55	Θ_0	0.441	0.439	0.439	0.437
				Θ_1	0.125	0.130	0.129	0.127

Table 4.2: Confidence interval width and coverage probability comparison for our method and Janes method using bootstrap: 1000 Monte Carlo each with 1000 sample size. Biomarker has a standard uniform distribution U(0,1).

K_1	K_2	K_3	K_4	Θ_1	**Method**	**95% CI**	**CI Width**	**Coverage**
0.25	0.75	0.75	0.25	0.233	Boot Norm.	(0.196 , 0.265)	0.071	0.940
					Boot Perc.	(0.196 , 0.266)	0.071	0.950
					Janes Emp.	(0.188 , 0.272)	0.083	0.960
					Janes Mod.	(0.193 , 0.271)	0.077	0.950
0.10	0.90	0.90	0.10	0.347	Boot Norm.	(0.317 , 0.370)	0.054	0.970
					Boot Perc.	(0.318 , 0.372)	0.054	0.960
					Janes Emp.	(0.306 , 0.382)	0.077	0.940
					Janes Mod.	(0.309 , 0.381)	0.072	0.970
0.10	0.55	0.90	0.45	0.433	Boot Norm.	(0.393 , 0.464)	0.072	0.970
					Boot Perc.	(0.394 , 0.467)	0.072	0.980
					Janes Emp.	(0.386 , 0.476)	0.090	0.970
					Janes Mod.	(0.389 , 0.473)	0.084	0.970
0.25	0.75	0.50	0.50	0.117	Boot Norm.	(0.078 , 0.153)	0.076	0.930
					Boot Perc.	(0.079 , 0.155)	0.076	0.950
					Janes Emp.	(0.073 , 0.157)	0.084	0.950
					Janes Mod.	(0.079 , 0.156)	0.077	0.930
0.90	0.45	0.10	0.55	0.095	Boot Norm.	(0.070 , 0.125)	0.056	0.950
					Boot Perc.	(0.072 , 0.127)	0.056	0.950
					Janes Emp.	(0.067 , 0.133)	0.067	0.970
					Janes Mod.	(0.070 , 0.128)	0.058	0.980

Table 4.3: Sample size estimation using SWIRL method for biomarker stratified and biomarker strategy designs. Biomaker has $U(0,1)$ and $N(0,1)$ distributions.

Biomarker stratified design					Uniform (0,1)		Normal (0,1)	
K_1	K_2	K_3	K_4	W_{targ}	**SWIRL** $\hat{n}$	**95% CI** $\hat{n}$	**SWIRL** $\hat{n}$	**95% CI** $\hat{n}$
0.25	0.75	0.75	0.25	0.20	119	(109 , 129)	114	(104 , 125)
				0.15	215	(207 , 223)	197	(188 , 207)
				0.10	489	(484 , 494)	436	(430 , 442)
0.10	0.90	0.90	0.10	0.20	64	(50 , 78)	80	(71 , 89)
				0.15	120	(107 , 133)	132	(124 , 140)
				0.10	280	(270 , 290)	281	(275 , 288)
0.10	0.55	0.90	0.45	0.20	127	(119 , 135)	129	(120 , 138)
				0.15	230	(224 , 238)	230	(222 , 237)
				0.10	527	(522 , 531)	516	(511 , 521)
0.25	0.75	0.50	0.50	0.20	142	(132 , 151)	133	(120 , 145)
				0.15	257	(249 , 264)	236	(226 , 246)
				0.10	584	(579 , 589)	533	(526 , 539)
0.90	0.45	0.10	0.55	0.20	78	(63 , 94)	58	(43 , 72)
				0.15	137	(122, 151)	106	(92 , 119)
				0.10	303	(292 , 313)	242	(231 , 253)
0.60	0.40	0.50	0.50	0.20	340	(301 ,378)	130	(97 , 161)
				0.15	477	(446 , 508)	246	(220 , 272)
				0.10	870	(823 , 917)	580	(562 , 597)

Table 4.4: Sample size estimation using SWIRL method for biomarker stratified and biomarker strategy designs. Biomaker has $U(0,1)$ and $N(0,1)$ distributions.

Biomarker Strategy Design					Uniform (0,1)		Normal (0,1)	
K_1	K_2	K_3	K_4	W_{targ}	**SWIRL** $\hat{n}$	**95% CI** $\hat{n}$	**SWIRL** $\hat{n}$	**95% CI** $\hat{n}$
0.25	0.75	0.75	0.25	0.20	166	(154 , 179)	152	(140 , 165)
				0.15	293	(283 , 303)	262	(252 , 273)
				0.10	655	(647 , 663)	557	(570 , 584)
0.10	0.90	0.90	0.10	0.20	90	(68 , 113)	108	(93 , 123)
				0.15	166	(146 , 186)	176	(162 , 188)
				0.10	382	(368 , 395)	369	(359 , 378)
0.10	0.55	0.90	0.45	0.20	181	(165 , 197)	175	(166 , 184)
				0.15	318	(305 , 330)	307	(300 , 314)
				0.10	709	(698 , 719)	684	(678 , 689)
0.25	0.75	0.50	0.50	0.20	187	(168 , 206)	160	(145 , 174)
				0.15	325	(310 , 340)	284	(272 , 296)
				0.10	719	(706 , 733)	637	(628 , 646)
0.90	0.45	0.10	0.55	0.20	124	(103 , 144)	84	(69 , 99)
				0.15	198	(180 , 216)	147	(134 , 160)
				0.10	412	(399 , 424)	326	(316 , 336)
0.60	0.40	0.50	0.50	0.20	469	(429 , 509)	167	(136 , 199)
				0.15	707	(662 , 751)	318	(294 , 342)
				0.10	1386	(1258 , 1513)	747	(724 , 769)

Table 4.5: Monte Carlo evaluation of the SWIRL sample size estimation method.

Biomarker stratified design. Biomarker has Uniform (0,1) distribution								
K_1	K_2	K_3	K_4	**SWIRL** $\hat{n}$	W_{targ}	**Est.Width**	**Width Range**	**Width IQR**
0.25	0.75	0.75	0.25	119	0.20	0.201	(0.154 , 0.262)	(0.191 , 0.213)
				215	0.15	0.151	(0.129 , 0.176)	(0.144 , 0.159)
				489	0.10	0.101	(0.084 , 0.117)	(0.095 , 0.105)
0.10	0.90	0.90	0.10	64	0.20	0.209	(0.129 , 0.319)	(0.175 , 0.236)
				120	0.15	0.154	(0.091 , 0.209)	(0.140 , 0.167)
				280	0.10	0.101	(0.074 , 0.131)	(0.093 , 0.107)
0.10	0.55	0.90	0.45	127	0.20	0.202	(0.154 , 0.271)	(0.182 , 0.218)
				230	0.15	0.154	(0.119 , 0.193)	(0.144 , 0.166)
				527	0.10	0.101	(0.084 , 0.117)	(0.096 , 0.105)
0.25	0.75	0.50	0.50	142	0.20	0.205	(0.121 , 0.309)	(0.186 , 0.222)
				257	0.15	0.151	(0.104 , 0.190)	(0.142 , 0.163)
				584	0.10	0.100	(0.077 , 0.119)	(0.096 , 0.104)
0.90	0.45	0.10	0.55	78	0.20	0.286	(0.122 , 0.913)	(0.187 , 0.248)
				137	0.15	0.156	(0.105 , 0.696)	(0.134 , 0.159)
				303	0.10	0.104	(0.074 , 0.263)	(0.093 , 0.107)
0.60	0.40	0.50	0.50	340	0.20	0.214	(0.102 , 0.567)	(0.159 , 0.258)
				477	0.15	0.166	(0.091 , 0.379)	(0.123 , 0.193)
				870	0.10	0.097	(0.067 , 0.183)	(0.082 , 0.108)

Table 4.6: Monte Carlo evaluation of the SWIRL sample size estimation method.

Biomarker strategy design. Biomarker has Normal (0,1) Distribution								
K_1	K_2	K_3	K_4	**SWIRL** $\hat{n}$	W_{targ}	**Est.Width**	**Width Range**	**Width IQR**
0.25	0.75	0.75	0.25	152	0.20	0.201	(0.140 , 0.338(	(0.180 , 0.219)
				262	0.15	0.150	(0.116 , 0.197)	(0.138 , 0.160)
				557	0.10	0.103	(0.086 , 0.125)	(0.096 , 0.108)
0.10	0.90	0.90	0.10	108	0.20	0.190	(0.100 , 0.397)	(0.152 , 0.213)
				176	0.15	0.149	(0.092 , 0.209)	(0.129 , 0.160)
				369	0.10	0.098	(0.072 ,0.131)	(0.088 , 0.107)
0.10	0.55	0.90	0.45	175	0.20	0.206	(0.139 , 0.291)	(0.179 , 0.225)
				307	0.15	0.149	(0.112 , 0.189)	(0.138 , 0.159)
				684	0.10	0.100	(0.081 , 0.125)	(0.094 , 0.106)
0.25	0.75	0.50	0.50	160	0.20	0.200	(0.084 , 0.311)	(0.176 , 0.226)
				284	0.15	0.149	(0.108 , 0.209)	(0.136 , 0.159)
				637	0.10	0.099	(0.075 , 0.134)	(0.092 , 0.105)
0.90	0.45	0.10	0.55	84	0.20	0.203	(0.094 , 0.385)	(0.173 , 0.228)
				147	0.15	0.146	(0.101 , 0.207)	(0.129 , 0.158)
				326	0.10	0.101	(0.082 , 0.127)	(0.093 , 0.107)
0.60	0.40	0.50	0.50	167	0.20	0.205	(0.092 , 0.335)	(0.167 , 0.246)
				318	0.15	0.150	(0.066 , 0.227)	(0.127 , 0.172)
				747	0.10	0.102	(0.044 , 0.291)	(0.090 , 0.113)

Chapter 5

Reproducibility Metric for Evaluating Clinical Utility of Treatment Selection Biomarkers

Henok G. Woldu and Kevin K. Dobbin

To be submitted to Journal of Biostatistics.

Abstract

An originally validated predictive biomarker often undergoes a modification stage for numerous reasons. This nullifies previous outcome-biomarker relationship studies and mandates researchers to repeat the process. However, this is costly and time consuming and leads many initially promising biomarkers into a dead end. In this paper, we propose a reproducibility metric Δ_r that measures the difference in clinical performance between the original biomarker and the modified biomarker. This metric does not require that one observe the outcome associated with the modified biomarker and makes the evaluation process easy and less expensive. Proofs for the asymptotic results of Δ_r are provided. Monte Carlo methods are used to construct 95% CI for Δ_r. Ki67 reproducibility data is used to show its application. An R package RMPB is made available via Github for its implementation.

Keywords : reproducibility; Monte Carlo; ICC; CCC ; Rao-Blackwellization

5.1 Introduction

Initially validated predictive biomarker are often modified in the middle of the development stage. Platform migration, addition of laboratories, cost reduction, sample preparation simplification, and change in reagents are a few among the many reasons that result in biomarker modification. Biomarker modification however, causes all the earlier biomarker clinical performance studies to be invalid. To evaluate the clinical performance of the modified biomarker researchers must "make a fresh start", but this process is hardly feasible since it needs repeating the study which is time consuming and very expensive. This puts many initially promising biomarkers in a dead end. To the best of our understanding, currently there are no statistical methods to assess the impact of biomarker modification on patient outcome.

Predictive biomarkers also called treatment selection biomarkers are used to identify a subgroups of patients who are more likely to respond to a given treatment (Sargent and Allegra, 2002, Simon and Maitournam, 2004, Simon, 2008). Once the clinical utility of a predictive biomarker is validated, it can help physicians for recommending the best treatment for patients thereby improving the health of a patient. Among colorectal cancer patients for example, the KRAS status of a patient is used to identify whether the patient will benefit from Epidermal growth factor receptor (EGFR) inhibitor treatment or not (Amado et al., 2008, Mehta et al., 2010). Similarly, the OnctotypeDX assay is used to recommend if chemotherapy will benefit them after breast cancer surgery (Harris et al., 2016). However, these biomarkers need biostatical methods used to evaluate them more easily and quickly before they can be used in the clinical setting for making treatment decisions.

For a binary clinical endpoint, the relationship between a predictive biomarker and outcome is assessed using metrics like sensitivity, specificity, negative and positive predictive values and the area under receiver operating characteristic (ROC) curve (AUC) (Søreide,

2009, Bharti and Bharti, 2009). Biomarker by treatment interaction has also been a commonly used metric for assessing predictive biomarker performance (Byar, 1985, Buyse, 2007, Taube et al., 2009, Freidlin et al., 2010, Tajik et al., 2013). However, none of these have a clear clinical interpretation in terms of the health gain for a patient. To close this gap, a metric Θ, which measures the decrease in the expected event rate that results from a biomarker guided treatment was developed (Song and Pepe, 2004, Vickers et al., 2007, Brinkley et al., 2010, Janes et al., 2014a).

A predictive biomarker whose clinical utility was initially validated using the metric Θ, often goes through modification for the aforementioned reasons and others in the middle stage before being used in the final stage. If the original assay is changed for any reason, we call it a **modified assay**. This modification leads each patient to have two or more measurements and leaves unanswered the question on whether the biomarker performance will still be similar under the modified assay or not. In the medical field, assessing the agreement between two or more measurements is commonly known as reliability or inter-rater agreement (Kottner et al., 2011) while in engineering it is called a gauge repeatability and reproducibility study (Burdick et al., 2005, Ruiz Espejo, 2006).

Traditionally, measurement of reproducibility has been done using Pearson correlation coefficient, paired t-test, least square analysis of slope (=1) and intercept (=0) and coefficient of variation. However, none of these methods can assess the desired reproducibility characteristics, **precision** and **accuracy** at the same time (Bland and Altman, 1986, Lawrence and Lin, 1989, Müller and Büttner, 1994).

Pearson correlation coefficient only measures the strength the linear relationship between the two measurements but fails to detect any departure from the 45^o line. It is common for two measurments to have a high values of Pearson correlation coefficient but poor agreement (Bland and Altman, 1986). Paired t-test as elaborated in Lin (1989) would fail to detect a poor agreement in pairs of data such as $(1, 3.5), (2.5, 3), (3, 3), (4, 3)$ and $(5, 3)$. This data set

will result in a small test statistic and fail to reject the null hypothesis of good agreement. The least square approach for testing the slope (=1) and intercept (=0) also gives a misleading conclusion. More scattered data have a lower chance that the null hypothesis (slope=1 and/or intercept=0) would be rejected. On the other hand, highly reproducible results could result in rejecting the null hypothesis due to small standard error (Lawrence and Lin, 1989, Obuchowski et al., 2015). The Bland-Altman plot is another graphical method used to assessed the agreement of two measurements (Bland and Altman, 1986; 1999).

In the medical field the two commonly used reproducibility indices are the concordance correlation coefficient (CCC) and the intraclass correlation coefficient (ICC). Lin (1989) developed the concordance correlation coefficient and has been in use a lot since then. The CCC which assess agreement without the ANOVA assumptions includes precision and accuracy components. For pairs of n samples (X_{i1}, X_{i2}), that are independently sampled from bivariate normal with means μ_1 and μ_2 and respective variances and covariance $\sigma_1^2, \sigma_2^2, \sigma_{12}$, the $CCC = \frac{2\sigma_{12}}{\sigma_1^2+\sigma_2^2+(\mu_1-\mu_2)^2} = \rho C_b$. Here $C_b = [(v + 1/v + u^2)/2]^{-1}$, where $v = \sigma_1/\sigma_2$ and $u = (\mu_1 - \mu_2)/\sqrt{\sigma_1\sigma_2}$. The Pearson correlation coefficient ρ measures the precision component (how far each observation deviate from the best fit line) and the C_b component measures how far the best fit line deviates from the 45^o line (accuracy). This original work was latter extended to include general situations where there are more than two observers for data without replication and for data with replication (Chen and Barnhart, 2008).

The intraclass correlation coefficient (Fisher, 1925) is another widely used metric in biomedical research to assess reproducibility of measurements among raters, labs, technicians, or devices. The original ICC was based on the one-way analysis of variance (ANOVA) design, where there are only the subjects and observer (or lab or device) effects in the model. From the results of a one-way ANOVA table one can calculate, $ICC = \frac{\sigma_b^2}{\sigma_b^2+\sigma_e^2}$, where σ_b^2 is the between subjects variablility and σ_e^2 is the within subject variability. The original ICC (which we will call it here as ICC_1) was further extended to the second and third ICCs

(ICC_2, ICC_3), which are based on two-way ANOVA model with and without interaction respectively (McGraw and Wong, 1996, Shrout and Fleiss, 1979, Bartko, 1966).

However, all the aforementioned reproducibility metrics, Pearson correlation coefficient, regression line, the graphical Bland-Altman plot, ICC and CCC cannot be used to assess the change in Θ that results when an assay is modified. A high value of ICC between an original assay and a modified assay does not mean that the two assays will have the same clinical utility performance as measured by Θ. When the original assay is observed along with the outcome of interest, a method developed by Song and Pepe (2004) and Janes et al. (2014a) can be used to get an estimate for Θ. However, estimating Θ under the modified assay is not straight forward because we do not observed a new outcome associated with the modified assay.

In this paper, we propose a new reproducibility metric Δ_r which is an estimate of the difference in Θ under the two scenarios (original assay vs modified assay observed). Implementation of this method is demonstrated using the Ki67 reproducibility study. An R package Reproducibility Metric for Predictive Biomarkers (RMPB) is made available via github.

The rest of this paper is organized as follows. A motivational example will be presented in Section 5.2. In Section 5.3 we set the scenario and introduce the notations. The mathematical derivation for estimating Δ_r is provided in Section 5.4. In Sections 5.5 and 5.7, simulation results and application using KI67 reproducibility study are presented respectively. This paper concludes in Section 5.7 with discussion.

5.2 Motivating Context

In breast cancer research and management, Ki67 has the ability to assess immunonohistochemical proliferation (Viale et al., 2008, Dowsett et al., 2011, Goldhirsch et al., 2011).

Ki67 is a nuclear proliferation biomarker used to identify the growth fraction of a given cell population (Yerushalmi et al., 2010). However, the potential use of this biomarker in clinical decision making is still limited due to lack of reproducibility in measurement (Harris et al., 2007). To set a universal outline for measuring Ki67 and identify the main factors that are bottle necks for the consistency of the measurment, Polley et al. (2013) conducted an international reproducibility study of Ki67. In the study, one hundred breast cancer cases where measured in eight different labs and Ki67 score was recorded for 100 patients in each of eight labs.

Polley et al. (2013) used the intraclass correlation coefficient (ICC) as a measure of agreement between the these labs. However, this alone is not a good enough metric to measure the clinical utility of the biomarks for two main reasons: (1) a high value of ICC does not mean that the two biomarkers have similar clinical performance when assessed using the metric Θ and (2)to directly compare the clinical performance of two biomarkers using Θ we need to observe the outcome associated with each biomarker. Observing an outcome for a second time, however, is time consuming and costly. Using the metric Δ_r we developed, we assessed the reproducibility of Ki64 data obtained from the eight labs.

5.3 Settings and Notations

Consider a randomized control trial where half of the subjects are assigned to active treatment and the other half to placebo (or standard of care). For simplicity we will denote the treatment assignment as A such that $A = 1$ if the subject is assigned to active treatment and $A = 0$ otherwise. Let the clinical end point of interest be a binary indicator denoted by Y such that $Y = 1$ represents the occurrence of a bad outcome while $Y = 0$ for a favorable outcome. Further define the continuous candidate biomarker X which is measured at baseline from each subject.

5.3.1 Assumptions For Estimating Δ_r

For the purpose of developing the reproducibility metric Δ_r later, the following assumptions need to be taken into consideration: (a) an observed outcome, whether $Y = 0$ or $Y = 1$ for a subject i is independent of other subjects treatment assignment; (b) treatment assignment is independent of an individual's biomarker value, i.e, $A \perp X$; (c) a given treatment is either useful or of no harm and (d) the relationship between the outcome Y , the treatment assignment A, the continuous biomarker X along with the biomarker-treatment interaction is represented using a multiple logistic regression model as:

$$Ln\left[\frac{Pr(Y=1|A,X)}{1-Pr(Y=1|A,X)}\right] = \beta_0 + \beta_1 X + \beta_2 A + \beta_3 AX. \tag{5.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ are the logistic model parameters.

5.3.2 Optimal Treatment Decision Criterion

The first step in predictive biomarker evaluation process is to set the optimal treatment rule or algorithm. Based on this rule, treatment assignment will depend on the subject's biomarker value. Let the absolute treatment effect be represented by $\Delta(X) = P(Y = 1|A = 0, X) - P(Y = 1|A = 1, X)$. From this, the optimal treatment is set in such a way that subjects will be treated if and only if $\Delta(X) \geq 0$ but not treated (assigned to placebo) otherwise. This method of setting the optimal treatment rule was previously used by Brinkley et al. (2010) and Janes et al. (2014). A simple algebraic manipulation of equation (5.1) will lead this optimal rule to be: treat a subject if $X < \frac{-\beta_2}{\beta_3}$ and do not treat a subject if $X \geq \frac{-\beta_2}{\beta_3}$ assuming $\beta_3 > 0$. When $\beta_3 < 0$, the reverse will be true.

5.4 Assessing Reproducibility Of Two Biomarkers

Let the original biomarker be X, which we will consider it to be the **gold standard** and the modified biomarker W, which we call it a **"modified assay"**. Further we will assume $W = X + U$ where $U \sim \mathcal{N}(0, \sigma_e^2)$. Assessing reproducibility of these two biomarkers using previously developed and studied metrics (like ICC and CCC) is not enough in our context. A high value of ICC between X and W, does not necessary mean that they both have the same clinical utility in guiding treatment for patients. From figure 5.1, we can see that the relationship between ICC and the parameter of interest Θ_1 is not proportional. This indicates reproducibility assessment of X and W is not be fully captured using metrics like ICC alone. A modified assay is deemed to reproduce the results of a gold standard biomarker if it resulted in the same or very similar values of Θ_1 as the gold standard.

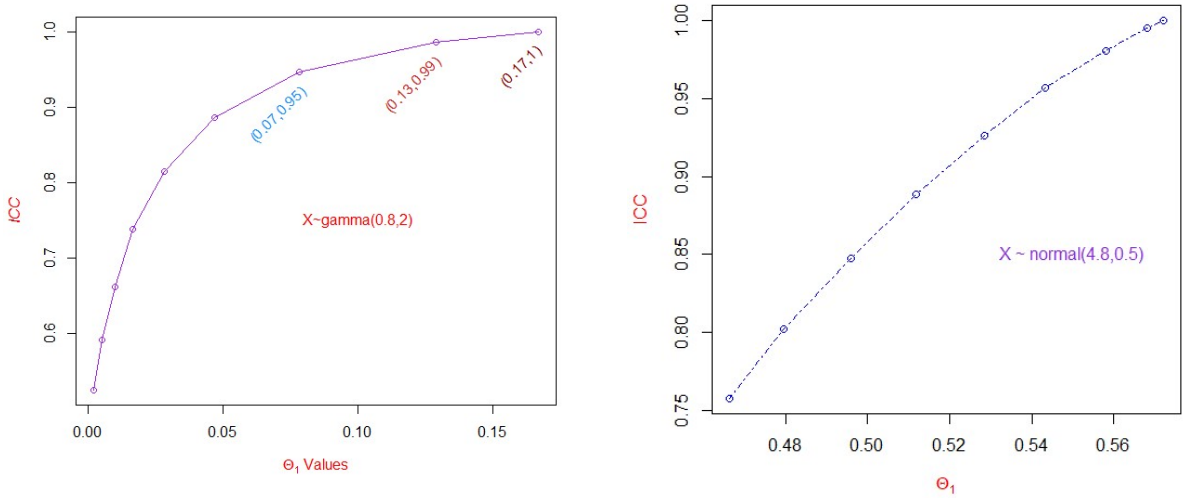

Figure 5.1: Shows the relationship between ICC and Θ. The left plot shows a scenario where the gold standard biomarker (X) is assumed to have a gamma distribution and the one on the right assuming a normally distributed biomarker. In both cases, the modified biomarker W is simulated such that $W = X + U$ where $U \sim \mathcal{N}(0, \sigma_e^2)$ represents the error term.

5.4.1 Estimating Reproducibility Metric Δ_r

Estimation of Δ_r involves two steps. In step one, assuming the original biomarker X is observed, we estimate Θ and denote is as Θ_x. Assuming the modified assay W is observed, in step two, we estimate Θ and denote it as Θ_w.

5.4.2 Estimating Θ_x

Let the gold standard biomarker be $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ with a probability density function given by $f(x)$. The probabilities of unfavorable outcomes when the default is "Treat All" and "Treat None" respectively are given as:

$$Pr(Y = 1|A = 0) = \int \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} f(x)dx \tag{5.2}$$

$$Pr(Y = 1|A = 1) = \int \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}}{1 + e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}} f(x)dx \tag{5.3}$$

The initial step in assessing the performance of a predictive biomarker is specifying a classifier rule. That is to develop a rule for recommending subjects whether a given treatment will benefit them or not based on their biomarker values. Let $\Delta_x(x)$ represent the absolute treatment effect. Then $\Delta_x(x)$ can be written as:

$$\begin{aligned} \Delta_x(x) &= Pr(Y = 1|A = 0, X = x) - Pr(Y = 1|A = 1, X = x) \\ &= \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}}{1 + e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}} \end{aligned} \tag{5.4}$$

Based on equation (4), the optimal classifier rule for recommending treatment is, treat a subject if $\Delta_x(x) \geq 0$ and recommend against taking a treatment if $\Delta_x(x) < 0$. This treatment

decision rule is similar was previous used by Zhang et al. (2012), Brinkley et al. (2010), and Janes et al. (2014). Assuming the coefficient $\beta_3 > 0$, the treatment rule of equation (5.4) is written as: treat a subject if $X < \frac{-\beta_2}{\beta_3}$ and to placebo if $X \geq \frac{-\beta_2}{\beta_3}$. After establishing the classifier rule, the next step in evaluating predictive biomarker performance is to quantify the decrease in the proportion of unfavorable outcome under established classifier based therapy which is obtained by estimating Θ_x. Let the current default treatment be "Treat All" and the coefficient $\beta_3 > 0$, then

$$\begin{aligned}
\hat{\Theta}_x^T &= \left[(\hat{P}(Y=1|A=1)\right] \\
&\quad - \left[\hat{P}(Y=1|A=1,\Delta_x(x)\geq 0)\hat{P}(\Delta_x(x)\geq 0) + \hat{P}(Y=1|A=0,\Delta_x(x)<0)\hat{P}(\Delta_x(x)<0)\right] \\
&= \left(\hat{P}(Y=1|A=1,\Delta_x(x)<0) - \hat{P}(Y=1|A=0,\Delta_x(x)<0)\right)\hat{P}(\Delta_x(x)<0) \\
&= \int_{-\hat{\beta}_2/\hat{\beta}_3}^{\infty} \left[\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}} - \frac{e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)x}}{1+e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)x}}\right] f(x)dx \\
&= \int_{-\hat{\beta}_2/\hat{\beta}_3}^{\infty} \left[\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}} - \frac{e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)x}}{1+e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)x}}\right] \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}\, dx \qquad (5.5)
\end{aligned}$$

such that $f(x)$ can be estimated from the normal probability density function given the biomarker values in the data. The notations in $\hat{\Theta}_x^T$ are chosen in such a way that superscript T represent the default treatment assumed is " treat all" and the subscript X represent that the observed biomarker is X. When the observed biomarker is W the subscript is changed to W from X and when the default treatment assumed is "treat none" the superscript is change to P from T.

5.4.3 Estimating Θ_w

Let the modified assay W be related to the gold standard biomarker X linearly by the equation $W = g(X) + U$ where U is randomly distributed independent error term. Further,

let the absolute treatment effect when W is observed instead of X be $\Delta_w(w)$, such that:

$$\Delta_w(w) \quad = \quad P(Y=1|A=0,) - P(Y=1|A=1,W) \tag{5.6}$$

Applying the rule of total probability, we can further write $\Delta_w(w)$ as

$$\begin{aligned}
\Delta_w(w) &= \int_{-\infty}^{\infty} \{P(Y=1|T=0,X=x) - P(Y=1|T=1,X=x)\} f(x|w)dx|w \\
&= \int_{-\infty}^{\infty} \left\{ \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)x}} \right\} f(x|w)dx|w \\
&= \int_{-\infty}^{\infty} \Delta_x(x) f(x|w)dx|w
\end{aligned} \tag{5.7}$$

If we assume the current default treatment is “Treat All", the decrease in the proportion of unfavorable event rate as a result of the marker guided therapy, can be shown to be:

$$\begin{aligned}
\hat{\Theta}_w^T &= \left(\hat{P}(Y=1|T=1,\Delta_w(w)<0) - \hat{P}(Y-1|T=0,\Delta_w(w)<0)\right)\hat{P}(\Delta_w(w)<0) \\
&= \int_{-\hat{\beta}_2/\hat{\beta}_3}^{\infty}\int_{-\infty}^{\infty} \left[\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}} - \frac{e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)x}}{1+e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)x}} \right] f(x|w)dx|w f(w)dw \\
&= \int_{-\hat{\beta}_2/\hat{\beta}_3}^{\infty}\int_{-\infty}^{\infty} \left[\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}} - \frac{e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)x}}{1+e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)x}} \right] f(x,w)dxdw
\end{aligned} \tag{5.8}$$

The joint density $f(x,w)$ in equation (5.8) can be estimated using parametric bivariate density function $\hat{f}(x,w)$. Let the gold standard biomarker $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and the modified assay be $W = X + U$. If we assume $U \sim \mathcal{N}(0, \sigma_e^2)$, then $W \sim \mathcal{N}(\mu_x, \sigma_w^2 = \sigma_x^2 + \sigma_e^2)$ and the joint distribution of X, W will be a bivariate normal such that;

$$\begin{pmatrix} X \\ W \end{pmatrix} \quad \sim \quad N\left[\begin{pmatrix} \mu_x \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{xw} \\ \sigma_{xw} & \sigma_w^2 \end{pmatrix} \right] \tag{5.9}$$

Then equation (5.8) can be further written as:

$$\begin{aligned}\hat{\Theta}_w^T &= \int_{-\hat{\beta}_2/\hat{\beta}_3}^{\infty}\int_{-\infty}^{\infty}\left[\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}} - \frac{e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)x}}{1+e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)x}}\right] \\ &\quad \frac{\exp\left\{-\frac{1}{2(1-\hat{\rho}^2)}\left[\left(\frac{x-\hat{\mu}_x}{\hat{\sigma}_x}\right)^2 - 2\hat{\rho}\left(\frac{x-\hat{\mu}_x}{\hat{\sigma}_x}\right)\left(\frac{w-\hat{\mu}_x}{\hat{\sigma}_w}\right) + \left(\frac{w-\hat{\mu}_x}{\hat{\sigma}_w}\right)^2\right]\right\}}{2\pi\hat{\sigma}_x\hat{\sigma}_w\sqrt{1-\hat{\rho}^2}} dxdw \end{aligned} \tag{5.10}$$

Estimation of the double integrals as in equation (5.10) can be done numerically or with a Monte Carlo. The Monte Carlo estimator of equation (5.10) will have the form:

$$\hat{\Theta}_w^T = \sum_{w_i:trt\ better}\sum_{x_i}\left\{\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_i}} - \frac{e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)x_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)x_i}}\right\} \tag{5.11}$$

where w_i*:trt better* represents the optimal treatment under the modified biomarker as stated in equation (5.7). When the sample n for the assay comparison is not large, the standard error associated with the estimator might be large since estimating the joint distribution $\hat{f}(x,w)$ will be noisy. From our simulation studies we have seen that ICC may be nearly sufficient, suggesting that we can condition on the correlation coefficient between x and w in order to reduce the Monte Carlo variation using Rao-Blackwellization method (Robert, 2004, Dobbin and Ionan, 2015).

Let ρ represent the correlation between X and W. Given the joint distribution of X and W follows a bivariate normal as in equation (5.9), if we observe W instead of X, the conditional expectation of X given W is written as;

$$\begin{aligned}\tilde{x}_i &= E[X_i|W_i] \\ &= \mu_x + \rho\frac{\sigma_x}{\sigma_w}(W_i-\mu_x)\end{aligned} \tag{5.12}$$

The Monte Carlo estimator of $\hat{\Theta}_w^T$ in equation (5.11) is then obtained by

$$\hat{\Theta}_w^T = \sum_{w_i:trt\ \ better} \left\{ \frac{e^{\hat{\beta}_0+\hat{\beta}_1\tilde{x}_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_1\tilde{x}_i}} - \frac{e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)\tilde{x}_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_2+(\hat{\beta}_1+\hat{\beta}_3)\tilde{x}_i}} \right\} \tag{5.13}$$

where $\tilde{x}_i$ is obtained using equation (5.12). The parameters in equation (5.12) are replaced by their respective estimators such that $\hat{\rho} = r$ and $r = ICC$.

5.4.4 Estimating Δ_r

Finally the estimates of $\hat{\Theta}_x^T$ and $\hat{\Theta}_w^T$ are combined to get an estimator of the reproducibility metric $\hat{\Delta}_r^T$ as:

$$\begin{aligned} \hat{\Delta}_r^T &= \int\int \left\{ \hat{P}(Y=1|X,A) - \hat{P}(Y=1|W,A) \right\} \hat{f}(x,w)dxdw \\ &= \hat{\Theta}_x^T - \hat{\Theta}_w^T \end{aligned} \tag{5.14}$$

Under ideal conditions the modified assay W is said to reproduce the results of the gold standard biomarker X perfectly if $\hat{\Delta}_r^T = 0$. The higher the value $\hat{\Delta}_r^T$ is different from 0, the lesser will W be considered as substitutes for X. A confidence interval for Δ_r is constructed using Monte Carlo methods as in (Robert, 2004) or bootstrap methods (Efron and Tibshirani, 1994, Davison and Hinkley, 1997). When the default treatment is “Treat None" and the coefficient $\beta_3 < 0$, the extension of the above procedures is given in the supplementary materials.

5.5 Simulation Study

Using equation (5.1) following the steps detailed below, we first generated m data sets each with sample size n.

Step 1: Convert the clinician input values K_1 to K_4 to model parameters β_0 to β_3. Details are given in Appendix (A.0.3).

Step 2: Use randomization to assign subjects to $A = 0$ or $A = 1$ with probability 0.5

Step 3: Generate the gold standard biomarker X_i from a given probability distribution $f(x)$, $i = 1, 2, ..., n$.

Step 4: Generate the modified biomarker $W_i = X_i \ + \mathcal{N}(0, \sigma_e^2)$, $i = 1, 2, ..., n$.

Step 5: Calculate the probability of an event such that: $p = \frac{e^{\beta_0 \ + \ \beta_1 X \ + \ \beta_2 A \ + \ \beta_3 AX}}{1+e^{\beta_0 \ + \ \beta_1 X \ + \ \beta_2 A \ + \ \beta_3 AX}}$.

Step 6: Generate the outcome Y from a binomial distribution with success probability p obtained from step 5 above.

Step 7: Under the normality assumption for the biomarker and the error term, get $\tilde{X}_i$ such that: $\tilde{X}_i = E[X_i|W_i] = \mu_x \ + \ \rho\frac{\sigma_x}{\sigma_w}(W_i - \mu_x)$, where ρ is the correlation between X and W.

The $K_1 - K_4$ clinician values are chosen to reflect the biomarker performance and we can use the equations given in chapters 3 & 4 to convert them to the model parameters $\beta_0 - \beta_3$. This step can be skipped if someone has any prior information about the βs' to relate the outcome with the covariates of equation (5.1). If ρ is not previously known, in step 5, it can replaced by its estimator $\hat{\rho}$. In our first simulation we generate a biomarker $X \sim \mathcal{N}(4.8, 3.24)$. This is done to mimic the OncotypeDX biomarkes as given in (Janes et al., 2014a). Three scenarios are used to choose different K value combination to reflect differences in the biomarker clinical utility performance performance. The higher the value of $\hat{\Theta}_x^T$ the better the biomarker performance. The modified assay W is given such that, $W = X + \sigma_e^2$, where the values of σ_e^2 are given in the first column of Table 5.1. The estimated value of the reproducibility metric $\hat{\Delta}_r^T$ with its standard error and 95% confidence interval are presented for different values of σ_e^2.

In Table 5.1 of the first scenario, the value of $\hat{\Delta}_r^T$ ranges from 2.1% when $\sigma_e^2 = 0.3$ to 7.7% when $\sigma_e^2 = 1.8$. To make a conclusion whether the modified biomarker can be considered as a substitute for the gold standard biomarker, in addition to the values of $\hat{\Delta}_r^T$, one has to take

the disease under study into account. A 2.1% difference in the clinical outcome could mean a lot in cancer research studies but might be of less significance in a non-life threatening disease. Results of $\hat{\Delta}_r^T$ under a moderate and weak biomarker performance assumptions are provided in Tables 5.2 and 5.3 for scenarios 2 and 3 respectively.

5.6 Application to Ki67 Reproducibility Study

The nuclear proliferation marker Ki67 can be utilized for different purposes in clinical breast cancer management. However, the interlaboratory inconsistency limited its broader application. To study the interlaboratory consistency of Ki67, eight laboratories received 100 breast cancer cases, one set stained by the central laboratory (Experiment A) and a second set stained by the participating laboratory(Experiment B).

In order to assess, the interlaboratory reproducibility in terms of change in Θ, the reproducibility metirc $\hat{\Delta}_r^T$ was estimated. The standard deviation of the Ki67 score was minimum for laboratory E in both Experiments and we set this lab to be our reference point. The Ki67 score range from 0 to 100 for each lab. The data have neither outcome nor treatment assignment covariates. Assuming a random 1:1 treatment assignment and using the $K_1 - K_4$ values, we first generated a binary outcome for each lab. In the result tables, Scenario 1 refers generally to a strong predictive biomarker performance while Scenario 2 be weaker.

Table 5.4 show results for $\hat{\Delta}_r^T$, its standard error and 95% confidence interval from Experiment A assuming a stronger biomarker performance. For lab E, $\hat{\Theta}_x^T(SE) = 0.328(0.057)$. We considered this to be Θ obtained when the observed biomarker is the gold standard. Measurement of Ki67 score obtained from the rest of the laboratories (A, B, C, D, F, G, H) are assumed to be measurements of Ki67 under modification. The estimated values of $\hat{\Delta}_r^T$ range from 0.015 to 0.135. Looking at Table 5.4, the difference in Θ between the reference lab (lab E) and lab F is only 1.5% and one could consider the measures from these two labs

as consistent. On the other hand the difference in Θ between lab E and lab B is 13.5% which shows the incosistency in measurement between these two labs. Similar results are presented in Table 5.5 for a less stronger biomarker.

For experiment B, where the staining was done by each participating laboratory, similar trends are observed in the estimated values of $\hat{\Delta}_r^T$ as shown in Table 5.6. For the reference lab (lab E) under Scenario 1 $\hat{\Theta}_x^T(SE) = 0.345(0.054)$. The values of $\hat{\Delta}_r^T$ range from 0.021 to 0.189. From these results lab G measurements can be considered more consistent with the reference lab (lab E) with $\hat{\Delta}_r^T(SE) = 0.021(0.008)$ and 95% $CI = (0.003, 0.039)$. On the other hand lab B measurments looks less consistent when compared to lab E measurements $\hat{\Delta}_r^T(SE) = 0.189(0.037)$. Similar results under the weaker biomarker are shown in Table 5.7.

5.7 Discussion

Evaluating predictive biomarkers in a quick and less costly manner is of great importance for cancer researchers. A metric Θ developed by (Song and Pepe, 2004, Brinkley et al., 2010, Janes et al., 2014a) has been favored as measure of predictive biomarker clinical utility in the past years. This metric measures how well the biomarker guided therapy in reducing the expected proportion of population event rate in comparison to the default treatment (non biomarker guided treatment). However, when the original biomarker is modified for reasons such as to reduce the cost of assay validation and preparation, previous outcome-biomarker performance studies became invalid. Biomarker clinical utility evaluation under the modified biomarker then has to start over which is costly and time consuming as one needs to wait to observe the outcome again. In this paper we proposed a reproducibility metric Δ_r which assesses the performance of the modified assay in comparison to the original assay without needing researcher to wait for new outcome to occur. Confidence intervals for Δ_r are constructed using Monte Carlo Methods. An R package RMPB (Reproducibility

Metric for Predictive Biomarker) is made available via Github. Additionally, the asymptotic results for Δ are provided.

Assessing the agreement between two measurements in medical research is commonly done using metrics such as Pearson correlation coefficient, paired t-test, linear regression, Bald-Altman plot (Bland and Altman, 1986), ICC(Fisher, 1925) and CCC (Lawrence and Lin, 1989). Using simulation, we have shown that, a high value of ICC between two biomarkers does not mean the two biomarkers have similar performance as measured by the metric Θ. This was demonstrated by plotting ICC vs Θ. Since both ICC and Θ have values which range from 0 to 1, we would expect a straight line through the 45^0 if one metric could be used as a substitute for the other for the purpose of assessing biomarker performance. For a binary clinical endpoint, the relationship between a predictive biomarker and outcome have been assessed also using metrics like sensitivity, specificity, negative and positive predictive values and area under receiver operating characteristic (ROC) curve (AUC) (Søreide, 2009, Bharti and Bharti, 2009). Again we can not use these metrics directly for our intended purpose because (a) the outcome-biomarker association measured using these metric can not be directly translated to clinical biomarker performance measure metric Θ and (b) to estimate these metrics under the modified biomarker, one also need to wait to observe the new outcome.

In this paper, we have assumed both the original (gold standard) and the modified biomarker to have a normal distribution. The modified biomarker is considered as a sum of the original biomarker plus a normal error term with zero mean. But all the steps followed to estimate Δ_r under these assumptions could be extended (1) when the biomarker has a non-normal distribution and (2) when the error term has a distribution which is not normal. Similarly, even though this paper assumes a binary outcome, the steps outlined in this paper can be extended easily when clinical outcome of interest is continuous, count and time to event.

Acknowledgment

We thank Lisa McShane for providing the Ki67 data.

Table 5.1: An Estimate of Δ_r^T along its standard error and 95% CI under scenario 1. A 500 Monte Carlo samples each with 300 sample size were used.

Scenario 1: K_1= 0.25 K_2= 0.75 K_3= 0.75 K_4= 0.25				
	Θ_x^T	$\hat{\Theta}_w^T$	$\hat{\Delta}_r^T$	
σ_e^2	Estimate (SE)	Estimate (SE)	Estimate (SE)	95% CI
0.0	0.245(0.029)	0.245(0.029)	0.000(0.000)	
0.3		0.224(0.027)	0.021(0.009)	(0.004,0.039)
0.6		0.210(0.029)	0.035(0.012)	(0.012,0.057)
0.9		0.197(0.028)	0.048(0.137)	(0.021,0.073)
1.2		0.186(0.027)	0.059(0.015)	(0.029,0.086)
1.5		0.176(0.028)	0.069(0.016)	(0.038,0.097)
1.8		0.168(0.027)	0.077(0.017)	(0.045,0.106)

Table 5.2: An Estimate of Δ_r^T along its standard error and 95% CI under scenario 2. A 500 Monte Carlo samples each with 300 sample size were used.

Scenario 2: K1= 0.10 ; K2= 0.60 ; K3= 0.45 ; k4= 0.25				
	Θ_x^T	$\hat{\Theta}_w^T$	$\hat{\Delta}_r^T$	
σ_e^2	Estimate (SE)	Estimate (SE)	Estimate (SE)	95% CI
0.0	0.178(0.024)	0.178(0.024)	0.000(0.000)	
0.3		0.164(0.033)	0.014(0.008)	(0.001,0.031)
0.6		0.153(0.032)	0.025(0.011)	(0.005,0.049)
0.9		0.144(0.032)	0.034(0.013)	(0.011,0.062)
1.2		0.136(0.031)	0.042(0.014)	(0.016,0.073)
1.5		0.129(0.030)	0.049(0.015)	(0.022,0.082)
1.8		0.123(0.031)	0.055(0.016)	(0.026,0.089)

Table 5.3: An Estimate of Δ_r^T along its standard error and 95% CI under scenario 3. A 500 Monte Carlo samples each with 300 sample size were used.

Scenario 3: K1= 0.15 ; K2= 0.30 ; K3= 0.25 ; k4= 0.10				
	Θ_x^T	$\hat{\Theta}_w^T$	$\hat{\Delta}_r^T$	
σ_e^2	Estimate (SE)	Estimate (SE)	Estimate (SE)	95% CI
0.0	0.072 (0.026)	0.072(0.026)	0.000(0.000)	
0.3		0.068(0.025)	0.003(0.006)	(0.000,0.015)
0.6		0.065(0.024)	0.007(0.008)	(0.001,0.022)
0.9		0.062(0.025)	0.100(0.009)	(0.001,0.028)
1.2		0.059 (0.025)	0.012(0.011)	(0.002,0.033)
1.5		0.057(0.023)	0.015(0.012)	(0.002,0.038)
1.8		0.055(0.026)	0.017(0.013)	(0.002,0.041)

Table 5.4: An Estimate of Δ_r^T along its standard error and 95% CI under scenario 1 from the Ki67 reproducibility study Experiment A. We used lab E measurements as a gold standard. A 1000 Monte Carlo simulation each with 100 sample size was used to construct the 95% CI for Δ_r^T .

Scenario 1: K_1= 0.25 K_2= 0.75 K_3= 0.75 K_4= 0.25					
		Θ_x^T	$\hat{\Theta}_w^T$	$\hat{\Delta}_r^T$	
LAB	σ_e	Estimate (SE)	Estimate (SE)	Estimate (SE)	95% CI
E	0.00	0.328(0.057)	0.328(0.057)	0.000(0.000)	
F	0.92		0.313(0.053)	0.015(0.008)	(0.001,0.030)
A	3.65		0.265(0.050)	0.060(0.019)	(0.024,0.096)
H	3.69		0.264(0.050)	0.061(0.018)	(0.024,0.097)
G	4.61		0.252(0.049)	0.073(0.022)	(0.032,0.115)
D	5.74		0.238(0.048)	0.087(0.025)	(0.041,0.136)
C	8.02		0.213(0.047)	0.112(0.029)	(0.057,0.170)
B	10.58		0.190(0.046)	0.135(0.033)	(0.073,0.203)

Table 5.5: An Estimate of Δ_r^T along its standard error and 95% CI under scenario 2 from the Ki67 reproducibility study Experiment A. We used lab E measurements as a gold standard. A 1000 Monte Carlo simulation each with 100 sample size was used to construct the 95% CI for Δ_r^T .

Scenario 2: K_1= 0.10 K_2= 0.60 K_3= 0.45 K_4= 0.25					
		Θ_x^T	$\hat{\Theta}_w^T$	$\hat{\Delta}_r^T$	
LAB	σ_e	Estimate (SE)	Estimate (SE)	Estimate (SE)	95% CI
E	0.00	0.236 (0.065)	0.236(0.065)	0.000(0.000)	
F	0.92		0.223 (0.063)	0.012(0.006)	(0.002,0.025)
A	3.65		0.190(0.059)	0.046(0.018)	(0.015,0.082)
H	3.69		0.189(0.058)	0.047(0.017)	(0.015,0.082)
G	4.61		0.179(0.058)	0.056(0.021)	(0.019,0.099)
D	5.74		0.168(0.057)	0.067(0.025)	(0.023,0.116)
C	8.02		0.149(0.055)	0.087(0.031)	(0.031,0.146)
B	10.58		0.132(0.054)	0.106(0.036)	(0.039,0.176)

Table 5.6: An Estimate of Δ_r^T along its standard error and 95% CI under scenario 1 from the Ki67 reproducibility study Experiment B. We used lab E measurements as a gold standard. A 1000 Monte Carlo simulation each with 100 sample size was used to construct the 95% CI for Δ_r^T .

Scenario 1: K_1= 0.25 K_2= 0.75 K_3= 0.75 K_4= 0.25					
		Θ_x^T	$\hat{\Theta}_w^T$	$\hat{\Delta}_r^T$	
LAB	σ_e	Estimate (SE)	Estimate (SE)	Estimate (SE)	95% CI
E	0.00	0.345 (0.054)	0.345(0.054)	0.000(0.000)	
G	1.11		0.324(0.049)	0.021 (0.008)	(0.003,0.039)
A	2.23		0.302(0.048)	0.043(0.013)	(0.017,0.072)
F	2.27		0.291(0.047)	0.054(0.016)	(0.022,0.089)
C	4.31		0.268(0.046)	0.077(0.020)	(0.036,0.122)
D	6.18		0.242(0.045)	0.102(0.025)	(0.054,0.157)
H	6.20		0.241(0.045)	0.101(0.024)	(0.053,0.157)
B	16.17		0.156(0.042)	0.189(0.037)	(0.118,0.272)

Table 5.7: An Estimate of Δ_r^T along its standard error and 95% CI under scenario 2 from the Ki67 reproducibility study Experiment B. We used lab E measurements as a gold standard. A 1000 Monte Carlo simulation each with 100 sample size was used to construct the 95% CI for Δ_r^T .

Scenario 2: K_1= 0.10 K_2= 0.60 K_3= 0.45 K_4= 0.25					
		Θ_x^T	$\hat{\Theta}_w^T$	$\hat{\Delta}_r^T$	
LAB	σ_e	Estimate (SE)	Estimate (SE)	Estimate (SE)	95% CI
E	0	0.250(0.068)	0.250(0.068)	0.000(0.000)	
G	1.11		0.234 (0.066)	0.016(0.007)	(0.002,0.032)
A	2.23		0.220(0.064)	0.031(0.012)	(0.008,0.056)
F	2.27		0.212(0.063)	0.039(0.015)	(0.010,0.070)
C	4.31		0.195(0.061)	0.057(0.021)	(0.016,0.099)
D	6.18		0.175(0.058)	0.077(0.027)	(0.024,0.133)
H	6.20		0.175(0.059)	0.077(0.027)	(0.024,0.133)
B	16.17		0.106(0.054)	0.145(0.046)	(0.058,0.235)

Chapter 6

Summary and Future Research

6.1 Summary

In this dissertation we proposed three statistical methodologies for assessing the clinical utility and reproducibility of predictive biomarkers. One is proposing a metric Ψ_B using Bayesian decision theoretic framework; the second is proposing a sample size estimation method SWIRL; and the third is proposing a reproducibility metric Δ_r.

A predictive biomarker utility performance metric Ψ was proposed using Bayesian decision theory framework. Early phase clinical trial data are usually small and the maximum likelihood based estimator are biased and inefficient. This on the other hand leads to conclusions which are flawed. However, adding experts' prior information and data collected from the early phase studies together, Bayesian methods were used to estimate the Ψ and overcome the problem. Novel equations were used to convert clinician(expert) information to useful priors. Estimation of Ψ was outlined in a general framework so that monetary drug costs and negative side effects of the drug can be taken into consideration during the evaluation process. A more efficient MCMC algorithm Hamiltonian Monte Carlo (HMC) was used to get estimate for the posterior mean of Ψ along its standard error and 95% credible

interval.

When the primary goal is to evaluate the clinical performance of a treatment selection biomarker, sample size determination is a key part of the study design. For a binary clinical endpoint, Θ which measures the decrease in the proportion of unfavorable outcomes that results from biomarker guided therapy has been advocated as a metric for evaluating the marker's performance. However, a sample size estimation method was lacking to supplement the biomarker study design. A novel sample size estimation algorithm, Squared Width Inversion Linear Regression (SWILR), is proposed to determine a sample size n so that the 95% CI mean width of Θ is less than the user defined length (W_{targ}). This is the first sample size method developed for estimating the the predictive biomarker clinical utility performance metric Θ. With the SWIRL algorithm, m data sets with an increment of c are first generated from the multiple logistic regression model given in equation (4.1). Under each data set, the mean 95% CI width (w) of Θ is estimated either using the equations we developed in chapter 4 or using the Janes et. al (2014) method and a linear regression is then fitted with n as outcome and $\frac{1}{w^2}$ as a covariate. The fitted regression line is finally used to estimate a sample size n for a user defined 95% CI mean width (W_{targ}) of the metric Θ. The linearity assumption for the implementation of the SWIRL algorithm comes from the asymptotic distribution proof for Θ which is given in Appendix (A.0.1).

A treatment selection biomarker undergoes three stages of validation and evaluation before it can be used for decision making. Of the three stages, in the middle development stage, the biomarker or assay is required to be modified for reasons such as simplification of its sample preparation, minimizing the cost, migrating the assay platform etc. Modification of the assay however, invalidates previous outcome-assay performance studies as measured by metrics such as Θ. Under the modified assay, estimating Θ is not straightforward because we do not observe the outcome. Existing reproducibility metrics such as the Pearson correlation coefficient (ρ), intraclass correlation coefficient (ICC), concordance correlation coefficient

(CCC) and others are not suitable metrics for this purpose. This is because a high values of ICC or ρ between the orginal assay (X) and the modified assay (W) does not directly translate to Θ_x and Θ_w being equal. To assess the change in Θ under the modified assay we proposed a reproducibility metric Δ_r. The key advantage of Δ_r is, it does not need observation of the clinical outcome under the modified biomarker. This helps biomarker researchers to assess the effect of assay modification on the clinical performance of the assay with less time and at a much lower cost.

6.2 Direction for Future Research

Now it is becoming a common practice to use two or more predictive biomarkers' information simultaneously to make clinical decision. In the asthma clinical trial for example, three different biomarkes (fractional exhaled nitric oxide [FeNO], blood eosinophils and periostin) were initially under consideration (Korenblat et al., 2018). As such the methods proposed in this dissertation can be extended to take this issue into consideration. Let the outcome of interest be $\mathbf{y}$, which is binary, such that, $\mathbf{y} \in \{0, 1\}$. Further let the input variables be represented by a vector $\mathbf{x}$ such that $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \mathbf{T})$ where $\mathbf{x_1}$ and $\mathbf{x_2}$ denote the biomarkers BMK_1 and BMK_2 respectively which are measured at baseline from each subject and $\mathbf{T}$ for the treatment assignment, placebo or active group. Additionally, $\theta \in \Theta$ will represent the parameter subspace that relate the outcome Y with the inputs $\mathbf{x}$. The natural approach to represent the relationship between the outcome Y and the vector of input $\mathbf{x}$ is through a log linear model, which can be written as,

$$Ln\left[\frac{Pr(Y_i = 1|X_{1i}, X_{2i}, T_i)}{1 - Pr(Y_i = 1|X_{1i}, X_{2i}, T_i))}\right] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 A_i + \beta_4 X_{1i} A_i + \beta_5 X_{2i} A_i + \beta_6 X_{1i} X_{2i} \tag{6.1}$$

where $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ and β_6 in the model represent the biomarkers, treatment, biomarker by treatment interaction and biomarker by biomarker interaction effects respectively. Now let, both biomarkers X_1 and X_2 have a known joint probability density function each given by $f(x_1, x_2)$ where $X_1 \in (-\infty, \infty)$ and $X_2 \in (-\infty, \infty)$. The probability of unfavorable outcome given treatment is written as :

$$\begin{aligned} Pr(Y=1|T=1) &= \int\int \left\{ \frac{e^{\beta_0+\beta_3+(\beta_1+\beta_4+\beta_6X_2)X_1+(\beta_2+\beta_5)X_2}}{1+e^{\beta_0+\beta_3+(\beta_1+\beta_4+\beta_6X_2)X_1+(\beta_2+\beta_5)X_2}} \right\} f(x_1,x_2)dx_1dx_2 \\ Pr(Y=1|T=0) &= \int\int \left\{ \frac{e^{\beta_0+(\beta_1+\beta_6X_2)X_1+\beta_2X_2}}{1+e^{\beta_0+(\beta_1+\beta_6X_2)X_1+\beta_2X_2}} \right\} f(x_1,x_2)dx_1dx_2 \end{aligned}$$

6.2.1 Optimal Treatment Assignment

from equation (1) above it is straightforward to show that:

$$\begin{aligned} \frac{Odds(Y=1|T=1,X=x)}{Odds(Y=1|T=0,X=x)} &= \frac{exp\{\beta_0+\beta_3+(\beta_1+\beta_4+\beta_6X_2)X_1+(\beta_2+\beta_5)X_2\}}{exp\{\beta_0+(\beta_1+\beta_6X_2)X_1+\beta_2X_2\}} \\ &= exp\{\beta_3+\beta_4X_1+\beta_5X_2\} \end{aligned} \tag{6.2}$$

From equation (4.4) it is clear that, odds of an unfavorable outcome are greater among subjects in the active arm than the SOC arm if $\beta_3 + \beta_4X_1 + \beta_5X_2 > 0$. Based on this, the biomarker guided treatment decision (T_{opt}) can be set in such a way that :

$$T_{opt}(X=x) = \begin{cases} T=1 & : \beta_4X_1+\beta_5X_2 < -\beta_3 \\ T=0 & : \beta_4X_1+\beta_5X_2 \geq -\beta_3 \end{cases} \tag{6.3}$$

Depending on the sign of (β_4) coefficient a threshold for the marker guided therapy can be written as:

$$T_{opt}(X=x) \Rightarrow \texttt{if} : \beta_4 < 0 \begin{cases} T=1: & X_1 > \frac{-\beta_3-\beta_5 X_2}{\beta_4} \\ \\ T=0: & X_1 \leq \frac{-\beta_3-\beta_5 X_2}{\beta_4} \end{cases}$$

$$T_{opt}(X=x) \Rightarrow \texttt{if} : \beta_4 > 0 \begin{cases} T=1: & X_1 < \frac{-\beta_3-\beta_5 X_2}{\beta_4} \\ \\ T=0: & X_1 \geq \frac{-\beta_3-\beta_5 X_2}{\beta_4} \end{cases}$$

6.2.2 Estimation of Θ

Once the optimal treatment rule is set as in equation (6.3), one can proceed to estimate Θ using the procedures outlined in chapter 4. If we assume the default treatment is treat all for example, we can get an estimate of Θ_1 as:

$$\begin{aligned} \hat{\Theta}_1 &= Pr(Y=1|T=1,X_1,X_2) - Pr(Y=1|T_{opt},X_1,X_2) \\ &= \iint\limits_{trt:all} \left\{ \frac{e^{\beta_0+\beta_3+(\beta_1+\beta_4+\beta_6 X_2)X_1+(\beta_2+\beta_5)X_2}}{1+e^{\beta_0+\beta_3+(\beta_1+\beta_4+\beta_6 X_2)X_1+(\beta_2+\beta_5)X_2}} \right\} f(x_1,x_2)dx_1dx_2 - \\ &\quad \iint\limits_{trt:opt} \left\{ \frac{e^{\beta_0+\beta_3+(\beta_1+\beta_4+\beta_6 X_2)X_1+(\beta_2+\beta_5)X_2}}{1+e^{\beta_0+\beta_3+(\beta_1+\beta_4+\beta_6 X_2)X_1+(\beta_2+\beta_5)X_2}} \right\} f(x_1,x_2)dx_1dx_2 \end{aligned} \tag{6.4}$$

Similarly, when the default treatment is treat none, an estimate of Θ_0 can be obtained as:

$$\begin{aligned} \hat{\Theta}_0 &= Pr(Y=0|T=1,X_1,X_2) - Pr(Y=1|T_{opt},X_1,X_2) \\ &= \iint\limits_{trt:none} \left\{ \frac{e^{\beta_0+(\beta_1+\beta_6 X_2)X_1+\beta_2 X_2}}{1+e^{\beta_0+(\beta_1+\beta_6 X_2)X_1+\beta_2 X_2}} \right\} f(x_1,x_2)dx_1dx_2 - \\ &\quad \iint\limits_{trt:opt} \left\{ \frac{e^{\beta_0+\beta_3+(\beta_1+\beta_4+\beta_6 X_2)X_1+(\beta_2+\beta_5)X_2}}{1+e^{\beta_0+\beta_3+(\beta_1+\beta_4+\beta_6 X_2)X_1+(\beta_2+\beta_5)X_2}} \right\} f(x_1,x_2)dx_1dx_2 \end{aligned} \tag{6.5}$$

Bibliography

Kathy S Albain, William E Barlow, Steven Shak, Gabriel N Hortobagyi, Robert B Livingston, I-Tien Yeh, Peter Ravdin, Roberto Bugarini, Frederick L Baehner, Nancy E Davidson, et al. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *The lancet oncology*, 11(1): 55–65, 2010.

Douglas G Altman and J Martin Bland. Measurement in medicine: the analysis of method comparison studies. *The statistician*, pages 307–317, 1983.

Rafael G Amado, Michael Wolf, Marc Peeters, Eric Van Cutsem, Salvatore Siena, Daniel J Freeman, Todd Juan, Robert Sikorski, Sid Suggs, Robert Radinsky, et al. Wild-type kras is required for panitumumab efficacy in patients with metastatic colorectal cancer. *Journal of clinical oncology*, 26(10):1626–1634, 2008.

Shashi Amur, Lisa LaVange, Issam Zineh, S Buckman-Garner, and Janet Woodcock. Biomarker qualification: toward a multiple stakeholder framework for biomarker development, regulatory acceptance, and utilization. *Clinical Pharmacology and Therapeutics*, 98(1):34–46, 2015.

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

Stuart G Baker and Barnett S Kramer. Evaluating surrogate endpoints, prognostic markers, and predictive markers: some simple themes. *Clinical Trials*, 12(4):299–308, 2015.

John R Ball, Christine M Micheel, et al. *Evaluation of biomarkers and surrogate endpoints in chronic disease*. National Academies Press, 2010.

John J Bartko. The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, 19(1):3–11, 1966.

Eric D Bateman, Homer A Boushey, Jean Bousquet, William W Busse, Tim JH Clark, Romain A Pauwels, and Søren E Pedersen. Can guideline-defined asthma control be achieved? the gaining optimal asthma control study. *American journal of respiratory and critical care medicine*, 170(8):836–844, 2004.

Elisabeth H Bel. Clinical phenotypes of asthma. *Current opinion in pulmonary medicine*, 10(1):44–50, 2004.

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

Bhavneet Bharti and Sahul Bharti. Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research: trade-off between sensitivity and specificity with change of test cut-offs. *Journal of clinical pathology*, 62(11):1051–1051, 2009.

J Martin Bland and Douglas G Altman. Measuring agreement in method comparison studies. *Statistical methods in medical research*, 8(2):135–160, 1999.

J Martin Bland and DouglasG Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.

Marco Bonetti and Richard D Gelber. Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*, 5(3):465–481, 2004.

Patrick M Bossuyt and Tajik Parvin. Evaluating biomarkers for guiding treatment decisions. *EJIFCC*, 26(1):63, 2015.

Jason Brinkley, Anastasios Tsiatis, and Kevin J Anstrom. A generalized estimator of the attributable benefit of an optimal treatment regime. *Biometrics*, 66(2):512–522, 2010.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

Richard K Burdick, Connie M Borror, and Douglas C Montgomery. *Design and analysis of gauge R&R studies: making decisions with confidence intervals in random and mixed ANOVA models*, volume 17. SIAM, 2005.

Marc Buyse. Towards validation of statistically reliable biomarkers. *European Journal of Cancer Supplements*, 5(5):89–95, 2007.

David P Byar. Assessing apparent treatment—covariate interactions in randomized clinical trials. *Statistics in Medicine*, 4(3):255–263, 1985.

Tianxi Cai, Lu Tian, Peggy H Wong, and LJ Wei. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282, 2010.

Bradley P Carlin and Thomas A Louis. *Bayesian methods for data analysis*. CRC Press, 2008.

Bradley P Carlin and Thomas A Louis. *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall/CRC, 2010.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Chia-Cheng Chen and Huiman X Barnhart. Comparison of icc and ccc for assessing agreement for data without and with replications. *Computational Statistics & Data Analysis*, 53(2):554–564, 2008.

Jonathan Corren, Robert F Lemanske Jr, Nicola A Hanania, Phillip E Korenblat, Merdad V Parsey, Joseph R Arron, Jeffrey M Harris, Heleen Scheerens, Lawren C Wu, Zheng Su, et al. Lebrikizumab treatment in adults with asthma. *New England Journal of Medicine*, 365(12):1088–1098, 2011.

Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

Morris H DeGroot. *Optimal statistical decisions*, volume 82. John Wiley & Sons, 2005.

Eugene Demidenko. Sample size determination for logistic regression revisited. *Statistics in medicine*, 26(18):3385–3397, 2007.

Eugene Demidenko. Sample size and optimal design for logistic regression with binary interaction. *Statistics in medicine*, 27(1):36–46, 2008.

Kevin K Dobbin and Alexei C Ionan. Sample size methods for constructing confidence intervals for the intra-class correlation coefficient. *Computational Statistics & Data Analysis*, 85:67–83, 2015.

ALLAN DONNER. The use of correlation and regression in the analysis of family resemblance. *American journal of epidemiology*, 110(3):335–342, 1979.

Mitch Dowsett, Torsten O Nielsen, Roger A'Hern, John Bartlett, R Charles Coombes, Jack Cuzick, Matthew Ellis, N Lynn Henry, Judith C Hugh, Tracy Lively, et al. Assessment of ki67 in breast cancer: recommendations from the international ki67 in breast cancer working group. *Journal of the National Cancer Institute*, 103(22):1656–1664, 2011.

Jing Du, Young-Taek Park, Nawanan Theera-Ampornpunt, Jeffrey S McCullough, and Stuart M Speedie. The use of count data models in biomedical informatics evaluation research. *Journal of the American Medical Informatics Association*, 19(1):39–44, 2011.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

BM Fine and L Amler. Predictive biomarkers in the development of oncology drugs: a therapeutic industry perspective. *Clinical Pharmacology & Therapeutics*, 85(5):535–538, 2009.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

Boris Freidlin, Lisa M McShane, and Edward L Korn. Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute*, 2010.

Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. In *Readings in Computer Vision*, pages 564–584. Elsevier, 1987.

Charles Geyer. Introduction to markov chain monte carlo. *Handbook of markov chain monte carlo*, 20116022:45, 2011.

Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

Natasa Gisev, J Simon Bell, and Timothy F Chen. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9(3):330–338, 2013.

Oleg Gluz, Ulrike A Nitz, Matthias Christgen, Ronald E Kates, Steven Shak, Michael Clemens, Stefan Kraemer, Bahriye Aktas, Sherko Kuemmel, Toralf Reimer, et al. West german study group phase iii planb trial: first prospective outcome data for the 21-gene recurrence score assay and concordance of prognostic markers by central and local pathology assessment. *Journal of Clinical Oncology*, 34(20):2341–2349, 2016.

A 2011 Goldhirsch, William C Wood, Alan S Coates, Richard D Gelber, Beat Thürlimann, H-J Senn, and Panel members. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2011. *Annals of oncology*, 22(8):1736–1747, 2011.

Sylvan B Green. Patient heterogeneity and the need for randomized clinical trials. *Controlled Clinical Trials*, 3(3):189–198, 1982.

Lacey Gunter, Ji Zhu, and Susan Murphy. Variable selection for optimal decision making. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 149–154. Springer, 2007.

Nicola A Hanania, Michael Noonan, Jonathan Corren, Phillip Korenblat, Yanan Zheng, Saloumeh K Fischer, Melissa Cheu, Wendy S Putnam, Elaine Murray, Heleen Scheerens,

et al. Lebrikizumab in moderate-to-severe asthma: pooled data from two randomised placebo-controlled studies. *Thorax*, pages thoraxjnl–2014, 2015.

Lyndsay Harris, Herbert Fritsche, Robert Mennel, Larry Norton, Peter Ravdin, Sheila Taube, Mark R Somerfield, Daniel F Hayes, and Robert C Bast Jr. American society of clinical oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *Journal of clinical oncology*, 25(33):5287–5312, 2007.

Lyndsay N Harris, Nofisat Ismaila, Lisa M McShane, Fabrice Andre, Deborah E Collyar, Ana M Gonzalez-Angulo, Elizabeth H Hammond, Nicole M Kuderer, Minetta C Liu, Robert G Mennel, et al. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American society of clinical oncology clinical practice guideline. *Journal of Clinical Oncology*, 34(10):1134–1150, 2016.

Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

FY Hsieh. Sample size tables for logistic regression. *Statistics in medicine*, 8(7):795–802, 1989.

Mei-Chen Hu, Martina Pavlicova, and Edward V Nunes. Zero-inflated and hurdle models of count data with extra zeros: examples from an hiv-risk reduction intervention trial. *The American journal of drug and alcohol abuse*, 37(5):367–375, 2011.

HE Hua, TANG Wan, WANG Wenjuan, and CRITS-CHRISTOPH Paul. Structural zeroes and zero-inflated models. *Shanghai archives of psychiatry*, 26(4):236, 2014.

Ying Huang, Peter B Gilbert, and Holly Janes. Assessing treatment-selection markers using a potential outcomes framework. *Biometrics*, 68(3):687–696, 2012.

Holly Janes, Margaret S Pepe, Patrick M Bossuyt, and William E Barlow. Measuring the performance of markers for guiding treatment decisions. *Annals of internal medicine*, 154 (4):253–259, 2011.

Holly Janes, Marshall D Brown, Ying Huang, and Margaret S Pepe. An approach to evaluating and comparing biomarkers for patient treatment selection. *The international journal of biostatistics*, 10(1):99–121, 2014a.

Holly Janes, Margaret S Pepe, and Ying Huang. A framework for evaluating markers used to select patient treatment. *Medical Decision Making*, 34(2):159–167, 2014b.

Holly Janes, Marshall D Brown, and Margaret S Pepe. Designing a study to evaluate the benefit of a biomarker for selecting patient treatment. *Statistics in medicine*, 34(27): 3503–3515, 2015.

Hakjin Jang, Soobeom Lee, and Seong W Kim. Bayesian analysis for zero-inflated regression models with the power prior: Applications to road safety countermeasures. *Accident Analysis & Prevention*, 42(2):540–547, 2010.

NL Johnson and S Kotz. Distributions in statistics: Discrete distributionshoughton mifflin, 1969.

Christos S Karapetis, Shirin Khambata-Ford, Derek J Jonker, Chris J O'callaghan, Dongsheng Tu, Niall C Tebbutt, R John Simes, Haji Chalchal, Jeremy D Shapiro, Sonia Robitaille, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine*, 359(17):1757–1765, 2008.

Katy Klauenberg and Clemens Elster. Markov chain monte carlo methods: an introductory example. *Metrologia*, 53(1):S32, 2016.

Philip Korenblat, Edwin Kerwin, Igor Leshchenko, Karl Yen, Cecile TJ Holweg, Judith Anzures-Cabrera, Carmen Martin, Wendy S Putnam, Laura Governale, Julie Olsson, et al. Efficacy and safety of lebrikizumab in adult patients with mild-to-moderate asthma not receiving inhaled corticosteroids. *Respiratory Medicine*, 134:143–149, 2018.

Franzi Korner-Nievergelt, Tobias Roth, Stefanie Von Felten, Jérôme Guélat, Bettina Almasi, and Pius Korner-Nievergelt. *Bayesian data analysis in ecology using linear models with R, BUGS, and Stan*. Academic Press, 2015.

Jan Kottner, Laurent Audigé, Stig Brorson, Allan Donner, Byron J Gajewski, Asbjørn Hróbjartsson, Chris Roberts, Mohamed Shoukri, and David L Streiner. Guidelines for reporting reliability and agreement studies (grras) were proposed. *International journal of nursing studies*, 48(6):661–671, 2011.

Albert Koulman, Geoffrey A Lane, Scott J Harrison, and Dietrich A Volmer. From differentiating metabolites to biomarkers. *Analytical and bioanalytical chemistry*, 394(3):663–670, 2009.

Jan S Krouwer. Why bland–altman plots should use x, not (y+ x)/2 when x is a reference method. *Statistics in medicine*, 27(5):778–780, 2008.

Nicholas B La Thangue and David J Kerr. Predictive biomarkers: a paradigm shift towards personalized cancer medicine. *Nature reviews Clinical oncology*, 8(10):587–596, 2011.

John M Lachin. Sample size and power for a logrank test and cox proportional hazards model with multiple groups and strata, or a quantitative covariate with multiple strata. *Statistics in medicine*, 32(25):4413–4425, 2013.

Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

G Lavezzari and AW Womack. Industry perspectives on biomarker qualification. *Clinical Pharmacology & Therapeutics*, 99(2):208–213, 2016.

I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

H Liu and DANIEL A Powers. Bayesian inference for zero-inflated poisson regression models. *Journal of Statistics: Advances in Theory and Applications*, 7(2):155–188, 2012.

Liying Lu, Yingzi Fu, Peixiao Chu, and Xiaolin Zhang. A bayesian analysis of zero-inflated count data: An application to youth fitness survey. In *Computational Intelligence and Security (CIS), 2014 Tenth International Conference on*, pages 699–703. IEEE, 2014.

Scott M Lynch. *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer Science & Business Media, 2007.

Peter McCullagh. Generalized linear models. *European Journal of Operational Research*, 16 (3):285–292, 1984.

Kenneth O McGraw and Seok P Wong. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30, 1996.

Sunali Mehta, Andrew Shelling, Anita Muthukaruppan, Annette Lasham, Cherie Blenkiron, George Laking, and Cristin Print. Predictive and prognostic molecular markers for cancer medicine. *Therapeutic advances in medical oncology*, 2(2):125–148, 2010.

IW Molenaar. A. gelman, jb carlin, hs stern, and db rubin. bayesian data analysis. *PSYCHOMETRIKA*, 62:285–286, 1997.

John Mullahy. Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365, 1986.

Reinhold Müller and Petra Büttner. A critical discussion of intraclass correlation coefficients. *Statistics in medicine*, 13(23-24):2465–2476, 1994.

John Ashworth Nelder and R Jacob Baker. *Generalized linear models*. Wiley Online Library, 1972.

Nancy A Obuchowski, Anthony P Reeves, Erich P Huang, Xiao-Feng Wang, Andrew J Buckler, Hyun J Kim, Huiman X Barnhart, Edward F Jackson, Maryellen L Giger, Gene Pennello, et al. Quantitative imaging biomarkers: A review of statistical methods for computer algorithm comparisons. *Statistical methods in medical research*, 24(1):68–106, 2015.

Stephen G Pauker and Jerome P Kassirer. Clinical application of decision analysis: a detailed illustration. In *Seminars in nuclear medicine*, volume 8, pages 324–335. Elsevier, 1978.

Mei-Yin C Polley, Samuel CY Leung, Lisa M McShane, Dongxia Gao, Judith C Hugh, Mauro G Mastropasqua, Giuseppe Viale, Lila A Zabaglo, Frédérique Penault-Llorca, John MS Bartlett, et al. An international ki67 reproducibility study. *Journal of the National Cancer Institute*, 105(24):1897–1906, 2013.

Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011.

Howard Raiffa. Applied statistical decision theory. 1974.

Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

Christian Robert and George Casella. A short history of markov chain monte carlo: Subjective recollections from incomplete data. *Statistical Science*, pages 102–115, 2011.

Christian P Robert. *Monte carlo methods*. Wiley Online Library, 2004.

Patrick Royston and Willi Sauerbrei. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in medicine*, 23(16):2509–2525, 2004.

Yuanbin Ru, Garrett M Dancik, and Dan Theodorescu. Biomarkers for prognosis and treatment selection in advanced bladder cancer patients. *Current opinion in urology*, 21(5): 420, 2011.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Mariano Ruiz Espejo. Design of experiments for engineers and scientists, 2006.

Daniel Sargent and Carmen Allegra. Issues in clinical trial design for tumor marker studies. In *Seminars in oncology*, volume 29, pages 222–230. Elsevier, 2002.

GL Serfontein and AM Jaroszewicz. Estimation of gestational age at birth. comparison of two methods. *Archives of Disease in Childhood*, 53(6):509–511, 1978.

Babak Shahbaba, Shiwei Lan, Wesley O Johnson, and Radford M Neal. Split hamiltonian monte carlo. *Statistics and Computing*, 24(3):339–349, 2014.

Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

Salman Siddiqui and Christopher E Brightling. Airways disease: phenotyping heterogeneity using measures of airway inflammation. *Allergy, Asthma & Clinical Immunology*, 3(2):60, 2007.

Richard Simon. Lost in translation: problems and pitfalls in translating laboratory observations to clinical utility. *European Journal of Cancer*, 44(18):2707–2713, 2008.

Richard Simon and Aboubakar Maitournam. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*, 10(20):6759–6763, 2004.

Xiao Song and Margaret Sullivan Pepe. Evaluating markers for selecting a patient's treatment. *Biometrics*, 60(4):874–883, 2004.

Kjetil Søreide. Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. *Journal of clinical pathology*, 62(1):1–5, 2009.

David J Spiegelhalter, Andrew Thomas, Nicky G Best, Wally Gilks, and D Lunn. Bugs: Bayesian inference using gibbs sampling. *Version 0.5,(version ii) http://www. mrc-bsu. cam. ac. uk/bugs*, 19, 1996.

Brian N Swanson. Delivery of high-quality biomarker assays. *Disease markers*, 18(2):47–56, 2002.

Parvin Tajik, Aleiko H Zwinderman, Ben W Mol, and Patrick M Bossuyt. Trial designs for personalizing cancer care: a systematic review and classification. *Clinical Cancer Research*, 19(17):4578–4588, 2013.

Sheila E Taube, Gary M Clark, Janet E Dancey, Lisa M McShane, Caroline C Sigman, and Steven I Gutman. A perspective on challenges and issues in biomarker development and drug and biomarker codevelopment. *Journal of the National Cancer Institute*, 2009.

Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

Giuseppe Viale, Anita Giobbie-Hurder, Meredith M Regan, Alan S Coates, Mauro G Mastropasqua, Patrizia Dell'Orto, Eugenio Maiorano, Gaëtan MacGrogan, Stephen G Braye,

Christian Öhlschlegel, et al. Prognostic and predictive value of centrally reviewed ki-67 labeling index in postmenopausal women with endocrine-responsive breast cancer: results from breast international group trial 1-98 comparing adjuvant tamoxifen with letrozole. *Journal of clinical oncology*, 26(34):5569, 2008.

Andrew J Vickers, Michael W Kattan, and Daniel J Sargent. Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials*, 8 (1):14, 2007.

Sally E Wenzel. Asthma: defining of the persistent adult phenotypes. *The Lancet*, 368(9537): 804–813, 2006.

Alice S Whittemore. Sample size for logistic regression with small response probability. *Journal of the American Statistical Association*, 76(373):27–32, 1981.

Stephen A Williams, David E Slavin, John A Wagner, and Christopher J Webster. A cost-effectiveness approach to the qualification and acceptance of biomarkers. *Nature Reviews Drug Discovery*, 5(11):897, 2006.

Rainer Winkelmann. *Econometric analysis of count data*. Springer Science & Business Media, 2008.

Rinat Yerushalmi, Ryan Woods, Peter M Ravdin, Malcolm M Hayes, and Karen A Gelmon. Ki67 in breast cancer: prognostic and predictive potential. *The lancet oncology*, 11(2): 174–183, 2010.

Salim Yusuf, Janet Wittes, Jeffrey Probstfield, and Herman A Tyroler. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Jama*, 266(1):93–98, 1991.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

GY Zou. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in medicine*, 31(29):3972–3981, 2012.

Appendix A

Chapter 4

A.0.1 Asymptotic Properties of Θ

Theorem: Let the biomarker X has a standard uniform distribution, i.e, $X \sim \mathcal{U}(0,1)$. Then $\Theta_0 = \frac{1}{n}\sum_{i=n}^{n} \Delta_i$ where $\Delta_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}$. Note that since X_i are *i.i.d.* random variables, then Δ_i are *i.i.d.* random variables. Also since , $-1 \leq \Delta_i \leq 1$, they have mean and variance. Let the mean and variance of Δ_is' be μ_Δ and σ^2_Δ respectively. Then

$$\begin{aligned} E\left[\sqrt{n}(\widehat{\Theta_0} - \mu_\Delta)\right] &\rightarrow 0 \\ Var\left[\sqrt{n}(\widehat{\Theta_0} - \mu_\Delta)\right] &\rightarrow \sigma^2_{\Theta_0} \end{aligned}$$

as $n \rightarrow \infty$.

Proof:

(i) Lets first look at the asymptotic claim for the expectation:

$$E\left[\sqrt{n}(\widehat{\Theta_0} - \mu_\Delta)\right] = E\left\{E\left[\sqrt{n}(\widehat{\Theta_0} - \mu_\Delta)|\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right]\right\}$$

Where the inner expectation is taken over the X_i and the out expectation is taken over the parameter estimates. Let

$$\begin{aligned} \eta_o &= E\left[\Delta_i|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right] \\ &= E\left[\widehat{\Theta_m}\right] \end{aligned}$$

then

$$E\left\{E\left[\sqrt{n}(\widehat{\Theta_m}-\mu_\Delta)|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right]\right\} = E\left[\sqrt{n}(\eta_o-\mu_\Delta)\right]$$

Nothing that η_o is a continous function of the maximum likelihood estimates (MLE), and the MLE converge to a normal distribution with mean zero, it follows that

$$E\left[\sqrt{n}(\eta_o-\mu_\Delta)\right] \to 0$$

as $n \to \infty$.

(ii) Now lets first look at the asymptotic claim for the variance:

As stated previously:

$$\begin{aligned} \Theta_0 &= \frac{1}{n}\sum_{i=n}^{n}\Delta_i \\ &= \frac{1}{n}\sum_{i=n}^{n}\left\{\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right\} \end{aligned} \tag{A.1}$$

Lets take the first term of Δ_i from Eq. (13) first.

$$VarE\left[\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)-\mu_\Delta\right\}\right] =$$
$$E\left[Var\left\{\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)-\mu_\Delta|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right\}\right\}\right] +$$

$$Var\left[E\left\{\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)-\mu_\Delta|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right)\right\}\right] \tag{A.2}$$

Not that, conditional on $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, the sums are sums of independent random variables. Taking the first term on the right hand side of Equation (14):

$$E\left[Var\left\{\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)-\mu_\Delta|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right)\right\}\right] =$$
$$E\left[\frac{1}{n}\sum_{i=n}^{n}Var\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right\}\right]$$

Since $-1 \leq \left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right) \leq 1$, the variance taken over X_i is absolutely bounded.

$$E\left[\frac{1}{n}\sum_{i=n}^{n}Var\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right\}\right] \to \sigma_o^2$$

Now turning to the second term in Equation (14)

$$Var\left[E\left\{\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)-\mu_\Delta|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right)\right\}\right]$$
$$= Var\left[\frac{1}{\sqrt{n}}n(\mu_n-\mu_\Delta)\right]$$
$$= Var\left[\frac{1}{\sqrt{n}}(\mu_n-\mu_\Delta)\right]$$

where $\mu_n = E\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right\} = \mu_n(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3)$.

The function $\mu_n(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3)$ is continuous in $(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3)$. Therefore, since $(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3)$ is maximum likelihood, under the usual condition s (ref) we have

$$\sqrt{n}\left[\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\\ \hat{\beta}_3\end{pmatrix}-\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\end{pmatrix}\right]\rightarrow Normal\sqrt{n}\left[\begin{pmatrix}0\\ 0\\ 0\\ 0\end{pmatrix},\Sigma_{\beta_0,\beta_1,\beta_2,\beta_3}\right]$$

It follow from (theorem ref) that

$$\sqrt{n}(\mu_n-\mu_\Delta)\rightarrow Normal(0,\sigma_\mu^2)$$

where $\sigma_\mu<\infty$

We have therefore shown that

$$Var[\sqrt{n}(\widehat{\Theta_m}-\mu_\Delta)]\rightarrow\sigma_\mu^2+\sigma_o^2=\sigma_{\Theta_0}^2$$

The proof for Θ_1 follows similarly.

This proof can easily be extended when the biomarker has a distribution other than uniform (Normal, Gamma etc).

A.0.2 End point calculation under optimal treatment for a biomarker with standard uniform distribution

Let the biomarker X has a continuous measure with probability density function $f(x)$. In Eq.(5) section 2.2, we have shown that:

$$T_{opt}(X=x) \;=\; \begin{cases} T=1 & :\beta_3 x<(-\beta_2)\\ T=0 & :\beta_3 x>(-\beta_2)\end{cases}$$

Depending on the sign of the coefficient β_3, the optimal treatment is further written as:

$$T_{opt}(X = x) \Rightarrow \texttt{if} : \beta_3 < 0 \begin{cases} T = 1 : & x > \frac{-\beta_2}{\beta_3} \\ T = 0 : & x \leq \frac{-\beta_2}{\beta_3} \end{cases}$$

$$T_{opt}(X = x) \Rightarrow \texttt{if} : \beta_3 > 0 \begin{cases} T = 1 : & x < \frac{-\beta_2}{\beta_3} \\ T = 0 : & x \geq \frac{-\beta_2}{\beta_3} \end{cases}$$

Let $\mathcal{A}_1 = \{x : \beta_3 x < (-\beta_2)\}$, and $\mathcal{A}_0 = \Re^1 \backslash \mathcal{A}_1$ (where "\" is the set difference symbol). The value of the end points for $\mathcal{A}_1$ & $\mathcal{A}_0$ will vary depending on the support of the probability density function for X. Let the end points for $\mathcal{A}_1$ & $\mathcal{A}_0$ are respectively (c, d) and (a, b) and $X \sim \mathcal{U}(0, 1)$. Then the possible end point values are as shown below.

Table A.1: Lower and Upper Integral limits calculation table for biomarker with U (0,1)

Parameters	Prob. Of Death/Relapse Picture	End Point Values
$\beta_3 > 0$, $0 < (-\beta_2/\beta_3) < 1$	Lines cross at $-\beta_2/\beta_3$ T=1 Prob. increase rel. to T=0	a = $(-\beta_2/\beta_3)$, b0 = 1 c = 0, d = $(-\beta_2/\beta_3)$
$\beta_3 > 0$, $0 < (\beta_2/\beta_3) < 1$	Lines cross at $-\beta_2/\beta_3$ T=1 Prob. increase rel. to T=0	a =0, b= (β_2/β_3) c =(β_2/β_3) , d = 1
$\beta_3 > 0$, $(\beta_2/\beta_3) < 0$	T=0 : Always better	a = 1, b = 0 c = d = 1
$\beta_3 < 0$, $(\beta_2/\beta_3) < 0$	T=1 : Always better	a = b = 0 c = 0, d = 1
$\beta_3 > 0$, $(\beta_2/\beta_3) < 1$	T=1 : Always better	a = b = 1 c = 0, d = 1
$\beta_3 < 0$, $(\beta_2/\beta_3) < 1$	T=0 : Always better	a = 0, b = 1 c = d = 1

Refere to figure 1

Similar development could be followed for other biomarker distributions considered.

A.0.3 Converting clinician inputs to model parameters

The natural approach to represent the relationship between the outcome ($Y \in \{0, 1\}$) and the covariates (T and X) along the interaction term ($T * X$) is using multiple logistic regression as:

$$Ln\left[\frac{Pr(Y=1|T=0,X)}{1-Pr(Y=1|T=0,X)}\right] = \beta_0 + \beta_1 x$$

$$
\begin{aligned}
& & = & \begin{cases} \beta_0 & : \beta_1 = 0 \\ \beta_0 + \beta_1 x & : \beta_1 \neq 0 \end{cases} \\
Ln\left[\frac{Pr(Y=1|T=1,X)}{1-Pr(Y=1|T=1,X)}\right] & = & \beta_0 + \beta_2 + (\beta_1+\beta_3)x \\
& = & \begin{cases} \beta_0 + \beta_2 & : \beta_1 + \beta_3 = 0 \\ \beta_0 + \beta_2 + (\beta_1+\beta_3)x & : \beta_1 + \beta_3 \neq 0 \end{cases}
\end{aligned}
$$

Let the clinician input values be K_1 to K_4. Restricting our focus to the 25^{th} and 75^{th} percentile of the marker value,we can write the following equations:

$$
\begin{aligned}
K_1 &= Ln\left[\frac{P(Y=1|A=0,x=F^{-1}(0.25))}{1-P(Y=1|A=0,x=F^{-1}(0.25))}\right] \\
&= \beta_0 + \beta_1 F^{-1}(-0.25) \\
K_2 &= Ln\left[\frac{P(Y=1|A=0,x=F^{-1}(0.75))}{1-P(Y=1|A=0,x=F^{-1}(0.75))}\right] \\
&= \beta_0 + \beta_1 F^{-1}(-0.75) \\
K_3 &= Ln\left[\frac{P(Y=1|A=1,x=F^{-1}(0.25))}{1-P(Y=1|A=1,x=F^{-1}(0.25))}\right] \\
&= \beta_0 + \beta_2 + (\beta_1+\beta_3) F^{-1}(-0.25) \\
K_4 &= Ln\left[\frac{P(Y=1|A=1,x=F^{-1}(0.75))}{1-P(Y=1|A=1,x=F^{-1}(0.75))}\right] \\
&= \beta_0 + \beta_2 + (\beta_1+\beta_3) F^{-1}(-0.75)
\end{aligned}
$$

If we let $F^{-1}(0.25)= z_1$, and $F^{-1}(0.75)= z_2$,

$$
\begin{aligned}
K_1 &= \beta_0 + \beta_1 z_1 \\
K_2 &= \beta_0 + \beta_1 z_2 \\
K_3 &= \beta_0 + \beta_2 + (\beta_1+\beta_3) z_1
\end{aligned}
$$

$$K_4 \;=\; \beta_0 + \beta_2 + (\beta_1 + \beta_3) z_2$$

Here we have four equations and four unknowns. Using MATLAB R2015a, the above equation can be solve in terms of the logistic model parameters as :

$$\beta_0 \;=\; \begin{cases} 0 & : K_1 = K_2 \\ \frac{K_2 * z_1 - K_1 * z_2}{z_1 - z_2} & : K_1 \neq K_2 \end{cases}$$

$$\beta_1 \;=\; \begin{cases} 0 & : K_1 = K_2 \\ \frac{K_1 - K_2}{z_1 - z_2} & : K_1 \neq K_2 \end{cases}$$

$$\beta_2 \;=\; \begin{cases} 0 & : K_1 = K_2, K_3 = K_4 \\ \frac{K_1 * z_2 - K_2 * z_1}{z_1 - z_2} & : K_1 \neq K_2, K_3 = K_4 \\ \frac{-K_3 * z_2 + K_4 * z_1}{z_1 - z_2} & : K_1 = K_2, K_3 \neq K_4 \\ \frac{K_1 * z_2 - K_2 * z_1 - K_3 * z_2 + K_4 * z_1}{z_1 - z_2} & : K_1 \neq K_2, K_3 \neq K_4 \end{cases}$$

$$\beta_3 \;=\; \begin{cases} 0 & : K_1 = K_2, K_3 = K_4 \\ \frac{K_2 - K_1}{z_1 - z_2} & : K_1 \neq K_2, K_3 = K_4 \\ \frac{K_3 - K_4}{z_1 - z_2} & : K_1 = K_2, K_3 \neq K_4 \\ \frac{K_2 - K_1 + K_3 - K_4}{z_1 - z_2} & : K_1 \neq K_2, K_3 \neq K_4 \end{cases}$$

Appendix B

Chapter 5

B.0.1 Asymptotic Properties of Δ_r

Theorem 1 : Let the biomarker X has a normal distribution, i.e, $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$. Using Monte Carlo, we get an estimator $\hat{\Theta}_x^T$ such that $\hat{\Theta}_x^T = \frac{1}{n}\sum_{i=n}^{n} \Theta_i^x$ where $\Theta_i^x = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}$. Note that since X_i are *i.i.d.* random variables, then Θ_i^x are *i.i.d.* random variables. Also since , $-1 \leq \Theta_i^x \leq 1$, they have mean and variance. Let the mean and variance of Θ_i^xs' be μ_{Θ_x} and $\sigma^2_{\Theta_x}$ respectively. Then

$$E\left[\sqrt{n}(\hat{\Theta}_x^T - \mu_{\Theta_x})\right] \rightarrow 0 \tag{B.1}$$

$$Var\left[\sqrt{n}(\hat{\Theta}_x^T - \mu_{\Theta_x})\right] \rightarrow \sigma^2_{\Theta_x} \tag{B.2}$$

as $n \rightarrow \infty$.

A detailed proofs of equations (18) and (19) is provided below:

Proof:

(i) Asymptotic convergence of the mean:

$$E\left[\sqrt{n}(\hat{\Theta}_x^T - \mu_{\Theta_x})\right] = E\left\{E\left[\sqrt{n}(\hat{\Theta}_x^T - \mu_{\Theta_x})|\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right]\right\}$$

where the inner expectation is taken over the X_i and the out expectation over the β parameter estimates. Now let,

$$\begin{aligned} \eta_x &= E\left[\Theta_i^x|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right] \\ &= E[\hat{\Theta}_m^x] \end{aligned}$$

then

$$E\left\{E\left[\sqrt{n}(\hat{\Theta}_m^x-\mu_{\Theta_x})|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right]\right\} = E\left[\sqrt{n}(\eta_x-\mu_{\Theta_x})\right]$$

$\hat{\beta}$s are maximum likelihood (MLE) estimates and η_x is a continuous function of these maximum likelihood estimates. Since maximum likehood estimates converge to a normal distribution with mean zero, it follows that;

$$E\left[\sqrt{n}(\eta_x-\mu_\Delta)\right] \to 0 \tag{B.3}$$

as $n \to \infty$.

(ii) Asymptotic convergence of the variance:

We know that,

$$\begin{aligned} \hat{\Theta}_x^T &= \frac{1}{n}\sum_{i=n}^{n}\Theta_i^x \\ &= \frac{1}{n}\sum_{i=n}^{n}\left\{\frac{e^{\beta_0+\beta_1X_i}}{1+e^{\beta_0+\beta_1X_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right\} \end{aligned} \tag{B.4}$$

Then the asymptotic variance of $\hat{\Theta}_x^T$ is given as:

$$\begin{aligned} &= Var\left[\sqrt{n}(\hat{\Theta}_x^T-\mu_{\Theta_x})\right] \\ &= Var\left[E\left[\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1X_i}}{1+e^{\beta_0+\beta_1X_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right) - \mu_{\Theta_x}\right\}\right]\right] \end{aligned}$$

$$= E\left[Var\left\{\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)-\mu_{\Theta_x}|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right)\right\}\right]+$$
$$Var\left[E\left\{\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)-\mu_{\Theta_x}|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right)\right\}\right] \quad \text{(B.5)}$$

Conditional on $\hat{\beta}s'$ the summations in equations (22) are sums of independent random variables. Now lets look at the first term of equation (22):

$$E\left[Var\left\{\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)-\mu_{\Theta_x}|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right)\right\}\right] =$$
$$E\left[\frac{1}{n}\sum_{i=n}^{n}Var\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right\}\right]$$

Since $-1 \leq \left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right) \leq 1$, the variance taken over X_i is absolutely bounded.

$$E\left[\frac{1}{n}\sum_{i=n}^{n}Var\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right\}\right] \to \sigma_o^2 \quad \text{(B.6)}$$

Now turning to the second term in equation (22)

$$Var\left[E\left\{\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)-\mu_{\Theta_x}|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right)\right\}\right]$$
$$= Var\left[\frac{1}{\sqrt{n}}n(\mu_n-\mu_{\Theta_x})\right]$$
$$= Var\left[\frac{1}{\sqrt{n}}(\mu_n-\mu_{\Theta_x})\right]$$

where $\mu_n = E\left\{\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right)|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\right\} = \mu_n(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3)$.

The function $\mu_n(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3)$ is continuous in $(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3)$. Therefore, since $(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3)$ is maximum likelihood, we have

$$\sqrt{n}\left[\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\\ \hat{\beta}_3\end{pmatrix}-\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\end{pmatrix}\right]\rightarrow Normal\sqrt{n}\left[\begin{pmatrix}0\\ 0\\ 0\\ 0\end{pmatrix},\Sigma_{\beta_0,\beta_1,\beta_2,\beta_3}\right]$$

It follow from (theorem ref) that

$$\sqrt{n}(\mu_n-\mu_{\Theta_x})\rightarrow Normal(0,\sigma_\mu^2) \tag{B.7}$$

where $\sigma_\mu<\infty$

Combining equations (23) and (24), we have shown that

$$Var[\sqrt{n}(\widehat{\Theta_m}-\mu_{\Theta_x}]\rightarrow\sigma_\mu^2+\sigma_o^2=\sigma_{\Theta_0}^2 \tag{B.8}$$

as $n\rightarrow\infty$.

Theorem 2 : Let the observed biomarker is W instead of X such that $W=X+U$ where $X\sim\mathcal{N}(\mu_x,\sigma_x^2)$ and $U\sim\mathcal{N}(0,\sigma_e^2)$. Since W is the sum of two normal, $W\sim\mathcal{N}(\mu_x,\sigma_w^2)$ such that $\sigma_w^2=\sigma_x^2+\sigma_e^2$. Using Monte Carlo, we get an estimator $\hat{\Theta}_w^T$ such that $\hat{\Theta}_w^T=\frac{1}{n}\sum_{i=n}^n\frac{1}{n}\sum_{i=n}^n\Theta_i^w$ where $\Theta_i^w=\frac{e^{\beta_0+\beta_1\tilde{X}_i}}{1+e^{\beta_0+\beta_1\tilde{X}_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}$ and $\tilde{x}_i=\mu_x+\hat{\rho}\frac{\hat{\sigma}_x}{\hat{\sigma}_w}(W_i-\mu_x)$. Note that since W_i are *i.i.d.* random variables, then Θ_i^w are *i.i.d.* random variables. Also since , $-1\leq\Theta_i^w\leq 1$, they have mean and variance. Let the mean and variance of Θ_i^ws' be μ_{Θ_w} and $\sigma_{\Theta_w}^2$ respectively. Then

$$E\left[\sqrt{n}(\hat{\Theta}_w^T-\mu_{\Theta_w})\right] \rightarrow 0 \tag{B.9}$$

$$Var\left[\sqrt{n}(\hat{\Theta}_w^T-\mu_{\Theta_w})\right] \rightarrow \sigma_{\Theta_w}^2 \tag{B.10}$$

as $n \to \infty$.

Asymptotic mean and variance proofs of equations (26) and (27) are given in detail below:

Proof:

(i) Asymptotic convergence of the mean:

$$E\left[\sqrt{n}(\hat{\Theta}_w^T - \mu_{\Theta_w})\right] = E\left\{E\left[\sqrt{n}(\hat{\Theta}_w^T - \mu_{\Theta_w})|\hat{\boldsymbol{B}}\right]\right\}$$

where $\hat{\boldsymbol{B}} = c(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$. The inner expectation is taken over the $\tilde{X}_i$ and the out expectation over the $\boldsymbol{B}$ parameter estimates. Now let,

$$\begin{aligned} \eta_w &= E\left[\Theta_i^w|\hat{\boldsymbol{B}}\right] \\ &= E[\hat{\Theta}_m^w] \end{aligned}$$

then

$$E\left\{E\left[\sqrt{n}(\hat{\Theta}_m^w - \mu_{\Theta_w})|\hat{\boldsymbol{B}}\right]\right\} = E\left[\sqrt{n}(\eta_w - \mu_{\Theta_w})\right]$$

$\hat{\boldsymbol{B}}$ is a vector of mle estimates and η_w is a continuous function of these mle estimates. Since mle estimates converge to a normal distribution with mean zero, it follows that;

$$E\left[\sqrt{n}(\eta_w - \mu_{\Theta_w})\right] \to 0 \tag{B.11}$$

as $n \to \infty$.

(ii) Asymptotic convergence of the variance:

We know that,

$$\hat{\Theta}_w^T = \frac{1}{n}\sum_{i=n}^{n}\sum_{i=n}^{n}\Theta_i^w$$

$$= \frac{1}{n}\sum_{i=n}^{n}\sum_{i=n}^{n}\left\{\frac{e^{\beta_0+\beta_1\tilde{X}_i}}{1+e^{\beta_0+\beta_1\tilde{X}_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}\right\} \tag{B.12}$$

Then the asymptotic variance of $\hat{\Theta}_w^T$ is given as:

$$\begin{aligned}
&= Var\left[\sqrt{n}(\hat{\Theta}_w^T - \mu_{\Theta_w})\right] \\
&= Var\left[E\left[\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1\tilde{X}_i}}{1+e^{\beta_0+\beta_1\tilde{X}_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}\right) - \mu_{\Theta_w}\right\}\right]\right] \\
&= E\left[Var\left\{\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1\tilde{X}_i}}{1+e^{\beta_0+\beta_1\tilde{X}_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}\right) - \mu_{\Theta_w}|\hat{\boldsymbol{B}}\right)\right\}\right] + \\
&\quad Var\left[E\left\{\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1\tilde{X}_i}}{1+e^{\beta_0+\beta_1\tilde{X}_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}\right) - \mu_{\Theta_w}|\hat{\boldsymbol{B}}\right)\right\}\right]
\end{aligned} \tag{B.13}$$

Conditional on $\hat{\boldsymbol{B}}$ the summations in equations (30) are sums of independent random variables. Now lets look at the first term of equation (30):

$$E\left[Var\left\{\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1\tilde{X}_i}}{1+e^{\beta_0+\beta_1\tilde{X}_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}\right) - \mu_{\Theta_w}|\hat{\boldsymbol{B}}\right)\right\}\right] = E\left[\frac{1}{n}\sum_{i=n}^{n}\sum_{i=n}^{n}Var\left\{\left(\frac{e^{\beta_0+\beta_1\tilde{X}_i}}{1+e^{\beta_0+\beta_1\tilde{X}_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}\right)|\hat{\boldsymbol{B}}\right\}\right]$$

Since $-1 \leq \left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_i}}\right) \leq 1$, the variance taken over $\tilde{X}_i$ is absolutely bounded.

$$E\left[\frac{1}{n}\sum_{i=n}^{n}\sum_{i=n}^{n}Var\left\{\left(\frac{e^{\beta_0+\beta_1\tilde{X}_i}}{1+e^{\beta_0+\beta_1\tilde{X}_i}} - \frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}\right)|\hat{\boldsymbol{B}}\right\}\right] \to \sigma_{w_1}^2 \tag{B.14}$$

Now turning to the second term in equation (22)

$$Var\left[E\left\{\frac{1}{\sqrt{n}}\sum_{i=n}^{n}\sum_{i=n}^{n}\left\{\left(\frac{e^{\beta_0+\beta_1\tilde{X}_i}}{1+e^{\beta_0+\beta_1\tilde{X}_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}\right)-\mu_{\Theta_w}|\hat{\boldsymbol{B}}\right)\right\}\right]$$
$$= Var\left[\frac{1}{\sqrt{n}}n(\mu_{w_2}-\mu_{\Theta_w})\right]$$
$$= Var\left[\frac{1}{\sqrt{n}}(\mu_{w_2}-\mu_{\Theta_w})\right]$$

where $\mu_{w_2} = E\left\{\left(\frac{e^{\beta_0+\beta_1\tilde{X}_i}}{1+e^{\beta_0+\beta_1\tilde{X}_i}}-\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}{1+e^{\beta_0+\beta_2+(\beta_1+\beta_3)\tilde{X}_i}}\right)|\hat{\boldsymbol{B}}\right\} = \mu_{w_2}(\hat{\boldsymbol{B}})$.

The function $\mu_{w_2}(\hat{\boldsymbol{B}})$ is continuous in $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$. Therefore, since $(\hat{\boldsymbol{B}})$ is maximum likelihood, we have

$$\sqrt{n}\left(\hat{\boldsymbol{B}}-\boldsymbol{B}\right) \rightarrow Normal\sqrt{n}\,(\mathbf{0}, \Sigma_{\boldsymbol{B}})$$

From this we can get;

$$\sqrt{n}(\mu_{w_1}-\mu_{\Theta_w}) \rightarrow Normal(0, \sigma^2_{w_2}) \tag{B.15}$$

where $\sigma_{w_2} < \infty$

From the results of equations (31) and (32), we have shown that

$$Var[\sqrt{n}(\hat{\Theta}^T_w-\mu_{\Theta_w}] \rightarrow \sigma^2_{w_1}+\sigma^2_{w_2} = \sigma^2_{\Theta_w} \tag{B.16}$$

as $n \rightarrow \infty$.

The asymptotic proof fro $\hat{\Delta}^T_r$, then follows from equations (20,25,28 and 33).

6.1 Supplementary Materials

Table S1: Posterior mean, standard error and 95% credible intervals of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ fitting a zero-inflated Poisson regression model assuming a $\mathcal{U}(0,1)$ biomarker. Data was generated from a standard Poisson model with sample size of 350. Coefficients used for data simulation are:$\beta_0 = -0.40$, $\beta_1 = 2.75$,$\beta_2 = 1.45$ and $\beta_3 = -3.00$.

Scenario 1: K-values $K_1 = 0.6$	$K_2 = 3.5$	$K_3 = 3.5$	$K_4 = 0.6$		
μ_{β_j}	$\beta_0 \sim N(-0.50, \sigma_j^2)$	$\beta_1 \sim N(1.76, \sigma_j^2)$	$\beta_2 \sim N(1.76, \sigma_j^2)$	$\beta_3 \sim N(-3.5, \sigma_j^2)$	
σ_j^2	0.5		10		100	
Post.	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.579(0.093)	(1.421,1.749)	1.596(0.095)	(1.431,1.767)	1.596(0.095)	(1.431,1.767)
Act	1.521(0.097)	(1.352,1.724)	1.529(0.099)	(1.358,1.738)	1.529(0.099)	(1.358,1.738)
Opt	0.959(0.056)	(0.854,1.075)	1.003(0.074)	(0.853,1.160)	1.003(0.074)	(0.852,1.160)
$\hat{\Psi}_{B_P}$	0.619(0.084)	(0.469,0.801)	0.594(0.094)	(0.425,0.792)	0.593(0.094)	(0.425,0.793)
$\hat{\Psi}_{B_T}$	0.563(0.073)	(0.428,0.687)	0.527(0.088)	(0.366,0.669)	0.526(0.088)	(0.368,0.669)

Scenario 2: K-values $K_1 = 0.6$	$K_2 = 3.5$	$K_3 = 2.5$	$K_4 = 1.5$		
μ_{β_j}	$\beta_0 \sim N(-0.50, \sigma_j^2)$	$\beta_1 \sim N(1.76, \sigma_j^2)$	$\beta_2 \sim N(1.42, \sigma_j^2)$	$\beta_3 \sim N(-2.27, \sigma_j^2)$	
σ_j^2	0.5		10		100	
Post.	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.584(0.093)	(1.422,1.750)	1.596(0.095)	(1.430,1.765)	1.596(0.095)	(1.431,1.765)
Act	1.519(0.098)	(1.353,1.724)	1.529(0.099)	(1.358,1.738)	1.529(0.099)	(1.358,1.736)
Opt	1.020(0.058)	(0.915,1.139)	1.003(0.074)	(0.851,1.161)	1.003(0.074)	(0.851,1.162)
$\hat{\Psi}_{B_P}$	0.563(0.083)	(0.411,0.738)	0.593(0.094)	(0.424,0.793)	0.593(0.095)	(0.425,0.792)
$\hat{\Psi}_{B_T}$	0.499(0.074)	(0.368,0.628)	0.526(0.088)	(0.367,0.669)	0.526(0.088)	(0.366,0.667)

Scenario 3: K-values $K_1 = 0.6$	$K_2 = 3.5$	$K_3 = 0.8$	$K_4 = 3.0$		
μ_{β_j}	$\beta_0 \sim N(-0.50, \sigma_j^2)$	$\beta_1 \sim N(1.76, \sigma_j^2)$	$\beta_2 \sim N(0.28, \sigma_j^2)$	$\beta_3 \sim N(-0.44, \sigma_j^2)$	
σ_j^2	0.5		10		100	
Post.	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.609(0.095)	(1.438,1.769)	1.596(0.095)	(1.431,1.765)	1.596(0.095)	(1.430,1.767)
Act	1.506(0.099)	(1.337,1.716)	1.529(0.099)	(1.359,1.738)	1.529(0.099)	(1.358,1.737)
Opt	1.131(0.062)	(1.024,1.259)	1.003(0.074)	(0.852,1.160)	1.003(0.074)	(0.852,1.162)
$\hat{\Psi}_{B_P}$	0.479(0.084)	(0.317,0.652)	0.593(0.094)	(0.424,0.792))	0.593(0.094)	(0.424,0.793)
$\hat{\Psi}_{B_T}$	0.376(0.073)	(0.244,0.515)	0.526(0.088)	(0.366,0.666)	0.526(0.088)	(0.366,0.668)

Table S2: Posterior mean, standard error and 95% credible intervals of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ fitting a negative binomial regression model assuming a $\mathcal{U}(0,1)$ biomarker. Data was generated from a standard Poisson model with sample size of 350. Coefficients used for data simulation are:$\beta_0 = -0.40$, $\beta_1 = 2.75$,$\beta_2 = 1.45$ and $\beta_3 = -3.00$.

Scenario 1: K-values	$K_1 = 0.6$	$K_2 = 3.5$	$K_3 = 3.5$	$K_4 = 0.6$	
μ_{β_j}	$\beta_0 \sim N(-0.50, \sigma_j^2)$	$\beta_1 \sim N(1.76, \sigma_j^2)$	$\beta_2 \sim N(1.76, \sigma_j^2)$	$\beta_3 \sim N(-3.5, \sigma_j^2)$	
σ_j^2	0.5		10	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.536(0.084)	(1.386,1.684)	1.549(0.086)	(1.394,1.695)	1.549(0.086)	(1.394,1.696)
Act	1.482(0.093)	(1.314,1.677)	1.486(0.096)	(1.315,1.685)	1.486(0.095)	(1.316,1.686)
Opt	0.931(0.053)	(0.821,1.046)	0.970(0.068)	(0.815,1.118)	0.971(0.068)	(0.816,1.118)
$\hat{\Psi}_{B_P}$	0.605(0.079)	(0.461,0.767)	0.579(0.089)	(0.420,0.771)	0.578(0.090	(0.419,0.772)
$\hat{\Psi}_{B_T}$	0.550(0.071)	(0.422,0.679)	0.516(0.086)	(0.364,0.652)	0.516(0.086)	(0.363,0.651)

Scenario 2: K-values	$K_1 = 0.6$	$K_2 = 3.5$	$K_3 = 2.5$	$K_4 = 1.5$	
μ_{β_j}	$\beta_0 \sim N(-0.50, \sigma_j^2)$	$\beta_1 \sim N(1.76, \sigma_j^2)$	$\beta_2 \sim N(1.42, \sigma_j^2)$	$\beta_3 \sim N(-2.27, \sigma_j^2)$	
σ_j^2	0.5		10	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.537(0.084)	(1.387,1.682)	1.549(0.085)	(1.394,1.695)	1.549(0.086)	(1.394,1.695)
Act	1.476(0.094)	(1.312,1.674)	1.486(0.096)	(1.315,1.686)	1.486(0.096)	(1.315,1.686)
Opt	0.989(0.054)	(0.879,1.108)	0.971(0.068)	(0.816,1.117)	0.971(0.068)	(0.815,1.118)
$\hat{\Psi}_{B_P}$	0.548(0.079)	(0.400,0.712)	0.578(0.090)	(0.421,0.770)	0.578(0.090)	(0.420,0.771)
$\hat{\Psi}_{B_T}$	0.486(0.071)	(0.358,0.619)	0.515(0.086)	(0.364,0.652)	0.515(0.086)	(0.363,0.652)

Scenario 3: K-values	$K_1 = 0.6$	$K_2 = 3.5$	$K_3 = 0.8$	$K_4 = 3.0$	
μ_{β_j}	$\beta_0 \sim N(-0.50, \sigma_j^2)$	$\beta_1 \sim N(1.76, \sigma_j^2)$	$\beta_2 \sim N(0.28, \sigma_j^2)$	$\beta_3 \sim N(-0.44, \sigma_j^2)$	
σ_j^2	0.5		10	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.554(0.084)	(1.404,1.699)	1.549(0.085)	(1.394,1.695)	1.549(0.086)	(1.395,1.695)
Act	1.454(0.096)	(1.287,1.656)	1.486(0.096)	(1.316,1.687)	1.486(0.096)	(1.316,1.685)
Opt	1.091(0.056)	(0.975,1.215)	0.971(0.068)	(0.816,1.121)	0.971(0.068)	(0.816,1.118)
$\hat{\Psi}_{B_P}$	0.463(0.079)	(0.310,0.629)	0.578(0.090)	(0.419,0.773)	0.579(0.090)	(0.419,0.771)
$\hat{\Psi}_{B_T}$	0.363(0.071)	(0.239,0.494)	0.515(0.086)	(0.363,0.652)	0.516(0.086)	(0.363,0.651)

Table S3: Posterior mean, standard error and 95% credible intervals of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ fitting a zero-inflated Poisson regression model assuming a $\mathcal{N}(4.8, 3.24)$ biomarker. Data was generated from a standard Poisson model with sample size of 350. Coefficients used for data simulation are: $\beta_0 = -0.10$, $\beta_1 = 0.08$,$\beta_2 = 0.65$ and $\beta_3 = -0.15$.

Scenario 1: K-values	$K_1 = 0.6$	$K_2 = 3.5$	$K_3 = 3.5$	$K_4 = 0.6$	
μ_{β_j}	$\beta_0 \sim N(-0.15, \sigma_j^2)$	$\beta_1 \sim N(0.11, \sigma_j^2)$	$\beta_2 \sim N(1.04, \sigma_j^2)$	$\beta_3 \sim N(-0.22, \sigma_j^2)$	
σ_j^2	0.2		5	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.330(0.075)	(1.189,1.481)	1.385(0.081)	(1.229,1.548)	1.385(0.081)	(1.229,1.548)
Act	1.292(0.073)	(1.151,1.412)	1.312(0.078)	(1.167,1.452)	1.312(0.078)	(1.167,1.452)
Opt	1.114(0.050)	(1.017,1.212)	1.177(0.083)	(1.048,1.297)	1.177(0.083)	(1.048,1.297)
$\hat{\Psi}_{B_P}$	0.217(0.075)	(0.101,0.350)	0.208(0.083)	(0.081,0.349)	0.208(0.083)	(0.081,0.349)
$\hat{\Psi}_{B_T}$	0.178(0.044)	(0.103,0.2750	0.135(0.066)	(0.024,0.279)	0.135(0.066)	(0.024,0.279)

Scenario 2: K-values	$K_1 = 0.6$	$K_2 = 3.0$	$K_3 = 2.5$	$K_4 = 0.6$	
μ_{β_j}	$\beta_0 \sim N(-0.18, \sigma_j^2)$	$\beta_1 \sim N(0.10, \sigma_j^2)$	$\beta_2 \sim N(0.80, \sigma_j^2)$	$\beta_3 \sim N(-0.18, \sigma_j^2)$	
σ_j^2	0.2		5	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.339(0.075)	(1.196,1.491)	1.386(0.081)	(1.231,1.548)	1.386(0.081)	(1.231,1.549)
Act	1.276(0.073)	(1.139,1.397)	1.312(0.078)	(1.169,1.452)	1.313(0.078)	(1.167,1.452)
Opt	1.135(0.051)	(1.032,1.234)	1.178(0.069)	(1.050,1.299)	1.178(0.069)	(1.048,1.298)
$\hat{\Psi}_{B_P}$	0.205(0.077)	(0.087,0.343)	0.208(0.083)	(0.079,0.349)	0.208(0.083)	(0.081,0.350)
$\hat{\Psi}_{B_T}$	0.142(0.042)	(0.074,0.239)	0.135(0.065)	(0.029,0.279)	0.135(0.066)	(0.028,0.281)

Scenario 3: K-values	$K_1 = 0.6$	$K_2 = 3.0$	$K_3 = 0.8$	$K_4 = 2.5$	
μ_{β_j}	$\beta_0 \sim N(-0.18, \sigma_j^2)$	$\beta_1 \sim N(0.10, \sigma_j^2)$	$\beta_2 \sim N(0.19, \sigma_j^2)$	$\beta_3 \sim N(-0.03, \sigma_j^2)$	
σ_j^2	0.2		5	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.370(0.075)	(1.223,1.523)	1.386(0.081)	(1.231,1.548)	1.386(0.081)	(1.231,1.548)
Act	1.246(0.073)	(1.114,1.365)	1.312(0.078)	(1.167,1.451)	1.312(0.078)	(1.168,1.452)
Opt	1.186(0.056)	(1.069,1.287)	1.178(0.069)	(1.050,1.296)	1.177(0.069)	(1.050,1.296)
$\hat{\Psi}_{B_P}$	0.184(0.087)	(0.059,0.344)	0.208(0.083)	(0.076,0.349)	0.208(0.083)	(0.075,0.349)
$\hat{\Psi}_{B_T}$	0.059(0.034)	(0.015,0.151)	0.135(0.066)	(0.027,0.278)	0.135(0.066)	(0.030,0.279)

Table S4: Posterior mean, standard error and 95% credible intervals of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ fitting a negative binomial regression model assuming a $\mathcal{N}(4.8, 3.24)$ biomarker. Data was generated from a standard Poisson model with sample size of 350. Coefficients used for data simulation are: $\beta_0 = -0.10$, $\beta_1 = 0.08$,$\beta_2 = 0.65$ and $\beta_3 = -0.15$.

Scenario 1: K-values	$K_1 = 0.6$	$K_2 = 3.5$	$K_3 = 3.5$	$K_4 = 0.6$		
μ_{β_j}	$\beta_0 \sim N(-0.15, \sigma_j^2)$	$\beta_1 \sim N(0.11, \sigma_j^2)$	$\beta_2 \sim N(1.04, \sigma_j^2)$	$\beta_3 \sim N(-0.22, \sigma_j^2)$		
σ_j^2	0.2		5		100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.291(0.071)	(1.155,1.424)	1.334(0.076)	(1.192,1.471)	1.335(0.076)	(1.192,1.471)
Act	1.253(0.071)	(1.122,1.379)	1.264(0.075)	(1.127,1.410)	1.264(0.075)	(1.127,1.411)
Opt	1.079(0.048)	(0.992,1.166)	1.132(0.065)	(1.009,1.242)	1.132(0.066)	(1.009,1.245)
$\hat{\Psi}_{B_P}$	0.211(0.073)	(0.101,0.340)	0.202(0.081)	(0.077,0.339)	0.202(0.081)	(0.078,0.338)
$\hat{\Psi}_{B_T}$	0.173(0.043)	(0.102,0.269)	0.132(0.065)	(0.028,0.274)	0.131(0.065)	(0.029,0.274)

Scenario 2: K-values	$K_1 = 0.6$	$K_2 = 3.0$	$K_3 = 2.5$	$K_4 = 0.6$		
μ_{β_j}	$\beta_0 \sim N(-0.18, \sigma_j^2)$	$\beta_1 \sim N(0.10, \sigma_j^2)$	$\beta_2 \sim N(0.80, \sigma_j^2)$	$\beta_3 \sim N(-0.18, \sigma_j^2)$		
σ_j^2	0.2		5		100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.230(0.071)	(1.161,1.432)	1.335(0.076)	(1.192,1.471)	1.335(0.076)	(1.191,1.473)
Act	1.238(0.071)	(1.109,1.363)	1.264(0.075)	(1.126,1.411)	1.264(0.075)	(1.126,1.411)
Opt	1.100(0.049)	(1.010,1.187)	1.132(0.065)	(1.009,1.242)	1.132(0.066)	(1.009,1.245)
$\hat{\Psi}_{B_P}$	0.199(0.076)	(0.086,0.336)	0.202(0.081)	(0.075,0.336)	0.202(0.081)	(0.076,0.339)
$\hat{\Psi}_{B_T}$	0.138(0.041)	(0.072,0.233)	0.132(0.064)	(0.031,0.274)	0.132(0.065)	(0.028,0.274)

Scenario 3: K-values	$K_1 = 0.6$	$K_2 = 3.0$	$K_3 = 0.8$	$K_4 = 2.5$		
μ_{β_j}	$\beta_0 \sim N(-0.18, \sigma_j^2)$	$\beta_1 \sim N(0.10, \sigma_j^2)$	$\beta_2 \sim N(0.19, \sigma_j^2)$	$\beta_3 \sim N(-0.03, \sigma_j^2)$		
σ_j^2	0.2		5		100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	1.328(0.072)	(1.183,1.458)	1.335(0.076)	(1.191,1.471)	1.335(0.076)	(1.191,1.471)
Act	1.207(0.072)	(1.082,1.332)	1.262(0.075)	(1.126,1.413)	1.264(0.075)	(1.126,1.411)
Opt	1.149(0.054)	(1.044,1.237)	1.132(0.066)	(1.010,1.242)	1.132(0.066)	(1.009,1.246)
$\hat{\Psi}_{B_P}$	0.179(0.085)	(0.053,0.336)	0.203(0.081)	(0.075,0.339)	0.202(0.081)	(0.076,0.339)
$\hat{\Psi}_{B_T}$	0.059(0.033)	(0.016,0.146)	0.132(0.064)	(0.031,0.274)	0.132(0.064)	(0.028,0.275)

Table S5: Posterior mean, standard error and 95% credible intervals of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ fitting a zero-inflated Poisson regression model to mirror the AA clinical trial study. Data was generated from a standard Poisson model with sample size of 460. Coefficients used for data simulation are: $\beta_0 = -9.85$, $\beta_1 = 2.20$,$\beta_2 = 5.33$ and $\beta_3 = -1.52$.

Scenario 1: K-values $K_1 = 0.10$	$K_2 = 2.50$	$K_3 = 1.50$	$K_4 = 0.40$		
μ_{β_j}	$\beta_0 \sim N(-3.72, \sigma_j^2)$	$\beta_1 \sim N(0.78, \sigma_j^2)$	$\beta_2 \sim N(5.01, \sigma_j^2)$	$\beta_3 \sim N(-1.27, \sigma_j^2)$	
σ_j^2	0.5		10	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	0.475(0.044)	(0.396,0.555)	0.501(0.043)	(0.418,0.588)	0.500(0.043)	(0.417,0.588)
Act	0.194(0.032)	(0.133,0.270)	0.226(0.035)	(0.165,0.306)	0.226(0.035)	(0.166,0.308)
Opt	0.182(0.027)	(0.129,0.242)	0.172(0.030)	(0.126,0.245)	0.171(0.031)	(0.124,0.244)
$\hat{\Psi}_{B_P}$	0.292(0.046)	(0.212,0.384)	0.329(0.039)	(0.258,0.404)	0.330(0.039)	(0.258,0.404)
$\hat{\Psi}_{B_T}$	0.012(0.005)	(0.003,0.028)	0.054(0.020)	(0.021,0.105)	0.055(0.021)	(0.022,0.107)

Scenario 2: K-values $K_1 = 0.10$	$K_2 = 2.50$	$K_3 = 1.00$	$K_4 = 0.30$		
μ_{β_j}	$\beta_0 \sim N(-3.72, \sigma_j^2)$	$\beta_1 \sim N(0.78, \sigma_j^2)$	$\beta_2 \sim N(4.24, \sigma_j^2)$	$\beta_3 \sim N(-1.07, \sigma_j^2)$	
σ_j^2	0.5		10	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	0.477(0.044)	(0.396,0.558)	0.501(0.043)	(0.418,0.587)	0.500(0.043)	(0.418,0.588)
Act	0.193(0.032)	(0.132,0.268)	0.226(0.035)	(0.166,0.306)	0.226(0.035)	(0.166,0.308)
Opt	0.186(0.029)	(0.130,0.249)	0.172(0.030)	(0.125,0.245)	0.171(0.031)	(0.124,0.244)
$\hat{\Psi}_{B_P}$	0.291(0.048)	(0.209,0.385)	0.329(0.039)	(0.258,0.404)	0.330(0.039)	(0.258,0.404)
$\hat{\Psi}_{B_T}$	0.008(0.004)	(0.002,0.021)	0.054(0.020)	(0.021,0.104)	0.055(0.021)	(0.022,0.107)

Scenario 3: K-values $K_1 = 0.1$	$K_2 = 2.50$	$K_3 = 0.40$	$K_4 = 1.25$		
μ_{β_j}	$\beta_0 \sim N(-3.72, \sigma_j^2)$	$\beta_1 \sim N(0.78, \sigma_j^2)$	$\beta_2 \sim N(2.30, \sigma_j^2)$	$\beta_3 \sim N(-0.51, \sigma_j^2)$	
σ_j^2	0.5		10	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	0.483(0.046)	(0.402,0.568)	0.501(0.043)	(0.418,0.588)	0.500(0.043)	(0.417,0.588)
Act	0.193(0.033)	(0.129,0.265)	0.226(0.035)	(0.166,0.307)	0.226(0.035)	(0.166,0.307)
Opt	0.192(0.032)	(0.128,0.262)	0.173(0.030)	(0.125,0.245)	0.171(0.031)	(0.124,0.245)
$\hat{\Psi}_{B_P}$	0.291(0.052)	(0.201,0.393)	0.328(0.039)	(0.257,0.404)	0.330(0.039)	(0.258,0.404)
$\hat{\Psi}_{B_T}$	0.002(0.001)	(0.000,0.006)	0.053(0.020)	(0.021,0.103)	0.055(0.021)	(0.022,0.107)

Table S6: Posterior mean, standard error and 95% credible intervals of $\hat{\Psi}_{B_P}$ and $\hat{\Psi}_{B_T}$ fitting a negative binomial regression model to mirror the AA clinical trial study. Data was generated from a standard Poisson model with sample size of 460. Coefficients used for data simulation are: $\beta_0 = -9.85$, $\beta_1 = 2.20$,$\beta_2 = 5.33$ and $\beta_3 = -1.52$.

Scenario 1: K-values $K_1 = 0.10$	$K_2 = 2.50$	$K_3 = 1.50$	$K_4 = 0.40$		
μ_{β_j}	$\beta_0 \sim N(-3.72, \sigma_j^2)$	$\beta_1 \sim N(0.78, \sigma_j^2)$	$\beta_2 \sim N(5.01, \sigma_j^2)$	$\beta_3 \sim N(-1.27, \sigma_j^2)$	
σ_j^2	0.5		10	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	0.438(0.044)	(0.349,0.517)	0.464(0.040)	(0.387,0.539)	0.464(0.040)	(0.387,0.539)
Act	0.179(0.029)	(0.122,0.248)	0.205(0.031)	(0.151,0.276)	0.205(0.031)	(0.151,0.276)
Opt	0.169(0.025)	(0.119,0.224)	0.155(0.027)	(0.111,0.220)	0.154(0.027)	(0.110,0.219)
$\hat{\Psi}_{B_P}$	0.269(0.044)	(0.191,0.344)	0.308(0.037)	(0.231,0.379)	0.309(0.037)	(0.232,0.381)
$\hat{\Psi}_{B_T}$	0.011(0.005)	(0.004,0.026)	0.049(0.018)	(0.019,0.093)	0.051(0.019)	(0.020,0.094)

Scenario 2: K-values $K_1 = 0.10$	$K_2 = 2.50$	$K_3 = 1.00$	$K_4 = 0.30$		
μ_{β_j}	$\beta_0 \sim N(-3.72, \sigma_j^2)$	$\beta_1 \sim N(0.78, \sigma_j^2)$	$\beta_2 \sim N(4.24, \sigma_j^2)$	$\beta_3 \sim N(-1.07, \sigma_j^2)$	
σ_j^2	0.5		10	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	0.439(0.044)	(0.349,0.519)	0.464(0.040)	(0.387,0.539)	0.464(0.040)	(0.388,0.540)
Act	0.179(0.029)	(0.121,0.247)	0.205(0.031)	(0.151,0.277)	0.205(0.031)	(0.151,0.277)
Opt	0.172(0.026)	(0.119,0.229)	0.155(0.027)	(0.111,0.221)	0.154(0.027)	(0.120,0.219)
$\hat{\Psi}_{B_P}$	0.267(0.046)	(0.187,0.344)	0.308(0.037)	(0.230,0.379)	0.310(0.037)	(0.232,0.380)
$\hat{\Psi}_{B_T}$	0.007(0.004)	(0.001,0.019)	0.050(0.018)	(0.020,0.093)	0.051(0.019)	(0.020,0.094)

Scenario 3: K-values $K_1 = 0.1$	$K_2 = 2.50$	$K_3 = 0.40$	$K_4 = 1.25$		
μ_{β_j}	$\beta_0 \sim N(-3.72, \sigma_j^2)$	$\beta_1 \sim N(0.78, \sigma_j^2)$	$\beta_2 \sim N(2.30, \sigma_j^2)$	$\beta_3 \sim N(-0.51, \sigma_j^2)$	
σ_j^2	0.5		10	100	
Posterior	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**	**mean(se)**	**95% CI**
soc	0.439(0.045)	(0.345,0.523)	0.463(0.040)	(0.387,0.540)	0.464(0.040)	(0.388,0.540)
Act	0.177(0.029)	(0.118,0.243)	0.205(0.031)	(0.150,0.276)	0.205(0.031)	(0.151,0.277)
Opt	0.176(0.028)	(0.118,0.239)	0.156(0.027)	(0.112,0.221)	0.154(0.027)	(0.110,0.219)
$\hat{\Psi}_{B_P}$	0.264(0.049)	(0.179,0.345)	0.308(0.037)	(0.230,0.379)	0.310(0.037)	(0.232,0.381)
$\hat{\Psi}_{B_T}$	0.001(0.001)	(0.000,0.006)	0.050(0.018)	(0.020,0.092)	0.051(0.019)	(0.020,0.094)

Table S7: Confidence interval width and coverage probability comparison for our method and Janes method using bootstrap: 1000 Monte Carlo each with 1000 sample size. Biomarker has a standard uniform distribution U(0,1).

k_1	k_2	k_3	k_4	**Estimate**	**Method**	**95% CI**	**Width of CI**	**Coverage**
0.25	0.75	0.75	0.25	0.2353	Boot Nor.	(0.1990 , 0.2668)	0.0677	0.95
					Boot Bas.	(0.1986 , 0.2672)	0.0686	0.95
					Janes Emp.	(0.1909 , 0.2739)	0.083	0.97
					Janes Mod.	(0.1954 , 0.3702)	0.0766	0.94
0.1	0.9	0.9	0.1	0.3456	Boot Nor.	(0.3189 , 0.3700)	0.0511	0.94
					Boot Bas.	(0.3183 , 0.3702)	0.0519	0.96
					Janes Emp.	(0.3059 , 0.3824)	0.0771	0.94
					Janes Mod.	(0.3101 , 0.3819)	0.0712	0.94
0.7	0.3	0.85	0.15	0.0595	Boot Nor.	(0.0275 , 0.0926)	0.065	0.94
					Boot Bas.	(0.0254 , 0.0912)	0.0657	0.96
					Janes Emp.	(0.0231 , 0.1007)	0.0775	0.93
					Janes Mod.	(0.0316 , 0.0962)	0.0646	0.95
0.1	0.55	0.9	0.45	0.4286	Boot Nor.	(0.3943 , 0.4651)	0.0688	0.95
					Boot Bas.	(0.3937 , 0.4636)	0.0699	0.96
					Janes Emp.	(0.3839 , 0.4741)	0.0901	0.95
					Janes Mod.	(0.3890 , 0.4716)	0.0826	0.93

Table S8: Confidence interval width and coverage probability comparison for our method and Janes method using bootstrap: 1000 Monte Carlo each with 1000 sample size. Biomarker has a standard normal distribution N(0,1).

K_1	K_2	K_3	K_4	Θ_1	**Method**	**95% CI**	**CI Width**	**Coverage**
0.25	0.75	0.75	0.25	0.247	Boot Norm.	(0.214 , 0.279)	0.065	0.970
					Boot Perc.	(0.215 , 0.281)	0.066	0.970
					Janes Emp.	(0.205 , 0.287)	0.082	0.980
					Janes Mod.	(0.211 , 0.287)	0.076	0.960
0.10	0.90	0.90	0.10	0.348	Boot Norm.	(0.322 , 0.373)	0.051	0.940
					Boot Perc.	(0.323 , 0.375)	0.052	0.940
					Janes Emp.	(0.309 , 0.386)	0.077	0.950
					Janes Mod.	(0.314 , 0.386)	0.072	0.940
0.10	0.55	0.90	0.45	0.441	Boot Norm.	(0.406 , 0.476)	0.070	0.970
					Boot Perc.	(0.406 , 0.477)	0.071	0.970
					Janes Emp.	(0.398 , 0.487)	0.089	0.980
					Janes Mod.	(0.401 , 0.486)	0.084	0.970
0.25	0.75	0.50	0.50	0.124	Boot Norm.	(0.089 , 0.160)	0.072	0.980
					Boot Perc.	(0.089 , 0.163)	0.073	0.970
					Janes Emp.	(0.081 , 0.167)	0.085	0.980
					Janes Mod.	(0.089 , 0.165)	0.075	0.980
0.90	0.45	0.10	0.55	0.125	Boot Norm.	(0.102 , 0.151)	0.049	0.960
					Boot Perc.	(0.103 , 0.153)	0.049	0.960
					Janes Emp.	(0.092 , 0.158)	0.066	0.970
					Janes Mod.	(0.098 , 0.155)	0.057	0.960
0.60	0.40	0.50	0.50	0.056	Boot Norm.	(0.019 , 0.098)	0.079	0.970
					Boot Perc.	(0.026 , 0.105)	0.078	0.970
					Janes Emp.	(0.015 , 0.105)	0.090	0.970
					Janes Mod.	(0.025 , 0.103)	0.078	0.960