Evaluation of Predictive Accuracy of Tests and Impact of Tests on Patient Outcomes

By

Mun Sang Yue

M.Eng., Imperial College of Science, Technology & Medicine, 1998

M.Sc., Stanford University, 2005

M.Sc., State University of New York at Stony Brook, 2012

Dissertation

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Biostatistics at Brown University

PROVIDENCE, RHODE ISLAND

MAY 2017

This dissertation by Mun Sang Yue is accepted in its present form
by the Department of Biostatistics as satisfying the
dissertation requirements for the degree of Doctor of Philosophy.

Date _____ _____

Constantine A. Gatsonis, Advisor

Recommended to the Graduate Council

Date _____ _____

Christopher Schmid, Reader

Date _____ _____

Tao Liu, Reader

Approved by the Graduate Council

Date _____ _____

Andrew G. Campbell, Dean of the Graduate School

Preface

The use of diagnostic tests and biomarkers is an essential part of medical care, and plays an important role in guiding therapy decisions in the era of precision medicine. In this dissertation, we address two major aspects of test evaluation, the assessment of predictive accuracy and the assessment of the impact of tests on patient outcomes. The predictive accuracy of tests is addressed in Chapters 1 and 2, while the impact of tests on patient outcomes is addressed in Chapter 3.

In the practice of evidence based medicine, the ability to synthesize evidence from primary studies of biomarkers is useful in optimizing health policy decision making. However methodological developments in the area of synthesizing predictive values have been limited. In Chapter 1, we put forth a new meta-analysis model to synthesize and compare predictive values of biomarkers. This model provides a joint summary assessment of the positive and negative predictive values.

In Chapter 2, we undertake a critical evaluation of the widespread use of hazard ratio as a summary measure of the prognostic performances of biomarkers. From the results of this study, we obtain a better understanding of the implications of using hazard ratio to summarize, and compare prognostic performances of biomarkers. This study also identifies essential information that should accompany the reporting of hazard ratio to allow proper evaluation of the prognostic performances of a biomarker.

A key challenge in evaluating the impact of diagnostic tests on patient outcomes, such

as morbidity, mortality, and health related quality of life, is that the pathway from test to outcomes typically involves subsequent disease management and treatment interventions. Modeling approaches, such as decision analysis and micro-simulation, are commonly used to study the impact of tests. Randomized studies (also known as diagnostic randomized controlled trials, DRCT) have also been utilized, but to a lesser extent than modeling. In addition to the large sample size typically required, DRCT studies are also prone to selection bias arising from noncompliance by study participants to assigned tests and interventions. Recent work has laid out the formal framework for evaluating DRCT designs, and derived formulas for sample size and power computations. However the impact of noncompliance has not been addressed. In Chapter 3, we adapt and apply modern methods in causal inference to estimate the causal outcomes of diagnostic tests in the presence of noncompliance. The performance of these causal estimates are evaluated via simulation of different scenarios.

Contents

List of Tables

List of Figures

Chapter 1

Meta-analysis of predictive values of biomarkers

Abstract

The evaluation of the predictive performance of biomarkers is a vital and growing area of research in precision medicine. However, statistical methods for meta-analysis of the predictive accuracy of tests, as measured by the positive and negative predictive values (PPV and NPV respectively), have received limited attention in the literature, in contrast to methods for meta-analysis of diagnostic accuracy. In this chapter, we propose a hierarchical summary predictive ROC (HSPROC) curve model to summarize estimates of PPV, NPV and disease prevalence jointly. The model accounts for the relationship between PPV and NPV stemming from the dependence on the threshold for test positivity, and also preserves the monotonicity of the summary predictive ROC curve. The HSPROC curves generated from the model can be used for comparison of different biomarkers. We applied the proposed method to two examples from the literature. The first is a meta-analysis of prognostic capabilities of biomarkers for rapid rule-out of acute myocardial infarction, and the second is re-

lated to biomarkers for acute pulmonary embolism. In both examples, comparisons of prognostic capabilities of the different biomarkers considered are illustrated.

1.1 Introduction

The use of biomarkers is an essential tool of precision medicine and has given rise to the rapidly growing literature of biomarker evaluation studies. A key objective of biomarker evaluation studies is the assessment of the predictive performance, with the hazard ratio as a commonly used metric for prediction of time-to-event outcomes (Altman et al., 2012), and the positive and negative predictive values (PPV and NPV respectively) as metrics for prediction of binary outcomes (Shiu and Gatsonis, 2008, Bossuyt et al., 2015). Recent examples of biomarker evaluation studies include Efficace et al. (2015), Chung et al. (2014), Anitei et al. (2014), Liao et al. (2012), Sørensen et al. (2013), Biliavska et al. (2013), etc. In this chapter, we focus on the meta-analysis of studies reporting estimates of PPV and NPV where data are reported in a form equivalent to a 2 by 2 table, with cut-offs unspecified and assumed varying from study to study.

1.1.1 Meta-analysis of diagnostic accuracy

The dependence of test performance metrics on the threshold for declaring a positive test result has led to the development of the Receiver Operating Characteristic (ROC) curve, and has also been incorporated in methods for the meta-analysis of studies reporting estimates of test sensitivity and specificity. The construction of a summary ROC (SROC) was proposed in Moses et al. (1993). This led to further developments in models like the proper summary ROC curve based on maximum likelihood estimation (Lloyd, 2000), the hierarchical summary ROC (HSROC) (Rutter and Gatsonis,

2001) that accounted for different thresholds used in different studies, and the bivariate random-effects model (Reitsma et al., 2005, Chu and Cole, 2006) to provide a "summary point" when performance estimates do not vary widely. The HSROC and bivariate random-effects models were shown to be mathematically equivalent in the absence of covariates (Harbord et al., 2007). A unifying framework for these two approaches and the choice of summary ROC curves were discussed in Arends et al. (2008). The HSROC and bivariate random-effects models are the preferred approach to meta-analyze diagnostic test accuracies (Macaskill et al., 2010, Trikalinos et al., 2012). Further developments expanded into cases where different studies used unequal number of ordered categories (Dukic and Gatsonis, 2003), and in cases where studies compared multiple index tests on the same participants in paired designs (Trikalinos et al., 2014).

Although sensitivity and specificity are theoretically independent of disease prevalence, these values have been shown to vary with empirical disease prevalence (Leeflang et al., 2013). Empirical disease prevalence can be viewed as a coarse marker for distinguishing study population characteristics with regards to disease spectrum. Differences in disease spectrum may lead to variations in observed test accuracy. A detailed discussion of possible relationships between empirical disease prevalence and test accuracy can be found in Leeflang et al. (2009). To account for variation in disease spectrum between studies, Chu et al. (2009) proposed an extension of the bivariate random-effects model to a tri-variate random-effects model to jointly model diagnostic accuracies and disease prevalence for studies that are prospectively designed. This method assumes a correlation structure between logit transformed sensitivity, specificity and prevalence using a multivariate normal distribution.

1.1.2 Meta-analysis of predictive values

An indirect approach to estimating summary predictive values was suggested by some authors (Macaskill et al., 2010, Trikalinos et al., 2012). In this approach, a meta-analysis of sensitivity and specificity is conducted first. This is followed by a transformation of the summary sensitivity and specificity to summary predictive values using a range of plausible prevalence values. Implicit in the indirect approach is the assumption that sensitivity and specificity are theoretically independent of disease prevalence.

Methods to directly provide a summary point for the predictive values have been proposed by Chu et al. (2009) using the alternative parameterization in the tri-variate random-effects model, and by Leeflang et al. (2012) where the predictive values are used in place of the diagnostic measures in the bivariate model. As with sensitivity and specificity, there exists a trade-off between PPV and NPV due to their dependence on the threshold for test positivity. In the literature, there are many instances where the thresholds used in the studies are not the same, or not even reported. In addition, measurements of biomarker can be heterogeneous due to the current state of biomarker assay processes (de Gramont et al., 2015). Under such circumstances, the reporting of a summary point may not be appropriate especially when meta-analyzing predictive values of a biomarker.

Some limitations of the indirect approach are best illustrated in the predictive ROC (PROC) space. The PROC curve consists of all possible pairs of PPV and (1-NPV) as the threshold for test positivity traverses its full range. The PROC curve was developed as a tool for characterizing the predictive performance of a test (Shiu and Gatsonis, 2008). The summary PROC curve derived from using the indirect approach may lead to a non-monotonic trade-off between PPV and NPV for a fixed disease prevalence. A biomarker should ordinarily have measurements that are mono-

tone with respect to the disease status. The lack of monotonicity in the PROC curve implies that there are situations whereby a test does not have a unique NPV for a given PPV, or vice versa. There will also be values of the threshold where trade-offs between PPV and NPV do not occur even when the corresponding trade-offs between sensitivity and specificity are occurring. Such behaviors are unrealistic and contradictory to the requirement of a monotone relationship between biomarker measurement and disease status.

In this chapter, we propose a Hierarchical Summary Predictive ROC (HSPROC) model. The HSPROC model overcomes the limitations highlighted in the preceding paragraphs, and produces a monotone summary predictive ROC curve for a binary test given the observed data from multiple studies. In the next section, we will introduce the basic assumptions, notation and definitions used throughout the chapter. Section 1.3 will recapitulate the key characteristics of the PROC curve, and the conditions required to achieve monotonicity. This will be followed by details of the model in Section 1.4. Application of the proposed method to an example on the meta-analysis of prognostic capabilities of biomarkers for rapid rule-out of acute myocardial infarction is presented in Section 1.5. More recently, Hattori and Zhou (2016) proposed a method to estimate predictive curves that are separate plots of PPV vs. threshold, and NPV vs. threshold. In Section 1.6, we will apply our proposed method to the same example used in that paper to highlight the differences between the two methods.

1.2 Assumptions, notation and definitions

We first describe the assumptions, mathematical notation and definitions related to diagnostic and predictive accuracies used in this chapter. Let D be a binary variable

denoting actual disease status, we define $D = 1$ as diseased and $D = 0$ as not diseased. Similarly for test outcome, T, we define $T = 1$ as a positive test and $T = 0$ as a negative test. We assume that Y is a continuous measurement of the biomarker, and without loss of generality, larger values of Y are more indicative of disease. For a given cutoff or decision threshold, c, $T = 1$ when $Y > c$. The true positive rate (TPR) and false positive rate (FPR) are defined as $TPR(c) = P(Y > c|D = 1) = \pi_1(c)$ and $FPR(c) = P(Y > c|D = 0) = \pi_0(c)$, respectively. The sensitivity of the test is thus equivalent to the TPR, and specificity of the test is equivalent to 1-FPR.

With disease prevalence of the population denoted as $p = P(D = 1)$,

$$\begin{aligned} PPV(c) = P(D = 1|T = 1) &= \frac{\pi_1(c)p}{\pi_1(c)p + \pi_0(c)(1-p)} \\ NPV(c) = P(D = 0|T = 0) &= \frac{(1-\pi_0(c))\,(1-p)}{(1-\pi_0(c))\,(1-p) + (1-\pi_1(c))\,p} \end{aligned}$$

For simplicity in notation, the dependence of TPR, FPR, π_1, π_0, PPV and NPV on c will be dropped in the rest of the chapter. The ROC curve is defined as $ROC(\cdot) = \{(FPR, TPR)_{c\in R}\}$ for a cutoff c belonging to the set R of all possible threshold values.

1.3 Predictive ROC (PROC) curve

The PROC curve was described in detail by Shiu and Gatsonis (2008), and only the key concepts will be summarized in this section.

The PROC curve is defined as $\{(1 - NPV, PPV)\}_{c\in R}$ for a cutoff c belonging to the set R of all possible threshold values at a fixed prevalence. Examples of PROC curves are shown in Figure 1.1 for the same test at different disease prevalences. From the plot, we observe that the PROC curve is implicitly a function of disease prevalence. The extreme points of the PROC curve correspond to threshold values

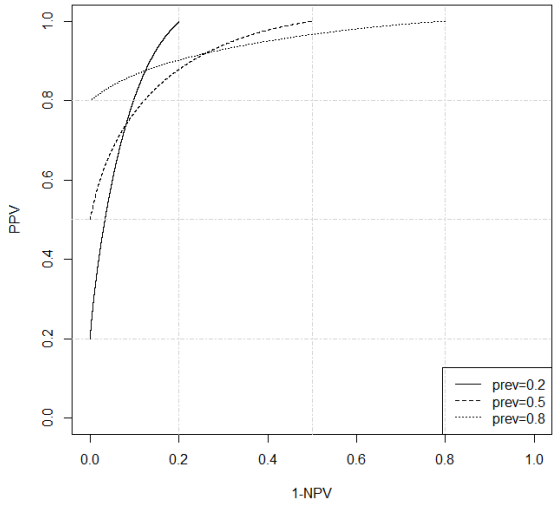


Figure 1.1: PROC curves generated using binormal model, $ROC(t) = \Phi\{a+b\Phi^{-1}(t)\}$. Here $a = 2$, $b = 1$, and t is the False Positive Rate (1-Specificity). Different values of disease prevalence used are shown in the legend.

resulting in all cases being classified as either positive or negative, and thus depend on the prevalence. This is in contrast to the ROC curve, which extends from the point (0,0) to (1,1) regardless of disease prevalence. In the PROC curve, the value of the threshold increases from the extreme where PPV equals to prevalence, to the other extreme of the curve where 1-NPV equals to the prevalence.

At a fixed prevalence rate, the variation in the PROC curves corresponding to different diagnostic performances as measured by the ROC curve is shown in the top row of Figure 1.2. A test with good performance measures will have PROC and ROC curves closer to the upper left corners of the respective plots. When TPR is equal to FPR for TPR and FPR $\in (0, 1)$, both PPV and 1-NPV are equal to the prevalence. When the PROC curve passes through this point, the corresponding ROC curve passes the 45° diagonal "guessing" line in the ROC plot, resulting in an improper ROC curve (Egan, 1975) as shown in the bottom row of Figure 1.2. The trajectories of a monotone versus a non-monotone PROC curve can be very different, and the lack of monotonicity in the PROC curve implies that there are situations wherein a test

does not have a unique NPV for a given PPV, and vice versa. A non-monotone PROC curve also implies that for some range of threshold values, a trade-off between PPV and NPV does not exist. It is hard to imagine a situation where this would occur in reality under the assumption that larger values of Y are more indicative of disease. Monotonicity of predictive accuracies with respect to the threshold is a common assumption in models used for predictive accuracy (Moskowitz and Pepe, 2004, Huang et al., 2007).

A necessary and sufficient condition for PPV to be monotone with respect to the threshold is that the conditional random variables $T|D = 1$ and $T|D = 0$ are hazard rate ordered, i.e $T|D = 0 \leq_{hr} T|D = 1$. Similarly for NPV, the necessary and sufficient condition is that the conditional random variables are reversed hazard rate ordered, i.e $T|D = 0 \leq_{rh} T|D = 1$. These conditions are simultaneously satisfied when the conditional random variables are likelihood ratio ordered, i.e $T|D = 0 \leq_{lr} T|D = 1$. It was shown in Egan (1975) that a concave ROC curve satisfies the likelihood ratio order. Thus an effective way to ensure that the PROC curve is monotone, while avoiding the complexity of modeling monotone PROC curves at different prevalence values, is to have a concave ROC curve. This will be the approach adopted in this chapter. Additional details on stochastic orders and the monotonicity conditions are included in Appendix 1.A.

In the ROC context, the use of models that may give rise to improper ROC curves, e.g. binormal models, has been investigated and discussed extensively. In support for such models, it has been argued that these anomalies usually occur over a small part of the ROC curve, and in some cases, occurs at the high end of the false positive range that is of little interest for practical purposes (Pepe, 2004). More recently however, Pesce et al. (2010) used decision analysis to argue for the use of proper ROC models, and Huang et al. (2013) also suggested the use of proper ROC models

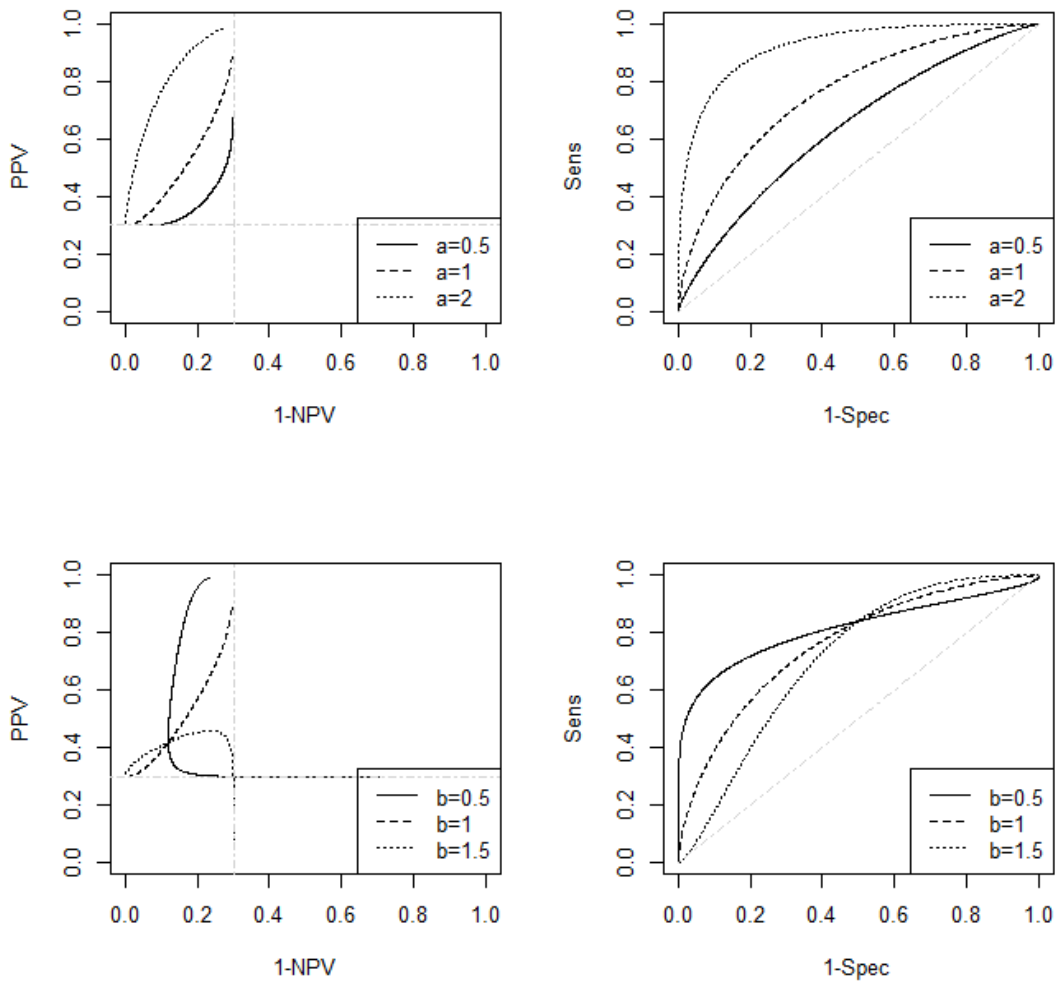


Figure 1.2: PROC curves and corresponding ROC curves for binormal ROC models. For the plots in the top row, the binormal model has parameters $b = 1$, $p = 0.3$, and values of a are as shown in the legend. For the plots in the bottom row, the binormal model has parameters $a = 1$, $p = 0.3$, and values of b are as shown in the legend. Under the binormal ROC model, the ROC curve is proper only when $b = 1$.

for biomarkers. For prognostic and predictive biomarkers, the key objective of the model is to capture the proper behavior of the predictive values with changes in threshold. We have shown that such anomalies, though small in the ROC space, are amplified in the PROC space, and will lead to PROC curves that contradict the underlying relationship between test measurement and disease condition. Thus with the insight obtained from the PROC curve, models that will ensure monotone PROC are more suitable for biomarkers.

1.4 HSPROC model

We assume the setting for the meta-analysis to involve K prospective studies with each study providing a 2 by 2 contingency table of summary data. For studies $k = 1 \dots K$, the notation used to represent the data in the k^{th} study are $(n_{k11}, n_{k10}, n_{k00}, n_{k01})$, representing the number of true positive, false positive, true negative and false negative cases respectively. Similarly, $(n_{k1.}, n_{k0.})$ are the number of cases with positive and negative test respectively. Empirical PPV and NPV for study k are thus determined by $\frac{n_{k11}}{n_{k1.}}$ and $\frac{n_{k00}}{n_{k0.}}$, respectively.

1.4.1 Bayesian hierarchical meta-regression model

Level I: Within-study variation

We assume a multinomial distribution for the counts in each study. Specifically we assume, for studies $k = 1 \dots K$, the within-study variation is defined as

$$[n_{k11}, n_{k10}, n_{k00}, n_{k01}] \sim \text{Mult}(N_k; q_{k1}, q_{k2}, q_{k3}, q_{k4})$$

where $N_k = \sum_i \sum_j n_{kij}$. The cell probabilities are expressed in terms of TPR, FPR and prevalence by $q_{k1} = \pi_{1_k} p_k,\ q_{k2} = \pi_{0_k}(1-p_k),\ q_{k3} = (1-\pi_{0_k})\,(1-p_k),$ and $q_{k4} = 1 - \sum_{j=1}^{3} q_{kj}$. We then define the model as

$$\text{logit}(\pi_{1_k}^{\beta}) = \theta_k + \alpha_k$$
$$\text{logit}(\pi_{0_k}^{\beta}) = \theta_k$$

The form of the above model is derived from assuming a Lomax distribution with a shape parameter of $1/\beta$ and scale parameter of e^{α_k} for the test measurement in the diseased population, and a Lomax distribution with the same shape parameter, but with a scale of 1 for the test measurement in the non-diseased population. The use of Lomax distributions was described in Campbell and Ratnaparkhi (1993) and Lloyd (2000), and important details for our application here are included in the Appendix 1.B.

In the above notation, θ_k denotes the positivity criteria as it affects both π_1 and π_0; α_k denotes the accuracy parameter as it results in a translation of the PROC curve, and also represents the difference between the true positive rate and false positive rate; and β influences the slope of the PROC curve. Under the hierarchical model used, the positivity criteria, accuracy parameter, and disease prevalence are allowed to vary across studies, while β is assumed to be constant across studies in the studied population.

Level II: Between-study variation

For the parameters that are allowed to vary across studies, we assume

$$\alpha_k|\mu_\alpha, \sigma_\alpha^2 \sim N(\mu_\alpha, \sigma_\alpha^2);\ \theta_k|\Theta, \sigma_\theta^2 \sim N(\Theta, \sigma_\theta^2);\ \text{and logit}(p_k)|\Pi, \sigma_\Pi^2 \sim N(\Pi, \sigma_\Pi^2)$$

Additional study level covariates that are deemed to affect the mean of α, θ and p can potentially be included into the model, but it should be noted that the number of covariates that can be effectively incorporated is limited by the remaining degrees of freedom available after specifying the basic hierarchical model. These study level covariates can be applied to explore the sources of variability related to differences in study design and execution, or to differences in patient groups and testing. As an example, the study level covariate Z is included in the Level II model via $\alpha_k|\mu_\alpha, \sigma_\alpha^2, \gamma_\alpha, Z_k \sim N(\mu_\alpha + \gamma_\alpha Z_k, \sigma_\alpha^2)$, with the corresponding prior for γ_α specified in Level III.

The concavity of the resulting ROC curve can be enforced by specifying likelihood and prior distributions for α that will result in positive support. Here we have chosen instead to let data determine the posterior distribution of α. In the event that the posterior distribution of α is in the negative region, resulting in a convex curve, a monotone transformation of the test result can be found to give a concave curve.

Level III

The remaining hyperparameters are assumed to be mutually independent. Values for the parameters in the prior distributions are chosen to better reflect plausible values while maintaining a relatively diffuse distribution, except for β. As stated earlier, β directly affects the slope of the PROC curve. As it turns out, the effect of β in the ROC space is on the symmetry of the curve. When $\beta = 1$, the resulting ROC curve is symmetric about the negative diagonal line of the ROC plot. A value of $\beta > 1$ will result in a ROC curve closer to the upper boundary of the ROC plot, while $0 < \beta < 1$ will lead to a ROC curve that is closer to the left vertical axis of the ROC plot. This perspective allows us to set a prior for β as $\text{Exp}(\log 2)$, which will have a median at 1 while maintaining the full support for β.

The remaining prior distributions are $\mu_\alpha \sim N(0, 100^2)$, $\sigma_\alpha \sim U(0, 100)$, $\Theta \sim N(0, 100^2)$, $\sigma_\theta \sim U(0, 100)$, $\Pi \sim N(0, 100^2)$, and $\sigma_\Pi \sim U(0, 100)$.

1.4.2 HSPROC model fitting

The proposed model can be implemented using open-source statistical computation software. For the examples described in Sections 1.5 and 1.6, the model was implemented in `JAGS` 4.2.0 (Plummer, 2003) via `R` 3.2.3 (R Core Team, 2015).

Since our interest is in the summary predictive performances of the test, and not the predictive performance of the test in a new study, each set of simulated draws for μ_α, Θ, β and Π are used to estimate the posterior distribution of functions of these parameters like PPV, NPV, disease prevalence, sensitivity and specificity. Inference could then be performed based on these posterior distributions.

The summary PROC curve given the observed data is generated by varying π_0, and using the medians of μ_a, β, and Π to compute the predictive values. Denoting the medians of the respective posterior distribution of the parameters by tilde, and for $\pi_0 \in [0, 1]$, the summary PROC curve is computed using the following expressions:

$$\Theta(\pi_0) = \text{logit}(\pi_0^{\tilde{\beta}}); \qquad \pi_1(\pi_0) = \left\{\text{logit}^{-1}\left[\tilde{\mu}_\alpha + \text{logit}\left(\pi_0^{\tilde{\beta}}\right)\right]\right\}^{1/\tilde{\beta}};$$
$$PPV(\pi_0) = \frac{1}{1 + \frac{1-\text{logit}^{-1}\tilde{\Pi}}{\text{logit}^{-1}\tilde{\Pi}} \frac{\pi_0}{\pi_1(\pi_0)}}; \qquad NPV(\pi_0) = 1 - \frac{1}{1 + \frac{1-\text{logit}^{-1}\tilde{\Pi}}{\text{logit}^{-1}\tilde{\Pi}} \frac{1-\pi_0}{1-\pi_1(\pi_0)}}$$

The summary PROC curve can be limited to the range of empirical FPR to avoid extrapolating beyond the range of data. Different values of the prevalence from its posterior distribution can also be used above to examine the behavior of the summary PROC curve under different prevalence. Where relevant, a summary point can be defined as the median PPV and NPV values from the respective posterior distribution.

1.5 Example 1: Meta-analysis of prognostic capabilities of biomarkers for rapid rule-out of acute myocardial infarction (AMI)

Lipinski et al. (2014) performed a meta-analysis to determine the prognostic capabilities of different biomarkers for rapid rule-out of acute myocardial infarction (AMI). Studies that assessed patients who presented to the emergency department with non-traumatic chest pain, and measured Copeptin levels were included for consideration. Case-control studies were excluded. 14 studies were eventually included in the meta-analysis with a total of 9,244 patients. Data were presented in terms of the number of true-positive, false-positive, false-negative, and true-negative. Some of the studies also presented data using Troponin, High Sensitive (HS) Troponin, and combinations of Copeptin with Troponin or HS Troponin. In the analysis, the authors computed the empirical diagnostic and predictive accuracies for each study. Each of these accuracies was then meta-analyzed separately using random-effects methods.

For the purpose of illustrating the utility of the proposed method to compare predictive capabilities of different biomarkers, we will restrict our analysis to Copeptin (13 studies) and Troponin (11 studies) only. For Copeptin, a number of studies had presented data using multiple cutpoints. To minimize the effect of correlation in the data used for illustration, only one cutpoint will be selected from these studies for analysis, and the main consideration for the choice of cutpoint is to induce greater heterogeneity in the data. For example, Balmelli/APACE presented data using Copeptin cutpoint of 10pmol/L and 14pmol/L, but only the data using the former cutpoint will be used as there are already more studies using 14pmol/L as cutpoint. Appendix 1.D.1 contains the data used for the meta-analysis of each biomarker. Results will be compared with the indirect approach. The HSROC model (Rutter and

Gatsonis, 2001) will be used in the indirect approach computation.

1.5.1 HSPROC model computations

Eight different chains were used with diffuse starting values for the parameters Θ, μ_α, β and Π, corresponding to $\Theta^{(0)} \in \{-30, 10\}$, $\mu_\alpha^{(0)} \in \{-10, 30\}$, $\beta^{(0)} \in \{0.1, 10\}$, and $\Pi^{(0)} \in \{-20, 10\}$ respectively. JAGS's automatic initial value generation function was used for the remaining parameters. An adaptation period of 1000 iterations was used, and this was followed by a burn-in of 100,000 iterations. 100,000 samples per chain were obtained from the sampling process after applying a thinning interval of 100. All Level III parameters were monitored. Convergence was assessed by examining the traceplots and using Gelman-Rubin diagnostic with an upper limit of the confidence interval of the potential scale reduction factor (PSRF) set at 1.01 for each parameter. The upper confidence interval PSRF for the monitored parameters and the corresponding multivariate PSRF attained were all 1.0.

1.5.2 Results from meta-analysis

The posterior distribution of the summary predictive values and prevalence is described in Appendix 1.D.2, Tables 1.3 and 1.4 for Copeptin and Troponin respectively. The posterior distribution of the summary sensitivity and specificity from the HSPROC model is also described in the appendix for completeness. In this particular meta-analysis, data for both biomarkers are from a similar population of patients and we would expect the posterior distribution of the prevalence to be similar. The estimated posterior mean and median of prevalence for both biomarkers are all approximately 0.2. The width of the 95% posterior interval is a reflection of the uncertainty involved in the constituent studies, and would be expected to be different

between the two biomarker data in this case.

Figure 1.3 shows the HSPROC curves, computed at the median prevalence of 0.2, for Copeptin and Troponin. All summary curves were restricted to the range of empirical FPR to avoid extrapolation. Empirical estimates for each study are also included in the plots. The corresponding summary PROC curves from the indirect approach are superimposed onto the respective plots for comparison. The summary PROC curves from the indirect approach are observed to be non-monotone for both biomarkers even when limited to the observed range of empirical FPR. The summary points obtained from using the primary parameterization of the tri-variate random-effects model (Chu et al., 2009) are also shown in the plots. For both biomarkers, the covariance structures selected based on AIC are referred to as the partially reduced model in Chu et al. (2009). The HSPROC curves at 2.5%, 50% and 97.5% quantiles of the posterior distribution of prevalence are shown in Figure 1.4 to provide a sense of the variation in the HSPROC curves with prevalence. Recall that each empirical point has a different empirical prevalence associated with it. It is observed that at different prevalence values, the shape of the HSPROC curves are influenced by those studies that have empirical prevalence values similar to the prevalence that the HSPROC curve assumed.

Differences in disease spectrum are known to affect predictive and diagnostic performances. To investigate if disease spectrum may have induced heterogeneity in the test accuracies, the logit transformed empirical prevalence $\text{logit}(\hat{p}_k)$ of each study was also included as a covariate to the accuracy parameter at the between study level. Specifically $\alpha_k|\mu_\alpha, \sigma_\alpha^2, \gamma_\alpha, Z_k \sim N(\mu_\alpha + \gamma_\alpha Z_k, \sigma_\alpha^2)$, where $Z_k = \text{logit}(\hat{p}_k)$. The resulting posterior distributions are shown in Appendix 1.D.2, Tables 1.5 and 1.6 for Copeptin and Troponin respectively. The posterior probability interval of γ_α for Copeptin is almost symmetrical about 0, providing greater confidence that disease spectrum

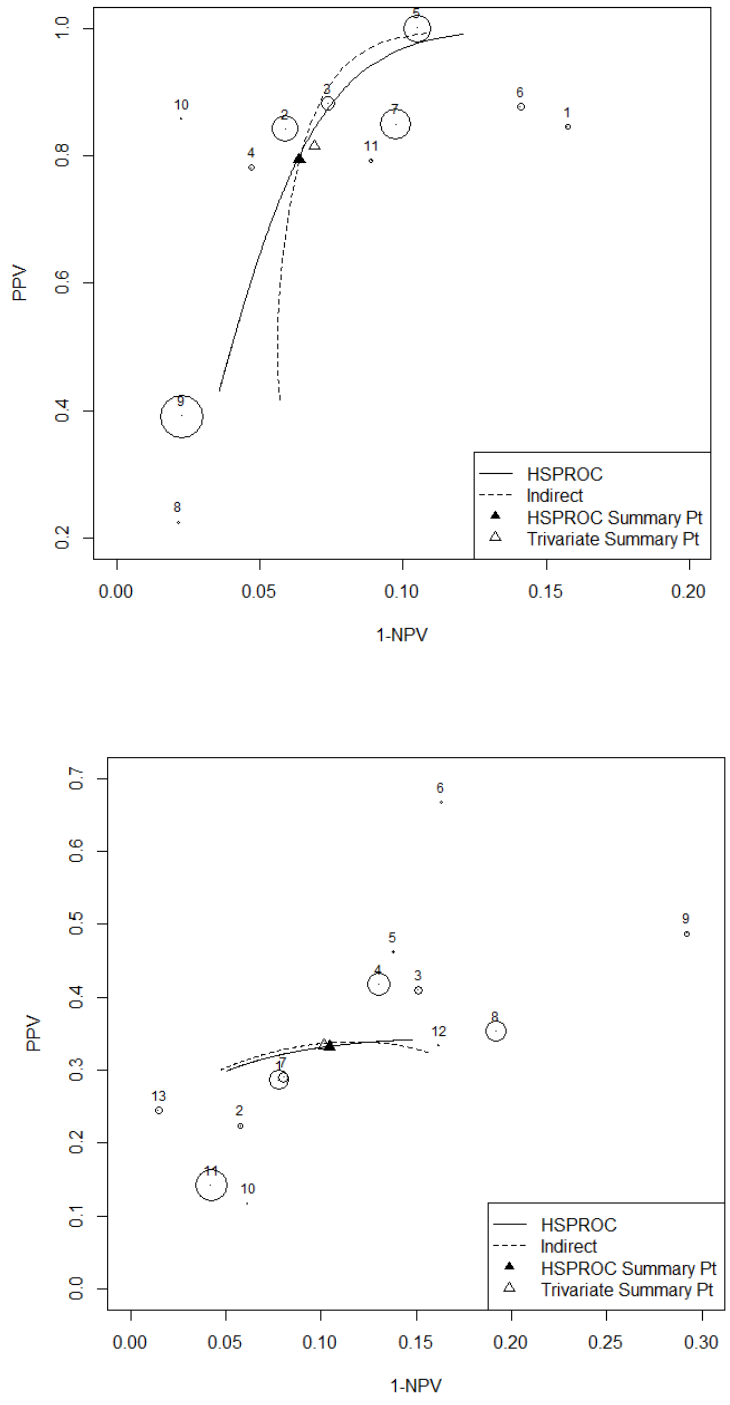


Figure 1.3: PROC curves for Troponin (top) and Copeptin (bottom) at a prevalence of 0.2. Empirical estimates are shown as dots and scaled by sample size.

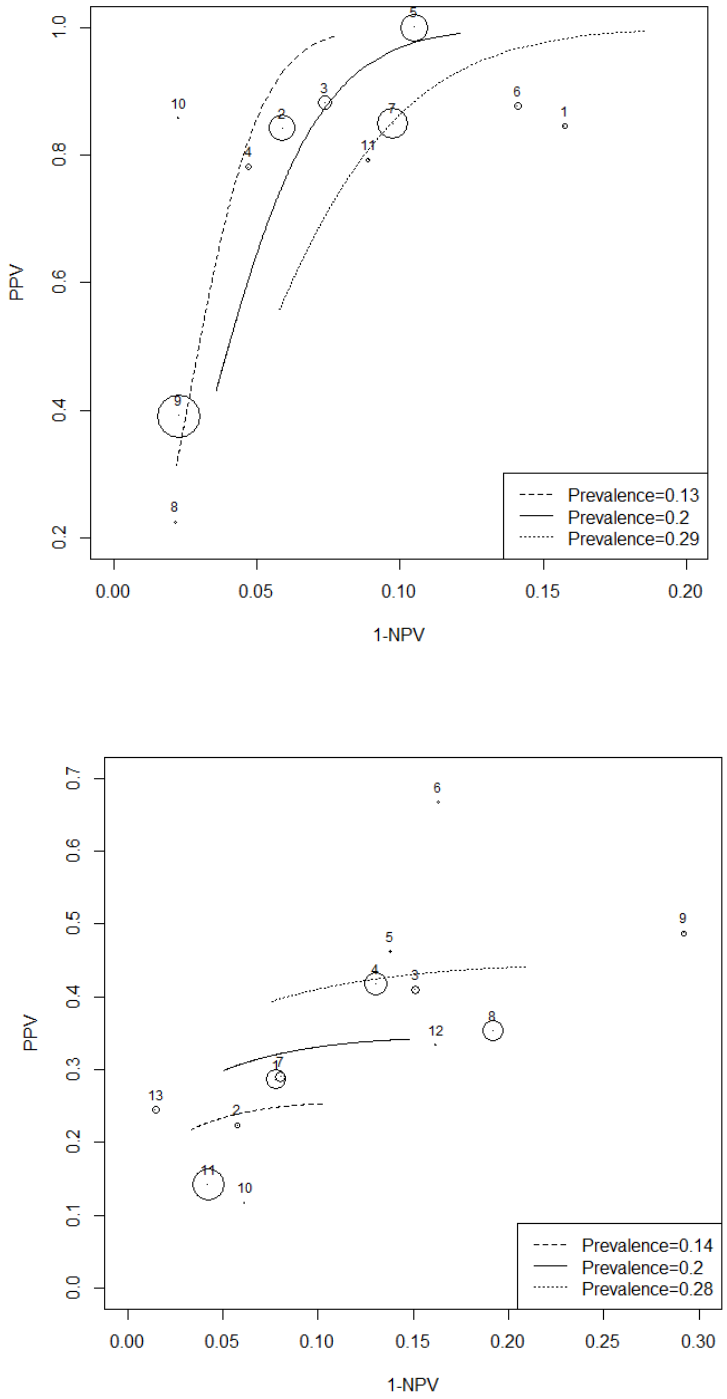


Figure 1.4: PROC curves at the 2.5%, 50%, and 97.5% quantiles of the posterior distribution of prevalence for Troponin (Top) and Copeptin (Bottom). Empirical estimates are shown as dots and scaled by sample size.

variation is not significant. For Troponin, approximatey 75% of the probability interval of γ_α is in the positive region, indicating that the effect of disease spectrum may not be ignored. To investigate this possibility, the studies for Troponin were divided into 2 subsets with similar empirical prevalences within each subset. Subset 1 contains studies with empirical prevalence less than the median value of the empirical prevalences (0.216), while the rest are in Subset 2. The resulting medians and 95% probability intervals of the posterior distributions of Subset 1's PPV and NPV are 0.631 $[0.246, 0.919]$, and 0.963 $[0.888, 0.998]$ respectively. Similarly for Subset 2, they are 0.899 $[0.628, 0.988]$, and 0.905 $[0.740, 0.992]$ respectively. The observation that stands out is the large 95% probability interval for Subset 1's PPV. The lack of precision in this estimate likely led to the posterior distribution for γ_α obtained for Troponin to have a large portion in the positive region. A closer scrutiny of the protocol for the studies in Subset 1 would be recommended to better understand the possible sources of heterogeneity.

When comparison of the predictive accuracies of different biomarkers is required, it is reasonable to assume that these different biomarkers are intended to be applied to the same population. The HSPROC curves of the biomarkers should therefore be compared at the prevalence rate of the target population. For a given prevalence rate, the summary PROC curve of a biomarker will completely dominate another if it resides to the upper left, where PPV and NPV are higher, and the curves do not intersect. If the summary PROC curves intersect, then the choice of biomarker may be made by limiting the comparison to the desired range of PPV or NPV. Using Copeptin and Troponin as an example, Figure 1.5 shows the summary PROC curves for the two modalities at various target prevalence values based on the full set of studies without adjusting for any between study covariates. In all cases, Troponin has better predictive values than Copeptin. It should be noted that we do not address the complexity arising from correlation induced by measurements of different biomarkers

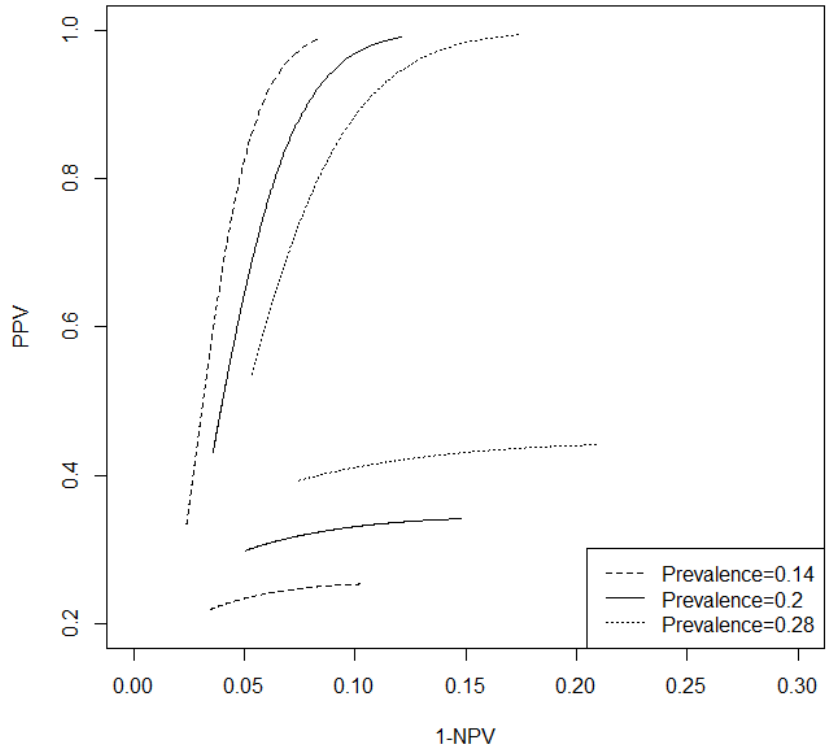


Figure 1.5: HSPROC curves for Copeptin and Troponin at different target values for prevalence. The set of HSPROC curves for Troponin are on the upper left of the plot.

taken from the same participants.

1.6 Example 2: Meta-analysis of prognostic capabilities of biomarkers for acute pulmonary embolism at high risk of short term death

In the second example, the data are from Becattini et al. (2007). The purpose of this example is to compare the results from the HSPROC model and the model proposed in Hattori and Zhou (2016). In the original study, the authors performed a meta-analysis on studies reporting the odds ratio on whether elevated serum troponin levels identify

patients with acute pulmonary embolism at high risk of short-term mortality. For the example here, the data consist of twenty studies of which 12 studies used Troponin I (total of 1303 patients), and the remaining 8 studies used Troponin T (total of 682 patients), Appendix 1.E.1.

Hattori and Zhou (2016) used the data from this study to illustrate their approach to estimate summary predictive curves, i.e. PPV vs. threshold, and NPV vs. threshold for Troponin I and T. They compared the prognostic capability of Troponin I vs. Troponin T using these estimated predictive curves under the assumption that thresholds used were known and allowed to vary across studies, similarity of disease prevalence in these studies, distribution of the test measurement is known and the same across all studies, and that fixed effect (equal effect) model was adequate. Asymptotic normal approximations were used to estimate confidence intervals for the summary predictive values. Both the summary positive and negative predictive curves, Figure 5 of Hattori and Zhou (2016), of Troponin T were found to be superior to the corresponding curves for Troponin I, and the authors concluded that Troponin T was superior to Troponin I.

We applied our model to this example using an approach similar to that described in Section 1.5. Detailed results are included in Appendix 1.E.2. A comparison of the two biomarkers at different target disease prevalence values is shown in Figure 1.6. From this plot, we observe that Troponin T has better prognostic capabilities than Troponin I across different disease prevalence values. The difference becomes larger at higher prevalence. The dependency of both PPV and NPV on positivity threshold can be easily discerned from the summary PROC curve at various target prevalences, and more importantly, monotone summary PROC curves are ensured. Investigation of sources of variability arising from between-study factors can also be easily performed in a similar manner to the earlier example in Section 1.5.

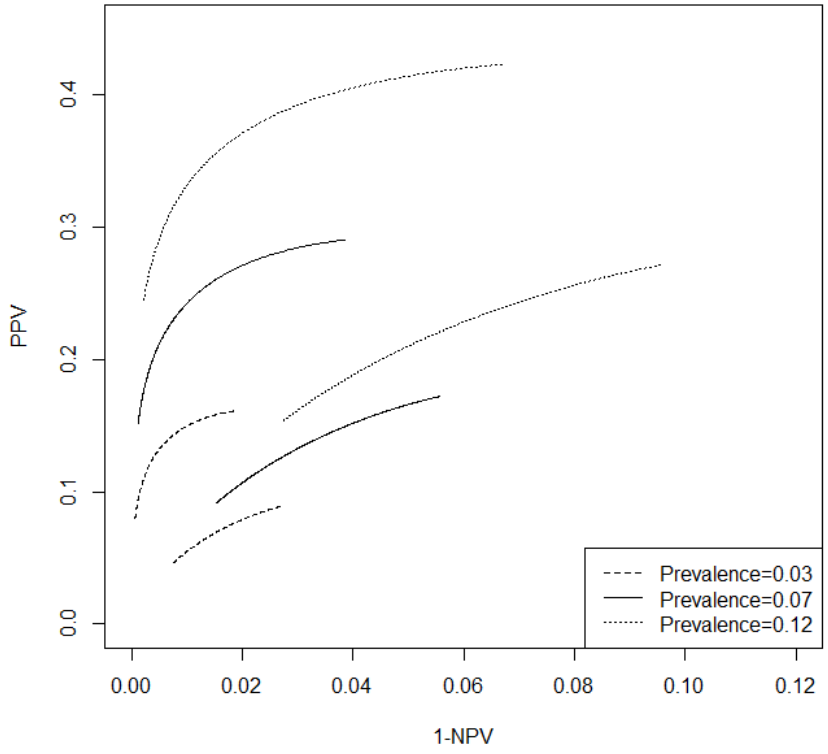


Figure 1.6: HSPROC curves for Troponin I and Troponin T at different target values for prevalence. The HSPROC curve for Troponin T is closer to the upper left of the plot at each value of prevalence.

For this set of data, the two different models arrive at the same conclusion that Troponin T has better prognostic capabilities than Troponin I. The results from HSPROC model, Figure 1.6, allow comparisons of both PPV and NPV jointly across different prevalence values between the 2 biomarkers. If the prevalence across studies in the data is truly similar as assumed in Hattori and Zhou (2016), the posterior distribution of the prevalence parameter in the HSPROC model will reflect that. The key difference between the models is that in the HSPROC model, the focus is on modeling the joint relationship between PPV, NPV, and prevalence. Modeling these key components jointly will ensure a well behaved predictive ROC curve, and show the interplay between PPV and NPV to indicate possible pairs of PPV and NPV that can be achieved. In addition, this model provides predictive performances at different values of prevalence.

1.7 Discussion

In this chapter, we have proposed a Bayesian hierarchical model to synthesize estimates of PPV, NPV, and disease prevalence jointly. This model accounts for effects of disease spectrum, and the dependence of both PPV and NPV on threshold. It assumes that positivity thresholds used across studies are different and unknown. Monotonicity of the summary PROC curve is ensured to yield physically meaningful behavior. Computation of summary PROC curves at different disease prevalence values to characterize the performance of the biomarker, and for comparing against other biomarkers under specific target disease prevalence can be easily performed.

We believe this model fills a gap in the existing methods for the meta-analysis of predictive accuracy for binary tests on binary outcomes. Existing approaches, also referred to as the indirect method in this chapter, have difficulties in producing well

behaved predictive values. We also noted that not all existing methods take advantage of additional information available from prospectively designed trials to account for prevalence, which serve as a proxy for disease spectrum, or account for differences in threshold used across studies. The joint dependency of the predictive values on threshold requires that predictive values be characterized as a pair. Separate assessments of each predictive value do not necessarily lead to a proper characterization of their joint dependence on the threshold. Thus, summary results for predictive values should be quantified and assessed jointly via a PROC curve, rather than separately through predictive curves. Examining each element of the predictive pair separately could easily let problems like the non-monotone behavior go undetected.

Our proposed model to meta-analyze predictive values of biomarkers assumes availability of data in the form of a 2 by 2 contingency table, which is often the case in practice. This model accounts for the prospective nature of the studies, and assumes a multinomial distribution for the observed cell counts in the contingency table. An underlying Lomax distribution for the distribution of the biomarker measurement, or a monotonously transformed version of the measurement, is assumed to ensure a monotone summary PROC curve. This is not necessarily a restriction on the general application of the model as the PROC curve is a function of both disease prevalence and the ROC curve. The latter pertains to the relationship between the survival functions of the diseased and non-diseased populations, and not the distributions themselves.

As noted earlier, the proposed model does not address the complexity arising from the use of correlated data. It is common for studies to measure multiple biomarkers on the same patient for greater trial efficiency. A similar situation is where results based on multiple cut-points are reported within the same study. Extension of the current model to allow for correlated data would be a relevant and useful topic for

future research. Other related topics would be the extension of the proposed model to a network meta-analysis setting, and settings where the reference standard is not perfect.

The use of a Bayesian hierarchical model provides flexibility and ease in obtaining estimates, and the corresponding posterior intervals. Computational demands of MCMC algorithm are less of a constraint with improving computational power. We believe the proposed model is therefore easily accessible to most meta-analysts.

1.A Monotonicity

1.A.1 Stochastic Orders

Let X and Y be two random variables, then under the usual stochastic order,

$$X \leq_{st} Y \iff P(X \geq x) \leq P(Y \geq x) \qquad \forall\, x \in (-\infty, \infty)$$

Let $r_X(t) = \dfrac{f_X(t)}{\bar{F}_X(t)}$, where $t \in \mathbb{R}$ and $\bar{F}_X(t) = 1 - F_X(t)$. Then X is said to be smaller than Y in the hazard rate order ($X \leq_{hr} Y$) when

$$r_X(t) \geq r_Y(t) \qquad t \in \mathbb{R}$$

Theorem 1.B.1 (Shaked and Shanthikumar, 2007) *If X and Y are two random variables such that $X \leq_{hr} Y$, then $X \leq_{st} Y$*

Let

$$\tilde{r}_X(t) = \frac{d}{dt} \log F_X(t) = \frac{f_X(t)}{F_X(t)}$$

Then X is said to be smaller than Y in the reversed hazard rate order ($X \leq_{rh} Y$) when

$$\tilde{r}_X(t) \leq \tilde{r}_Y(t) \qquad t \in \mathbb{R}$$

Theorem 1.B.42 (Shaked and Shanthikumar, 2007) *If X and Y are two random variables such that $X \leq_{rh} Y$, then $X \leq_{st} Y$*

Let X and Y be two random variables with densities f and g respectively. Then X is said to be smaller than Y in the likelihood ratio order ($X \leq_{lr} Y$) when

$$f(x)g(y) \geq f(y)g(x) \qquad \forall x \leq y$$

Theorem 1.C.1 (Shaked and Shanthikumar, 2007) *If X and Y are two random variables such that $X \leq_{lr} Y$, then $X \leq_{hr} Y$ and $X \leq_{rh} Y$, and therefore $X \leq_{st} Y$*

Remark 1.C.2 (Shaked and Shanthikumar, 2007) *Neither of the orders $\leq_{hr}$ and $\leq_{rh}$ (even if both hold simultaneously) implies the order $\leq_{lr}$*

1.A.2 Monotonicity of predictive values

Let Y denote the random variable $T|D=1$ with a distribution of G, and X denote the random variable $T|D=0$ with a distribution of F.

$$PPV(c) = \frac{[1-G(c)]p}{[1-G(c)]p + [1-F(c)](1-p)}$$
$$\frac{d}{dc}PPV(c) = -\frac{g(c)p}{[1-G(c)]p + [1-F(c)](1-p)} - \frac{[1-G(c)]p[-g(c)p - f(c)(1-p)]}{\{[1-G(c)]p + [1-F(c)](1-p)\}^2}$$

$$NPV(c) = \frac{F(c)(1-p)}{G(c)p + F(c)(1-p)}$$
$$\frac{d}{dc}NPV(c) = \frac{f(c)(1-p)}{G(c)p + F(c)(1-p)} - \frac{F(c)(1-p)[g(c)p + f(c)(1-p)]}{\{G(c)p + F(c)(1-p)\}^2}$$

Monotonicity of the predictive values with respect to the cut-off, c, can be determined by taking the first derivative of the predictive values with respect to the cut-off.

$$\frac{d}{dc}PPV(c) \geq 0$$

$$-g(c)\{[1-G(c)]p + [1-F(c)](1-p)\} + p[1-G(c)][g(c)p + f(c)(1-p)] \geq 0$$

$$g(c)[1-G(c)]p + g(c)[1-F(c)](1-p) \leq g(c)[1-G(c)]p + f(c)[1-G(c)](1-p)$$

$$g(c)[1-F(c)] \leq f(c)[1-G(c)]$$

$$\frac{g(c)}{1-G(c)} \leq \frac{f(c)}{1-F(c)}$$

$$\Longrightarrow X \leq_{hr} Y$$

$$\frac{d}{dc}NPV(c) \leq 0$$

$$f(c)(1-p)[G(c)p + F(c)(1-p)] \leq F(c)(1-p)[g(c)p + f(c)(1-p)]$$

$$f(c)G(c)p \leq F(c)g(c)p$$

$$\frac{f(c)}{F(c)} \leq \frac{g(c)}{G(c)}$$

$$\Longrightarrow X \leq_{rh} Y$$

From Theorem 1.C.1, we note that if $X \leq_{lr} Y$, then $X \leq_{hr} Y$ and $X \leq_{rh} Y$, thus satisfying the monotonicity conditions for PPV and NPV. The theorem further implies that $X \leq_{st} Y$, which is the fundamental assumption in diagnostic accuracy

1.B Proper ROC curve model

1.B.1 Lomax distribution

$$X \sim Lomax(\text{scale} = \lambda,\ \text{shape} = \kappa)$$

$$x \geq 0,\ \lambda > 0,\ \kappa > 0$$

$$\begin{aligned}
f_X(x) &= \frac{\kappa}{\lambda}\left(1 + \frac{x}{\lambda}\right)^{-(\kappa+1)} \\
F_X(x) &= 1 - \left(1 + \frac{x}{\lambda}\right)^{-\kappa} \\
E[X] &= \frac{\lambda}{\kappa - 1}, \qquad \kappa > 1 \\
\text{Var}(x) &= \frac{\lambda^2 \kappa}{(\kappa-1)^2(\kappa-2)}, \qquad \kappa > 2
\end{aligned}$$

1.B.2 ROC curve derivation

Let X be the result of a diagnostic test for the diseased population and that $X \sim Lomax(\lambda, \kappa)$. For the non-diseased population, let Y be the result of the diagnostic

test and $Y \sim Lomax(1, \kappa)$. Then

$$F_X(t) = 1 - \left(1 + \frac{t}{\lambda}\right)^{-\kappa}$$
$$F_Y(t) = 1 - (1+t)^{-\kappa}$$

$$\begin{aligned} TPR(t) = \pi_1(t) &= 1 - F_X(t) \\ &= \left(1 + \frac{t}{\lambda}\right)^{-\kappa} \\ FPR(t) = \pi_0(t) &= 1 - F_Y(t) \\ &= (1+t)^{-\kappa} \end{aligned}$$

$$\therefore ROC(\pi_0) = \left[1 + \frac{1}{\lambda}\left(\pi_0^{-1/\kappa} - 1\right)\right]^{-\kappa}$$

The derivation of the ROC curve model is as follows

$$\begin{aligned} \text{logit}\left(\pi_0^{1/\kappa}\right) &= \log \frac{\pi_0^{1/\kappa}}{1 - \pi_0^{1/\kappa}} \\ &= \log \pi_0^{1/\kappa} - \log\left[\pi_0^{1/\kappa}\left(\pi_0^{-1/\kappa} - 1\right)\right] \\ &= -\log\left(\pi_0^{-1/\kappa} - 1\right) \\ \text{logit}\left(\pi_1^{1/\kappa}\right) &= \log \frac{\pi_1^{1/\kappa}}{1 - \pi_1^{1/\kappa}} \\ &= \log\left[\frac{1}{1 + \frac{1}{\lambda}\left(\pi_0^{-1/\kappa} - 1\right)} \frac{1 + \frac{1}{\lambda}\left(\pi_0^{-1/\kappa} - 1\right)}{\frac{1}{\lambda}\left(\pi_0^{-1/\kappa} - 1\right)}\right] \\ &= \log \frac{\lambda}{\pi_0^{-1/\kappa} - 1} \\ &= \log \lambda + \text{logit}\pi_0^{1/\kappa} \end{aligned}$$

Let $\beta = \frac{1}{\kappa}$, and $\alpha = \log \lambda$, where $\beta > 0,\ -\infty < \alpha < \infty$, then

$$\text{logit}\left(\pi_0^{\beta}\right) = \theta$$
$$\text{logit}\left(\pi_1^{\beta}\right) = \theta + \alpha$$

and $ROC(\pi_0)$ can be expressed as $ROC(\pi_0) = \left[1 + e^{-\alpha}\left(\pi_0^{-\beta} - 1\right)\right]^{-1/\beta}$.

1.C JAGS Code

```
### JAGS Model for HSPROC
model{
     for (k in 1:K){
          ## Within Study Variation
          Y[k,] ~ dmulti(q[k,], N[k])
          q[k,1] <-  tpr[k] * p[k]
          q[k,2] <-  fpr[k] * (1-p[k])
          q[k,3] <-  (1-fpr[k]) * (1-p[k])
          q[k,4] <-  1 - q[k,1] - q[k,2] - q[k,3]
          tpr[k] <- g1[k]^(1/b)
          fpr[k] <- g0[k]^(1/b)

          logit(g1[k]) <- theta[k]+a[k]
          logit(g0[k]) <- theta[k]
          logit(p[k]) <- logitp[k]

          ## Between Study Variation
          a[k] ~ dnorm(mu_a, prec_a)
          theta[k] ~ dnorm(Theta, prec_theta)
          logitp[k] ~ dnorm(P, prec_P)
     }

     ## Priors
     mu_a ~ dnorm(0, 0.0001)
     prec_a <- pow(sigma_a, -2)
```

```
    sigma_a ~ dunif(0, 100)

    Theta ~ dunif(-10, 10)
    prec_theta <- pow(sigma_theta, -2)
    sigma_theta ~ dunif(0, 100)

    P ~ dnorm(0, 0.0001)
    prec_P <- pow(sigma_P, -2)
    sigma_P ~ dunif(0, 100)

    b ~ dexp(log(2))
}
```

1.D Data and Results for Example 1

1.D.1 Data

Table 1.1: Copeptin Data

Study	Cutpoint	TP	FP	FN	TN
Balmelli/APACE	10	131	327	58	685
Chenevier-Gobeaux	10.7	36	125	9	147
Giannitsis	10	95	137	41	230
Keller	13	172	240	127	847
Sebbane	13.1	36	42	16	100
Afzali	14	92	46	15	77
Charpentier	14	60	147	35	399
COPED	14	203	372	128	539
Eggers	14	57	60	71	172
Lotze	14	9	68	4	61
Maisel/CHOPIN	14	95	572	52	1183
Meune	14	7	14	6	31
Thelin	14	67	207	3	201

Table 1.2: Troponin Data

Study	Cutpoint	TP	FP	FN	TN
Troponin					
Afzali	Troponin I, 40 ng/L	87	16	20	107
Balmelli/APACE	Troponin T, 35 ng/L	127	24	61	973
Charpentier	Troponin I, 100 ng/L	52	7	43	539
Chenevier-Gobeaux	Troponin I, 140 or 60 ng/L	32	9	13	263
COPED	Troponin T, 30 ng/L	224	0	107	911
Eggers	Troponin I, 70 ng/L	92	13	36	219
Keller	Troponin T, 30 ng/L	185	33	114	1054
Lotze	Troponin T, 100 ng/L	11	38	2	91
Maisel/CHOPIN	Troponin I, 0.04 ng/mL	118	184	38	1627
Meune	Troponin I, 140 ng/L	12	2	1	43
Sebbane	Troponin I, 40 ng/L	38	10	13	133

1.D.2 Results

Table 1.3: Posterior distribution summaries of key parameter and diagnostic performance measures for Copeptin.

	Mean	2.5%	25%	50%	75%	97.5%
$\text{logit}^{-1}(\Pi)$	0.202	0.140	0.179	0.200	0.222	0.275
PPV	0.334	0.242	0.301	0.332	0.364	0.434
NPV	0.894	0.838	0.878	0.895	0.911	0.937
Sensitivity	0.692	0.575	0.654	0.693	0.730	0.804
Specificity	0.652	0.589	0.634	0.653	0.672	0.708

Table 1.4: Posterior distribution summaries of key parameter and diagnostic performance measures for Troponin.

	Mean	2.5%	25%	50%	75%	97.5%
$\text{logit}^{-1}(\Pi)$	0.203	0.131	0.176	0.200	0.226	0.291
PPV	0.786	0.607	0.738	0.793	0.842	0.919
NPV	0.934	0.888	0.923	0.936	0.947	0.965
Sensitivity	0.736	0.620	0.705	0.739	0.771	0.832
Specificity	0.948	0.888	0.936	0.952	0.965	0.984

Table 1.5: Posterior distribution summaries of parameters for Copeptin with $\text{logit}(\hat{p}_k)$ as between study covariate.

	Mean	SD	2.5%	25%	50%	75%	97.5%
Π	-1.389	0.211	-1.811	-1.522	-1.388	-1.255	-0.971
Θ	-6.242	3.155	-13.949	-7.906	-5.746	-4.021	-1.504
β	5.902	2.885	1.647	3.839	5.440	7.425	12.999
γ_α	0.185	0.712	-1.130	-0.220	0.135	0.537	1.777
μ_α	4.430	2.216	1.416	2.866	3.992	5.508	10.005
σ_Π	0.724	0.180	0.464	0.599	0.694	0.815	1.161
σ_α	1.218	0.640	0.441	0.777	1.068	1.488	2.874
σ_θ	1.700	0.911	0.527	1.058	1.506	2.118	4.014

Table 1.6: Posterior distribution summaries of parameters for Troponin with $\text{logit}(\hat{p}_k)$ as between study covariate.

	Mean	SD	2.5%	25%	50%	75%	97.5%
Π	-1.388	0.251	-1.890	-1.544	-1.388	-1.233	-0.889
Θ	1.114	1.532	-1.698	0.161	1.047	1.995	4.300
β	0.151	0.173	0.005	0.043	0.099	0.201	0.582
γ_α	0.157	0.261	-0.355	0.007	0.154	0.303	0.682
μ_α	2.767	0.551	1.952	2.448	2.696	2.990	4.003
σ_Π	0.786	0.221	0.480	0.633	0.745	0.892	1.330
σ_α	0.435	0.247	0.158	0.282	0.380	0.518	1.036
σ_θ	0.629	0.254	0.340	0.471	0.574	0.718	1.245

1.E Data and Results for Example 2

1.E.1 Data

Table 1.7: Troponin I Data

	Study	TP	FP	FN	TN
1	Meyer	0	14	0	22
2	Douketis	0	5	0	19
3	Kucher	4	24	1	62
4	Mehta	1	17	1	19
5	Enea	4	16	0	6
6	La Vecchia	5	9	1	33
7	Douketis	6	56	10	386
8	Binder	6	40	1	77
9	Yalamanchili	8	16	9	114
10	Scridon	23	50	5	63
11	Amorim	2	40	1	17
12	Hsu	12	50	8	40

Table 1.8: Troponin T Data

	Study	TP	FP	FN	TN
9	Giannitsis	8	10	1	37
10	Janata	5	36	0	65
11	Pruszczyk	8	24	0	32
12	Bova	7	19	1	33
13	Kostrubiec	9	30	6	55
14	Kline	2	18	0	161
15	Kaczynska	10	18	6	53
16	Tulevski	2	4	0	22

1.E.2 Results

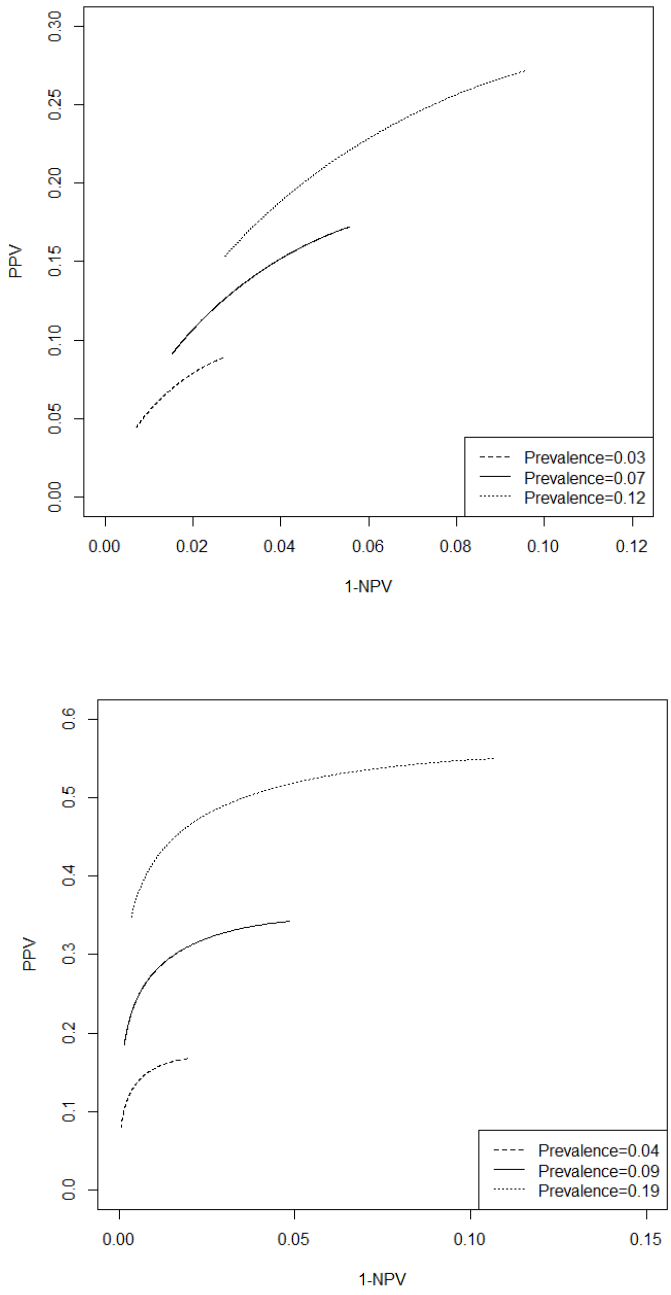


Figure 1.7: PROC curves at the 2.5%, 50%, and 97.5% quantiles of the posterior distribution of prevalence for Troponin I (Top) and Troponin T (Bottom). Empirical estimates are shown as dots and scaled by sample size.

Table 1.9: Posterior distribution summaries of key parameter and diagnostic performance measures for Troponin I.

	Mean	2.5%	25%	50%	75%	97.5%
$\text{logit}^{-1}(\Pi)$	0.073	0.035	0.059	0.072	0.086	0.122
PPV	0.137	0.063	0.108	0.133	0.162	0.232
NPV	0.965	0.928	0.957	0.968	0.977	0.991
Sensitivity	0.702	0.443	0.630	0.712	0.784	0.910
Specificity	0.648	0.491	0.605	0.653	0.696	0.776

Table 1.10: Posterior distribution summaries of key parameter and diagnostic performance measures for Troponin T.

	Mean	2.5%	25%	50%	75%	97.5%
$\text{logit}^{-1}(\Pi)$	0.095	0.037	0.070	0.089	0.112	0.187
PPV	0.256	0.100	0.194	0.248	0.308	0.464
NPV	0.983	0.928	0.976	0.991	0.998	1.000
Sensitivity	0.878	0.515	0.813	0.925	0.987	1.000
Specificity	0.736	0.606	0.705	0.741	0.773	0.838

Bibliography

D. G. Altman, L. M. McShane, W. Sauerbrei, and S. E. Taube. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. *PLoS medicine*, 9(5):e1001216, jan 2012. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001216.

M.-G. Anitei, G. Zeitoun, B. Mlecnik, F. Marliot, N. Haicheur, A.-M. Todosi, A. Kirilovsky, C. Lagorce, G. Bindea, D. Ferariu, M. Danciu, P. Bruneval, V. Scripcariu, J.-M. Chevallier, F. Zinzindohoué, A. Berger, J. Galon, and F. Pagès. Prognostic and Predictive Values of the Immunoscore in Patients with Rectal Cancer. *Clinical Cancer Research*, 20(7):1891–1899, apr 2014. doi: 10.1158/1078-0432.CCR-13-2830.

L. R. Arends, T. H. Hamza, J. C. van Houwelingen, M. H. Heijenbrok-Kal, M. G. M. Hunink, and T. Stijnen. Bivariate random effects meta-analysis of ROC curves. *Medical decision making : an international journal of the Society for Medical Decision Making*, 28(5):621–38, 2008. ISSN 0272-989X. doi: 10.1177/0272989X08319957.

C. Becattini, M. C. Vedovati, and G. Agnelli. Prognostic Value of Troponins in Acute Pulmonary Embolism. *Circulation*, 116(4):427 LP – 433, jul 2007.

I. Biliavska, T. A. Stamm, J. Martinez-Avila, T. W. J. Huizinga, R. B. M. Landewé, G. Steiner, D. Aletaha, J. S. Smolen, and K. P. Machold. Application of

the 2010 ACR/EULAR classification criteria in patients with very early inflammatory arthritis: analysis of sensitivity, specificity and predictive values in the SAVE study cohort. *Annals of the Rheumatic Diseases*, 72(8):1335–1341, aug 2013. doi: 10.1136/annrheumdis-2012-201909.

P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. Irwig, J. G. Lijmer, D. Moher, D. Rennie, H. C. W. de Vet, H. Y. Kressel, N. Rifai, R. M. Golub, D. G. Altman, L. Hooft, D. A. Korevaar, and J. F. Cohen. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*, 351, oct 2015.

G. Campbell and M. V. Ratnaparkhi. An application of Lomax distributions in receiver operating characteristic (ROC) curve analysis. *Communications in Statistics*, 22(6):1681–1697, 1993.

H. Chu and S. R. Cole. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology*, 59(12):1331–1332, 2006. doi: 10.1016/j.jclinepi.2006.06.011.

H. Chu, L. Nie, S. R. Cole, and C. Poole. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence : Alternative parameterizations and model selection. *Statistics in medicine*, 28(18):2384–2399, 2009. doi: 10.1002/sim.

C. H. Chung, Q. Zhang, C. S. Kong, J. Harris, E. J. Fertig, P. M. Harari, D. Wang, K. P. Redmond, G. Shenouda, A. Trotti, D. Raben, M. L. Gillison, R. C. Jordan, and Q.-T. Le. p16 Protein Expression and Human Papillomavirus Status As Prognostic Biomarkers of Nonoropharyngeal Head and Neck Squamous Cell Carcinoma. *Journal of Clinical Oncology*, 32(35):3930–3938, sep 2014. ISSN 0732-183X. doi: 10.1200/JCO.2013.54.5228.

A. de Gramont, S. Watson, L. M. Ellis, J. Rodon, J. Tabernero, A. de Gramont, and

S. R. Hamilton. Pragmatic issues in biomarker evaluation for targeted therapies in cancer. *Nat Rev Clin Oncol*, 12(4):197–212, apr 2015. ISSN 1759-4774. doi: 10.1038/nrclinonc.2014.202.

V. Dukic and C. Gatsonis. Meta-analysis of Diagnostic Test Accuracy Assessment Studies with Varying Number of Thresholds. *Biometrics*, 59(4):936–946, 2003. ISSN 0006341X. doi: 10.1111/j.0006-341X.2003.00108.x.

F. Efficace, G. Gaidano, M. Breccia, M. T. Voso, F. Cottone, E. Angelucci, G. Caocci, R. Stauder, D. Selleslag, M. Sprangers, U. Platzbecker, A. Ricco, G. Sanpaolo, O. Beyne-Rauzy, F. Buccisano, G. A. Palumbo, D. Bowen, K. Nguyen, P. Niscola, M. Vignetti, and F. Mandelli. Prognostic value of self-reported fatigue on overall survival in patients with myelodysplastic syndromes: a multicentre, prospective, observational, cohort study. *The Lancet Oncology*, 16(15):1506–1514, nov 2015. doi: 10.1016/S1470-2045(15)00206-5.

J. P. Egan. *Signal detection theory and ROC analysis*. New York: Academic Press, 1975.

R. M. Harbord, J. J. Deeks, M. Egger, P. Whiting, and J. A. C. Sterne. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics (Oxford, England)*, 8(2):239–51, 2007. ISSN 1465-4644. doi: 10.1093/biostatistics/kxl004.

S. Hattori and X.-H. Zhou. Evaluation of predictive capacities of biomarkers based on research synthesis. *Statistics in Medicine*, 35(25):4559–4572, nov 2016. ISSN 1097-0258. doi: 10.1002/sim.7018.

Y. Huang, M. S. Pepe, and Z. Feng. Evaluating the predictiveness of a continuous marker. *Biometrics*, 63(4):1181–1188, 2007. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2007.00814.x.

Y. Huang, M. S. Pepe, and Z. Feng. Logistic regression analysis with standardized

markers. *The Annals of Applied Statistics*, 7(3):1640–1662, 2013. ISSN 1932-6157. doi: 10.1214/13-AOAS634.

M. M. G. Leeflang, P. M. M. Bossuyt, and L. Irwig. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *Journal of Clinical Epidemiology*, 62(1):5–12, 2009. ISSN 08954356. doi: 10.1016/j.jclinepi.2008.04.007.

M. M. G. Leeflang, J. J. Deeks, A. W. S. Rutjes, J. B. Reitsma, and P. M. M. Bossuyt. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. *Journal of clinical epidemiology*, 65(10):1088–97, 2012. ISSN 1878-5921. doi: 10.1016/j.jclinepi.2012.03.006.

M. M. G. Leeflang, A. W. S. Rutjes, J. B. Reitsma, L. Hooft, and P. M. M. Bossuyt. Variation of a test's sensitivity and specificity with disease prevalence. *Canadian Medical Association Journal*, 185(11):E537–E544, 2013. ISSN 08203946. doi: 10.1503/cmaj.121286.

S. Liao, B. C. Penney, H. Zhang, K. Suzuki, and Y. Pu. Prognostic Value of the Quantitative Metabolic Volumetric Measurement on 18F-FDG PET/CT in Stage IV Nonsurgical Small-cell Lung Cancer. *Academic Radiology*, 19(1):69–77, 2012. ISSN 10766332. doi: 10.1016/j.acra.2011.08.020.

M. J. Lipinski, R. O. Escárcega, F. D'Ascenzo, M. A. Magalhães, N. C. Baker, R. Torguson, F. Chen, S. E. Epstein, Ò. Miró, P. Llorens, E. Giannitsis, U. Lotze, S. Lefebvre, M. Sebbane, J. P. Cristol, C. Chenevier-Gobeaux, C. Meune, K. M. Eggers, S. Charpentier, R. Twerenbold, C. Mueller, G. Biondi-Zoccai, and R. Waksman. A systematic review and collaborative meta-analysis to determine the incremental value of copeptin for rapid rule-out of acute myocardial infarction. *American Journal of Cardiology*, 113(9):1581–1591, 2014. ISSN 18791913. doi: 10.1016/j.amjcard.2014.01.436.

C. J. Lloyd. Regression models for convex ROC curves. *Biometrics*, 56(3):862–867, 2000. ISSN 0006-341X.

P. Macaskill, C. Gatsonis, J. Deeks, R. Harbord, and Y. Takwoingi. Analysing and Presenting Results. In J. Deeks, P. Bossuyt, and C. Gatsonis, editors, *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*, chapter 10. The Cochrane Collaboration, version 1 edition, 2010. URL `http://srdta.cochrane.org`.

L. E. Moses, D. Shapiro, and B. Littenberg. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in medicine*, 12(14):1293–1316, 1993. ISSN 0277-6715. doi: 10.1002/sim.4780121403.

C. S. Moskowitz and M. S. Pepe. Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics*, 5:113–127, 2004. ISSN 14654644. doi: 10.1093/biostatistics/5.1.113.

M. S. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2004.

L. L. Pesce, C. E. Metz, and K. S. Berbaum. On the Convexity of ROC Curves Estimated from Radiological Test Results. *Academic Radiology*, 17(8):960–968.e4, 2010. ISSN 10766332. doi: 10.1016/j.acra.2010.04.001.

M. Plummer. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *Distributed Statistical Computing (DSC 2003)*, pages 1–10, 2003.

R Core Team. R: A Language and Environment for Statistical Computing, 2015. ISSN 16000706.

J. B. Reitsma, A. S. Glas, A. W. S. Rutjes, R. J. P. M. Scholten, P. M. Bossuyt, and

A. H. Zwinderman. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology*, 58(10):982–90, oct 2005. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2005.02.022.

C. M. Rutter and C. A. Gatsonis. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in medicine*, 20(19):2865–2884, 2001. doi: 10.1002/sim.942.

M. Shaked and G. Shanthikumar. *Stochastic Orders*. Springer-Verlag New York, 1 edition, 2007.

S.-Y. Shiu and C. Gatsonis. The predictive receiver operating characteristic curve for the joint assessment of the positive and negative predictive values. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 366 (1874):2313–33, 2008. ISSN 1364-503X. doi: 10.1098/rsta.2008.0043.

P. D. Sørensen, E. H. Jakobsen, J. S. Madsen, E. B. Petersen, R. F. Andersen, B. Østergaard, and I. Brandslund. Serum HER-2: sensitivity, specificity, and predictive values for detecting metastatic recurrence in breast cancer patients. *Journal of Cancer Research and Clinical Oncology*, 139(6):1005–1013, 2013. ISSN 0171-5216. doi: 10.1007/s00432-013-1411-7.

T. A. Trikalinos, C. M. Balion, C. I. Coleman, L. Griffith, P. L. Santaguida, B. Vandermeer, and R. Fu. Meta-Analysis of Test Performance When There Is a Gold Standard. In S. Chang, D. Matchar, and G. Smetana, editors, *Methods Guide for Medical Test Reviews*, chapter 8. Rockville (MD): Agency for Healthcare Research and Quality (US), 2012. URL `http://www.ncbi.nlm.nih.gov/books/NBK98250/`.

T. A. Trikalinos, D. C. Hoaglin, K. M. Small, N. Terrin, and C. H. Schmid. Methods for the joint meta-analysis of multiple tests. *Research Synthesis Methods*, 5(4): 294–312, 2014. ISSN 17592879. doi: 10.1002/jrsm.1115.

Chapter 2

Implications of using hazard ratio to characterize performance of a prognostic biomarker

Abstract

The hazard ratio is commonly reported in studies evaluating prognostic biomarkers. In this chapter, we undertake a critical evaluation of the widespread use of hazard ratio as a summary measure of the prognostic performance of biomarkers. We use the framework of time-dependent receiver operating characteristic curves for this purpose. Under the proportional hazard assumption, we show that the same hazard ratio can result in very different levels of prognostic performance under different marker positivity rates, and baseline hazard functions. For example, a biomarker with a hazard ratio of 5.75 and marker positivity rate of 0.1 will at best only be able to correctly predict approximately 40% of patients who will have the clinical outcomes, and wrongly predict approximately 10% of patients who will not have the clinical outcomes. We show that differences in prognostic performance for different hazard ratios

will diminish with time, and provide examples from the literature to illustrate the inadequacies of using the hazard ratio alone to characterize prognostic performance of biomarkers. Essential information that should accompany the reporting of hazard ratios is identified to allow appropriate assessment of the prognostic performance of a biomarker.

2.1 Introduction

Biomarkers are commonly used to predict the course of disease and response to therapy. In 2001, the National Institutes of Health Biomarkers Definitions Working Group proposed to define a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Atkinson A.J. et al., 2001). In more recent years, new definitions have arisen to better distinguish the different roles that biomarkers play (Sargent and Mandrekar, 2013, Ballman, 2015). A predictive biomarker is defined as a marker, or combination of markers, that identify a specific treatment regimen that is effective only in a subgroup of patients. A prognostic biomarker separates populations with regard to clinical outcomes in either untreated patients or patients treated with standard treatment, and thus is useful in guiding clinical decisions on whom to treat. The primary endpoints in trials of prognostic and/or predictive biomarkers are usually overall, disease-free, or recurrence-free survival (Sargent et al., 2005), which typically lead to the use of survival analysis on data from such trials.

The Cox Proportional Hazard model, or Cox model (Cox, 1972), is the most commonly used method for analyzing survival outcomes. The model assumes that the ratio of the hazard rates of two groups is constant over time (proportional hazard),

and a linear relationship exists between the covariates and the log hazard function or log hazard rates. The hazard rate quantifies the likelihood a patient will experience an event during a defined interval of observation as a rate or percentage (Klein and Moeschberger, 2003). In a review of 50 studies reporting on prognostic tumor markers for cancer, Mallett et al. (2010) found that 98% of these studies used the Cox model. According to the reporting guideline for tumor marker prognostic studies (Altman et al., 2012), reporting a hazard ratio estimate or some other appropriate univariable association measures to show the unadjusted prognostic strength is recommended. The hazard ratio is also one of the two recommended outcome measures for survival time data in the CONSORT statement (Moher et al., 2012). Furthermore, in systematic reviews involving effect measures for time-to-event (survival) outcomes, both the Cochrane Handbook for Systematic Reviews of Interventions (The Cochrane Collaboration, 2011) and PRISMA (Liberati et al., 2009) suggest the use of hazard ratio as the summary measure.

On the other hand, authors have cautioned against the widespread and sometimes indiscriminate use of the hazard ratio. In the general context of survival analysis, limitations on the use of the hazard ratio as a summary measure have been discussed in Ware (2006), Hernán (2010), Blagoev et al. (2012). In the area of systematic reviews of prognostic tests, Rector et al. (2012) argued that the hazard ratio is just indicative of whether a more definitive evaluation of the prognostic biomarker is warranted, and has minimal direct impact on clinical practice. For situations where the proportional hazard assumption is not valid, discussions of alternatives to the hazard ratio can be found in Uno et al. (2014) and Uno et al. (2015). Nevertheless, we continue to see the use of hazard ratio as the primary measure for characterizing the performance of a prognostic biomarker in the literature even though no study has examined how well the hazard ratio can be used for this purpose.

In this study, we will focus on prognostic biomarkers and assume that the proportional hazard assumption is valid. The results can be extended to predictive biomarkers, but will need to bear in mind that the interpretations will be treatment specific. We aim to gain greater insights into the suitability of using the hazard ratio in characterizing the prognostic performance of a biomarker, by critically examining how a prognostic biomarker's hazard ratio translates to its prognostic performance over time. To answer this question, we will make use of the clinically intuitive framework of time-dependent Receiver Operating Characteristic (ROC) curves (Heagerty et al., 2000) to characterize the influence of the hazard ratio on prognostic performance.

The next section will provide technical details on the approach adopted in this chapter to study the implications of hazard ratios on prognostic capabilities of biomarkers. We will present and discuss results primarily for binary prognostic biomarkers. Unique features associated with continuous biomarkers will be highlighted. Two primary studies involving binary biomarkers (Chung et al., 2014, Efficace et al., 2015) and one other involving continuous biomarkers (Liao et al., 2012), are selected from the literature to help illustrate the implications on prognostic performance based on the reported hazard ratios. Each of these studies examined two or more biomarkers to predict the same outcome. We will conclude with a discussion of the findings. In the rest of the chapter, the terms prognostic biomarkers, biomarkers and markers are used interchangeably.

2.2 Methods

2.2.1 Proportional hazard model

In this chapter, the proportional hazard model is written as $\log h(t) = \log h_0(t) + \beta Y$, where $h(t)$ and $h_0(t)$ are the hazard and baseline hazard functions, respectively.

Y denotes biomarker measurement at baseline, with larger values of Y assumed to be more indicative of disease status or event occurrence. The proportional hazard assumption arises because the hazard function is proportional to the baseline hazard function by a factor of $\exp(\beta Y)$, and β is also known as the log hazard ratio. The nomenclature β and log hazard ratio will be used interchangeably from here on. The relationship between survival probability and the hazard function is given by $S_{T|Y=y}(t) = \exp\{-\int_0^t h(u;y)du\}$. In using the partial maximum likelihood estimation method to estimate β, the specification of the baseline hazard function is not required, and thus cannot be estimated directly.

Three different types of baseline hazard functions, namely constant, decreasing, and unimodal, are utilized in this chapter to examine the effect of different baseline hazard functions on the prognostic performance of biomarkers. The Generalized Weibull distribution (Mudholkar et al., 1996) is used to allow specification of different forms of the baseline hazard function, while maintaining the proportional hazard property. The quantile function of the generalized Weibull distribution used is

$$Q(u) = \begin{cases} b\left[\frac{1-(1-u)^{\lambda}}{\lambda}\right]^{1/a}, & \lambda \neq 0 \\ \\ b[-\log(1-u)]^{1/a}, & \lambda = 0 \end{cases}$$

where a and b are respectively the shape and scale parameters in the Weibull distribution, and λ is the third parameter defined to extend the Weibull distribution to contain unimodal and bathtub shaped hazard functions. The values of the parameters used for the different baseline hazard functions are summarized in Table 2.1. The proportional hazard assumption is attained with the specification $S_{T|Y=y}(t) = \left\{S_{T|Y=0}(t)\right\}^{\exp(\beta y)}$.

Table 2.1: Parameter values for baseline hazard functions

Baseline hazard function	a	b	λ
Constant	1	1	0
Decreasing	0.8	5	-1.5
Unimodal	3	5	-3

2.2.2 Time-dependent ROC curves

In classical ROC analysis of diagnostic tests, the binary target condition, often referred to as "disease status", is assumed known and fixed. However in settings where patients' disease status can change with time, i.e. patients are initially free of disease at the start of study but become diseased after time t, the classical approach is no longer suitable. Time-dependent ROC analysis (Heagerty et al., 2000), $ROC(t)$, was thus developed to address how well a prognostic marker measured at baseline can distinguish between patients who will become diseased, and those who will not in a follow-up interval $[0, t]$. In the time dependent ROC framework, different definitions of time dependent sensitivity and specificity were developed for specific purposes, and these were summarized in Heagerty and Zheng (2005).

Let T denote disease onset time. Under the time dependent ROC framework, the binary disease status D is defined as $D(t) = 1$ if $T \leq t$, and $D(t) = 0$ if $T > t$. The time dependent cumulative sensitivity and dynamic specificity are defined as

$$\text{cumulative sensitivity}(c; t) = P(Y > c | T \leq t)$$

$$\text{dynamic specificity}(c; t) = P(Y \leq c | T > t)$$

where c is the specified threshold value. Cumulative sensitivity is thus the probability that the biomarker measurement is above the specified threshold at baseline for those with onset of disease in the follow-up interval, and dynamic specificity is the probability that the measurement is below the threshold at baseline among those

with no disease onset in the same follow-up interval. For each time t, the time-dependent ROC curve is then obtained by plotting cumulative sensitivity against $1 -$ dynamic specificity for all possible threshold values. Other commonly used definitions from Heagerty and Zheng (2005) include

$$\text{incident sensitivity}(c;t) = P(Y > c|T = t)$$
$$\text{static specificity}(c;t^*) = P(Y \leq c|T > t^*)$$

where t^* is a fixed follow-up time point. For this study, we will use cumulative sensitivity and dynamic specificity for the time-dependent ROC as these definitions have greater clinical relevance.

2.2.3 Binary biomarkers

For a binary marker, $Y \in \{0, 1\}$, the cumulative sensitivity or cumulative True Positive Rate (TPR) from the time-dependent ROC framework is defined as

$$\begin{aligned} TPR(t) &= P(Y = 1|T \leq t) \\ &= \frac{[1 - S_{T|Y=1}(t)]P(Y = 1)}{[1 - S_{T|Y=1}(t)]P(Y = 1) + [1 - S_{T|Y=0}(t)]P(Y = 0)} \\ &= \frac{[1 - \{S_{T|Y=0}(t)\}^{exp(\beta)}]p}{[1 - \{S_{T|Y=0}(t)\}^{exp(\beta)}]p + [1 - S_{T|Y=0}(t)](1 - p)} \end{aligned}$$

where $p = P(Y = 1)$ is the marker positivity rate, and the third equality is due to the proportional hazard assumption. Similarly the dynamic False Positive Rate (FPR),

or $1-$ dynamic specificity, is

$$\begin{aligned} FPR(t) &= P(Y=1|T>t) \\ &= \frac{S_{T|Y=1}(t)p}{S_{T|Y=1}(t)p + S_{T|Y=0}(t)(1-p)} \\ &= \frac{\{S_{T|Y=0}(t)\}^{exp(\beta)}p}{\{S_{T|Y=0}(t)\}^{exp(\beta)}p + S_{T|Y=0}(t)(1-p)} \end{aligned}$$

Based on the definition of Y, β will be a non-negative real number. From the above expressions for cumulative $TPR(t)$ and dynamic $FPR(t)$, it is noted that for $0 < p < 1$, $\lim_{t\to\infty} TPR(t) = p$ and $\lim_{t\to\infty} FPR(t) = 0$ for the range of β assumed here.

Under an equal misclassification cost assumption, the prognostic performance of the binary prognostic marker can be measured using a time-varying version of the Youden's Index (Youden, 1950), $J(t) = TPR(t) - FPR(t)$. Note that, if an empirical $ROC(t)$ curve is drawn using the points (0,0), $(TPR(t), FPR(t))$, and (1,1), then the Youden's Index is equivalent to $2AUC(t) - 1$, where $AUC(t)$ is the area under the empirical $ROC(t)$ curve. Hence $J(t) = 0.8$ is equivalent to an empirical $AUC(t) = 0.9$.

2.2.4 Continuous biomarkers

When Y represents a continuous measurement of a biomarker, and denoting the decision threshold by c, the cumulative TPR is defined as

$$\begin{aligned} TPR(c;t) &= P(Y>c|T\le t) \\ &= \frac{\int_c^\infty \left\{1 - S_{T|Y=y}(t)\right\} f_Y(y)dy}{1 - S_T(t)} \\ &= \frac{\int_c^\infty \left\{1 - S_{T|Y=0}(t)^{exp(\beta y)}\right\} f_Y(y)dy}{\int_{-\infty}^\infty \left\{1 - S_{T|Y=0}(t)^{exp(\beta y)}\right\} f_Y(y)dy} \end{aligned}$$

and the dynamic FPR is defined as

$$\begin{aligned} FPR(c;t) &= P(Y > c | T > t) \\ &= \frac{\int_c^{\infty} S_{T|Y=y}(t) f_Y(y) dy}{S_T(t)} \\ &= \frac{\int_c^{\infty} \left\{ S_{T|Y=0}(t)^{exp(\beta y)} \right\} f_Y(y) dy}{\int_{-\infty}^{\infty} S_{T|Y=0}(t)^{exp(\beta y)} f_Y(y) dy} \end{aligned}$$

The time-dependent ROC curve is then defined as $ROC(v;t) = TPR(FPR^{-1}(v;t);t)$. A commonly used summary measure of the sensitivity and specificity over the range of possible thresholds for the marker is the time-dependent area under the ROC curve, $AUC(t)$.

$$AUC(t) = \int_0^1 ROC(v;t) dv$$

or

$$AUC(t) = \int_{-\infty}^{\infty} TPR(c;t) \left| \frac{\partial FPR(c;t)}{\partial c} \right| dc$$

From the expressions above, it is observed that

$$\begin{aligned} TPR(\infty;t) &= FPR(\infty;t) = 0 \\ TPR(-\infty;t) &= FPR(-\infty;t) = 1 \\ \lim_{t \to \infty} TPR(c;t) &= 1 - F_Y(c) \\ \lim_{t \to \infty} FPR(c;t) &= \begin{cases} 0 & , \beta > 0 \\ 1 - F_Y(c) & , \beta = 0 \end{cases} \end{aligned}$$

where $F_Y(y)$ is the cumulative distribution of the continuous marker. It can also be shown that $AUC(t) = 0.5$ when $\beta = 0$. For the purpose of this chapter, we assume that $Y \sim N(0,1)$. All numerical integrations were performed using adaptive quadrature in R 3.2.3 (R Core Team, 2015).

2.3 Results

For simplicity in notation, the dependence of TPR, FPR, ROC and AUC on t will be dropped in the rest of the chapter.

2.3.1 Binary biomarker

For a binary marker, both TPR and FPR are functions of the marker positivity rate (p), log hazard ratio (β), baseline hazard function and time (t). These relationships can be represented in a plot shown in Figure 2.1, where the dependence on time is through the marginal survival function. These curves were generated assuming a constant baseline hazard function with parameters given in Table 2.1. For both TPR and FPR curves, the horizontal line at the value of p represents the case when the biomarker has no discriminating ability, i.e. $\beta = 0$, and TPR is equal to FPR. The TPR lines (solid lines) for different values of β are above this horizontal $\beta = 0$ line, and will eventually converge towards the value p as time increases. On the other hand, FPR lines (dot-dash lines) originate from the value p, at the start and decrease towards zero as time increases. With increasing values of β, the TPR and FPR lines move further away from the reference horizontal line of $\beta = 0$. Three different colors (black, red, and green) are cycled through continuously to facilitate the identification of the corresponding log hazard ratio in the legend in Figure 2.1a. This legend applies to Figures 2.1, 2.2 and 2.3.

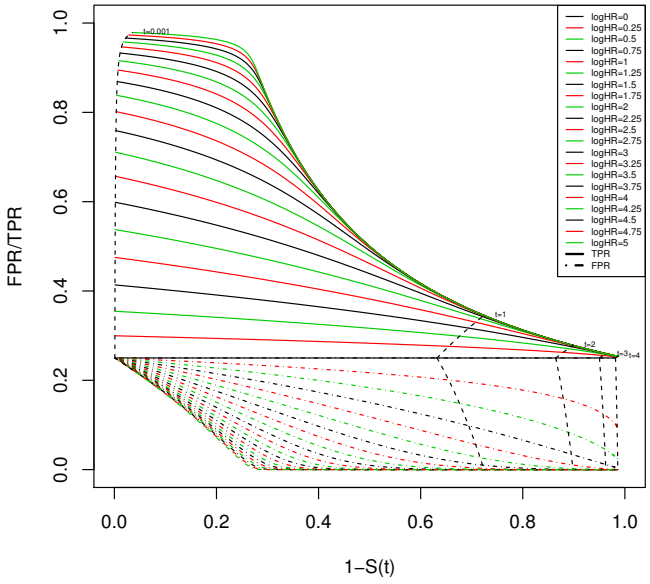
Intuitively the trends indicate that at the point where measurements of biomarker are taken, the patients will not have developed the event yet and will be in the event-free group. Thus FPR will be the same as the marker positivity rate. On the other hand, if time is allowed to increase infinitely, a majority of the patients will have the event eventually. The true positive rate will then approach the marker positivity rate, while

the false positive rate will decrease towards zero. Figure 2.1 shows two plots with different marker positivity rates to illustrate the similar behavioral trends of TPR and FPR with time. The actual time scale is also superimposed on each of the plots as isochronic dashed lines. The shifts in the isochronic dashed lines will depend on the baseline hazard function used. Keeping the marker positivity rate at 0.25, the results from using unimodal and decreasing baseline hazard functions are shown in Figure 2.2a and 2.2b respectively.

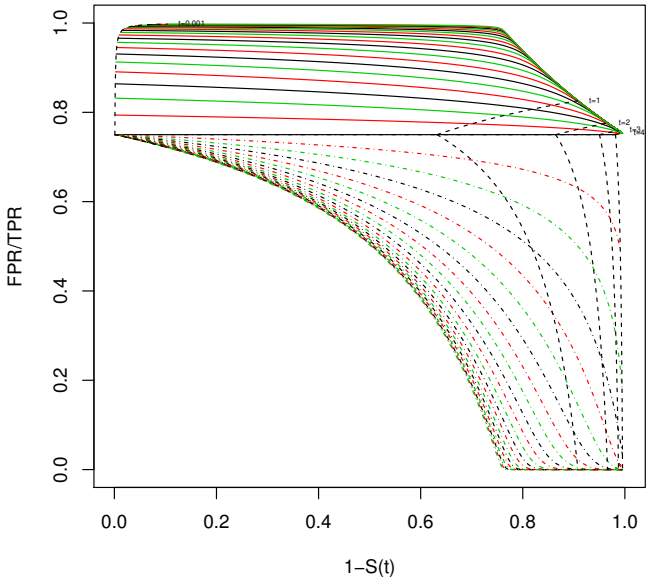
From the plots in Figure 2.1, we observed that at a fixed marker positivity rate, p, a larger hazard ratio will result in better cumulative sensitivity and dynamic specificity. However the improvements in specificity are much smaller than sensitivity in the initial period, as shown by how tightly the FPR curves are bunched up together in the early stages. This indicates that changes in specificity are not very sensitive to changes in β, but can be sensitive to changes in the value of p in the short term.

From Figure 2.2, it is observed that different baseline hazard functions will result in a geometrical translation of the isochronic lines in the plot. Higher baseline hazard rates will see faster rates of decrease in cumulative sensitivity toward the value of p, and faster rates of increase in dynamic specificity to 1, thus shifting the isochronic lines to the right. Conversely, lower baseline hazard rates will shift the isochronic lines to the left. Hence the shape of the plot for a given marker positivity rate is largely invariant to the form of the baseline hazard function.

More importantly, the plots allow us to make the observation that for a practical range of β from 0.5 to 1.5 (hazard ratio from 1.65 to 4.48), the cumulative sensitivity for a marker with a positivity rate of 0.25 is at most 0.6 with the corresponding dynamic specificity of approximately 0.75, regardless of the baseline hazard function. On the other hand, a marker with a positivity rate of 0.75 will have at most a cumulative sensitivity of approximately 0.9 with the corresponding dynamic specificity of 0.25

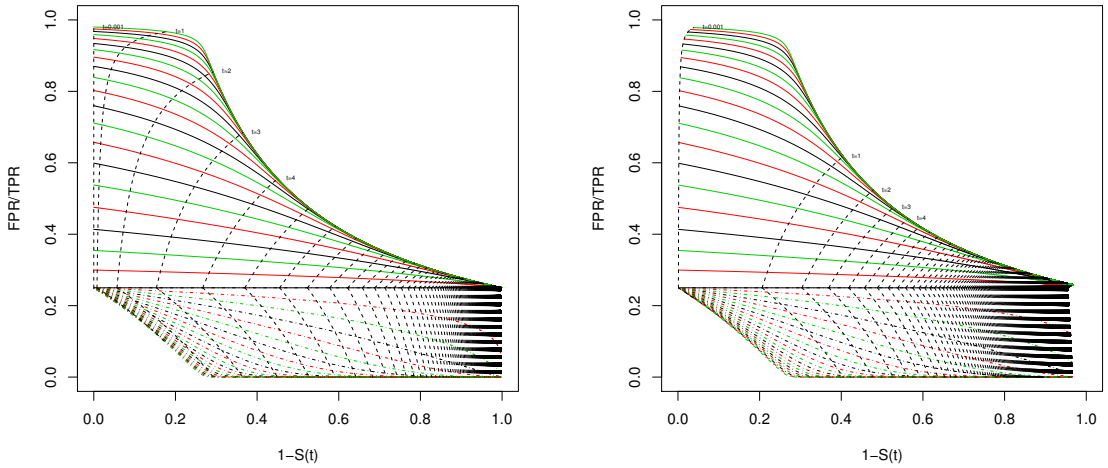


(a) Marker positivity rate, $p = 0.25$



(b) Marker positivity rate, $p = 0.75$

Figure 2.1: Variation of TPR and FPR with log hazard ratio (β), and time t with constant baseline hazard function for different marker positivity rates. Note that $S(t) = P(T \leq t)$. Different *color* codings refer to the different levels of β, starting from the horizontal line of $\beta = 0$ at the marker positivity rate. Different *line-type* codings are used to differentiate between TPR curves (solid), FPR curves (dot-dashed), and isochronic time curves (dashed).



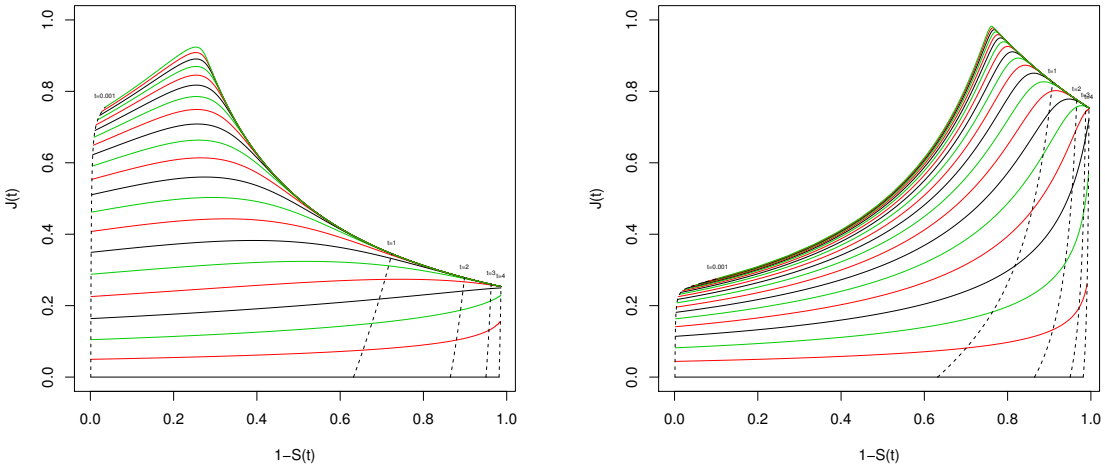
(a) Unimodal baseline hazard function (b) Decreasing baseline hazard function

Figure 2.2: Variation of TPR and FPR with log hazard ratio (β), and time (t) at marker positivity rate $p = 0.25$. Baseline hazard functions are as indicated in each sub-plot. Note that $S(t) = P(T \leq t)$. Different *color* codings refer to the different levels of β, starting from the horizontal line of $\beta = 0$ at the value of p in accordance to the legend in Figure 2.1. Different *line-type* codings are used to differentiate between TPR curves (solid), FPR curves (dot-dashed), and isochronic time curves (dashed).

for the same range of β. Thus the same value of β implies very different ranges of cumulative sensitivity/dynamic specificity for markers with different positivity rates.

The prognostic performance of biomarkers with different values of p, but the same baseline hazard function, are shown in Figure 2.3. From the figure, we observe that a maximum J exists for each value of log hazard ratio, and the occurrence of this optimal point approaches the value of $1 - S(t) = p$ with increasing log hazard ratio. Furthermore, the prognostic performance for different hazard ratios are observed to coalesce with increasing time. We also note that for a marker with prevalence of 0.25 and $\beta = 1.75$, $J(t = 1) \approx 0.35$. However, a marker with prevalence of 0.75 and for the same $\beta = 1.75$, $J(t = 1) \approx 0.8$. Hence even when baseline hazard functions are the same, larger hazard ratios do not necessarily imply better prognostic performance.

In another situation, we assume that we have two biomarkers Y_1 and Y_2 with the same

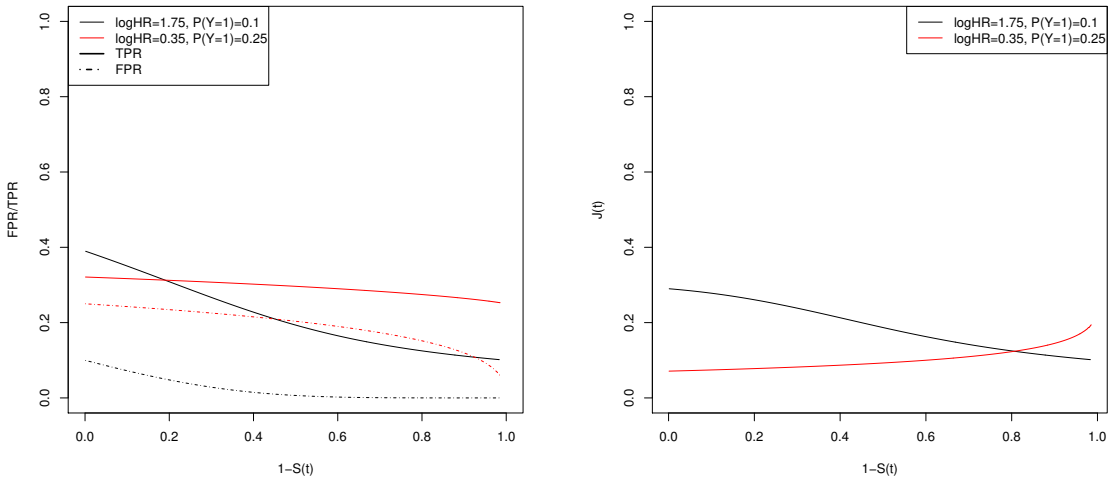


(a) Marker positivity rate, $p = 0.25$ (b) Marker positivity rate, $p = 0.75$

Figure 2.3: Variation of J with log hazard ratio (β), and $1 - S(t)$ with constant baseline hazard function at different values of p. The color codings are for different values of β according to the legend in Figure 2.1a.

constant baseline hazard function. The reported statistics are $\beta_1 = 1.75$, $\beta_2 = 0.35$, $P(Y_1 = 1) = 0.1$ and $P(Y_2 = 1) = 0.25$. The corresponding variations in TPR and FPR are shown in Figure 2.4a. In this case, the cumulative sensitivity of biomarker Y_1 varies from approximately 0.4 to 0.1, and the dynamic specificity varies from 0.9 to 1 with increasing time. For biomarker Y_2, the cumulative sensitivity varies from approximately 0.321 to 0.250, and dynamic specificity varies from 0.75 to 1 with increasing time. The prognostic performance of the two biomarkers are shown in Figure 2.4b. In this case, we noted crossings in the cumulative sensitivity curves and prognostic performance curves, implying that the superiority of the prognostic performance of one biomarker relative to the other can depend on the time frame of interest, which again will not be obvious from the hazard ratios alone.

So far in the examples that we have examined, we have assumed the same baseline hazard function to allow us to investigate the effects of log hazard ratio and marker positivity rate on the prognostic performance. However in an actual situation com-



(a) TPR and FPR (b) $J(t)$

Figure 2.4: Variation of performance measures for 2 binary biomarkers with $\beta_1 = 1.75$, $\beta_2 = 0.35$, $P(Y_1 = 1) = 0.1$ and $P(Y_2 = 1) = 0.25$.

paring the prognostic performance of two different binary biomarkers, the baseline hazard function for each of the biomarker will have to be provided. This function determines how fast or slow the performance measures travel along their respective contours in the corresponding $TPR(t)/FPR(t)$ vs. $1 - S(t)$ and $J(t)$ vs. $1 - S(t)$ plots. Clearly this will complicate the comparisons even further.

2.3.2 Continuous biomarker

The results for binary biomarkers can be extended to continuous biomarkers. In this setting, the cumulative sensitivity and dynamic specificity are also functions of the decision threshold c, and p is represented by the complement of the cumulative distribution function of the continuous biomarker at c. The variation in the time-dependent ROC curve with different values of log hazard ratio is shown in Figure 2.5 at a fixed time $t = 1$ for a constant baseline hazard function. We observe in Figure 2.5 that AUC increases with larger values of log hazard ratio for a fixed time. It is also

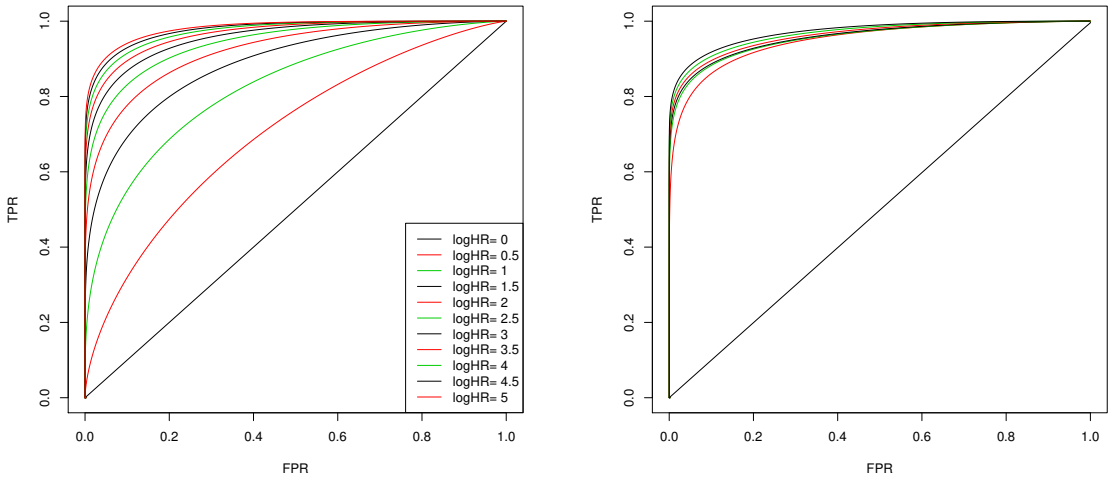
interesting to note that the ROC curves for different log hazard ratio values coalesce with the passage of time regardless of the initial values (except for $\beta = 0$), Figure 2.5b shown with $t = 10$. Intuition for this observation can be obtained from Figure 2.1 for the binary biomarker. Assuming that the 2 plots in Figure 2.1 represent two points on the ROC curve in Figure 2.5. The coalescence of the ROC curves is explained by the respective coalescence of the TPR and FPR curves. We can also infer from this that the time required for the ROC curves to coalesce will depend on the baseline hazard function. This observation implies that in the long run, biomarkers with larger hazard ratios have no prognostic advantages over biomarkers with lower hazard ratios, all yielding similar ROC curves for sufficiently large t. This is similar to the observations made for binary marker.

Another subtle observation is that the ROC curves will tend to become asymmetric with time and move closer to the vertical axis in the ROC plot. This could be due to the rate of decrease in FPR to zero being generally faster than the rate at which TPR decreases to the value of p, and also the rate of decrease in FPR is also generally faster for small p (large c) than large p (small c).

When we use a decreasing baseline hazard function, similar trends are observed, but occurring over a longer period of time. For comparison purposes, we show the patterns of the ROC curves at $t = 10$ in Figure 2.6a, and at $t = 1 \times 10^7$ in Figure 2.6b, which is when coalescence becomes obvious for the decreasing baseline hazard function used here. From Figures 2.5b and 2.6a, we note that $AUC(t = 10)$ for a biomarker with a constant baseline hazard rate and a log hazard ratio of 0.5, is greater than a biomarker that has a decreasing baseline hazard function and a log hazard ratio of 1.5. Hence it is possible to find examples to show that prognostic performance do not always commensurate with the magnitude of the hazard ratio.

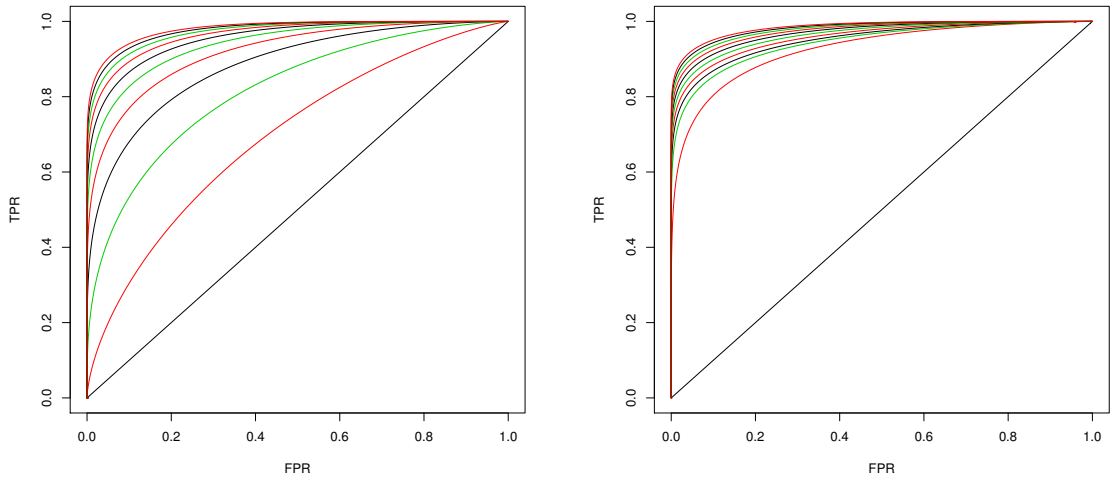
It is commonly known that the same continuous biomarker measured in different

scales can lead to different values for log hazard ratio. More importantly, different measurement scales may result in different baseline hazard function as the reference condition may have changed. The scales of measurement for continuous biomarkers can add another dimension of complexity to the comparison of the prognostic performance of biomarkers based on their hazard ratios alone. Thus distribution of the biomarker and the corresponding baseline hazard function are critical pieces of information required in evaluating prognostic performance.



(a) $t = 1$ (b) $t = 10$

Figure 2.5: Variation of ROC(t) curves with different values of log hazard ratio, β, at $t = 1$ and $t = 10$ with constant baseline hazard function. Different *color* codings refer to the different levels of β, starting from the diagonal line with a gradient of 1 for the ROC(t) curve with $\beta = 0$.



(a) $t = 10$ (b) $t = 10,000,000$

Figure 2.6: Variation of ROC(t) curves with different values of log hazard ratio, β, at $t = 10$ and $t = 10,000,000$ with decreasing baseline hazard function. Different *color* codings refer to the different levels of β, starting from the diagonal line with a gradient of 1 for the ROC(t) curve with $\beta = 0$. The legend in Figure 2.5a applies to the figures here.

2.4 Examples

Three recent studies from the literature are selected as examples to illustrate the implications arising from the use of hazard ratio to characterize prognostic performance of biomarker. None of the studies explicitly showed the results for testing of the proportional hazard assumption, and not all reported that the assumption was valid. For our purpose here, we assume that the proportional hazard assumption is valid in all three studies. Baseline survival functions were digitized and extracted using `Plot Digitizer` 2.6.8 (Huwaldt and Steinhorst). The data points are then fitted with a linear model, $-\log S(t) = \gamma_1 t + \gamma_2 t^2 + \gamma_3 t^3$. Analyses are performed for the reported time frame in the study to avoid extrapolation of data.

2.4.1 Prognostic biomarkers of nonoropharyngeal head and neck squamous cell carcinoma

In Chung et al. (2014), p16 protein expression and human papillomavirus (HPV) status were studied as prognostic biomarkers of Nonoropharyngeal Head and Neck Squamous Cell Carcinoma (non-OSPCC) in patients with non-OSPCC tumors who were enrolled onto three prospective Radiation Therapy Oncology Group clinical trials. Immunohistochemistry (IHC) was used to determine p16 expression, with positive expression defined as strong and diffuse nuclear and cytoplasmic staining in $\geq 70\%$ of the tumor cells. HPV status was determined by in situ hybridization (ISH) for a range of HPV types, and defined as positive when nuclear-specific staining was detected in the tumor cells. The reported marker positivity rate and hazard ratio for progression free survival outcome for positive p16 protein expression were estimated to be 0.238 and 0.63, respectively. The corresponding values for positive HPV status were estimated to be 0.104 and 0.77, respectively. These results indicated that positive status of the biomarkers was protective. In line with the convention used in general ROC analyses, coding status of the biomarkers is reversed such that a positive indication is associated with higher risk of non-OPSCC. Under this revised coding scheme for the biomarkers, the marker positivity rate and hazard ratio for negative p16 protein expression are estimated to be 0.762 and 1.587 respectively, and the marker positivity rate and hazard ratio for negative HPV status are estimated to be 0.896 and 1.299 respectively. From Figure 1 of Chung et al. (2014), the baseline survival functions for the reference groups based on p16 expression and HPV status were provided in sub-plots (A) and (C), respectively.

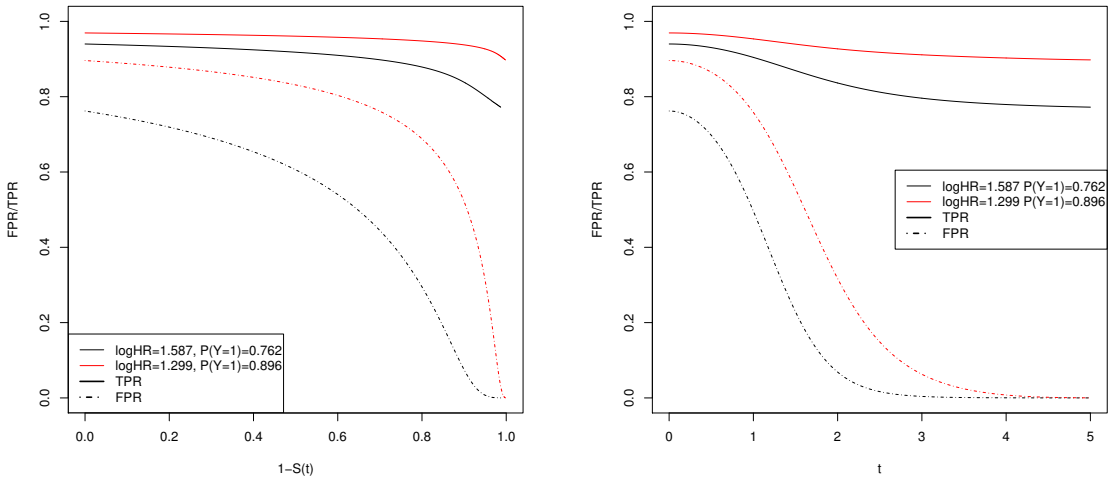
Denote Y_1 as negative p16 protein expression and Y_2 as negative HPV status with the respective parameters $\beta_1 = \log(1.587)$, $\beta_2 = \log(1.299)$, $P(Y_1 = 1) = 0.762$ and $P(Y_2 = 1) = 0.896$. The corresponding variations in TPR and FPR are shown in

Figure 2.7. As time increases, the cumulative sensitivity of biomarker Y_1 varies from approximately 0.835 to 0.762, and the dynamic specificity varies from approximately 0.238 to 1. For biomarker Y_2, the cumulative sensitivity varies from approximately 0.918 to 0.896, and dynamic specificity varies from approximately 0.104 to 1 with increasing time. The prognostic performance of the two biomarkers are shown in Figure 2.8.

From the above example comparing negative p16 protein expression and negative HPV status as prognostic biomarkers of non-OPSCC, while negative p16 protein expression has a higher hazard ratio than negative HPV status, it has better prognostic performance only up to approximately 2.5 years from baseline biomarker measurement. Beyond that, HPV performs better than p16 protein expression as a biomarker for non-OPSCC. On the other hand, when having high sensitivity is of utmost importance, then the results show that negative HPV status is a better prognostic biomarker.

2.4.2 Prognostic value of self-reported fatigue on myelodysplastic syndromes

In this study by Efficace et al. (2015), the authors investigated the prognostic value of self-reported fatigue on overall survival in patients with myelodysplastic syndromes. Patients were enrolled within 6 months of diagnosis with an intermediate-2-risk or high-risk score according to the International Prognostic Scoring System (IPSS). Self-reported fatigue score was obtained from the fatigue scale of the European Organisation for Research and Treatment of Cancer quality of life questionaire-core 30 (EORTC QLQ-C30) administered at baseline. The overall survival based on baseline patient's self-reported fatigue severity and IPSS risk group were provided in Figure 1 of Efficace et al. (2015). The fatigue score was dichotomized as low fatigue if the

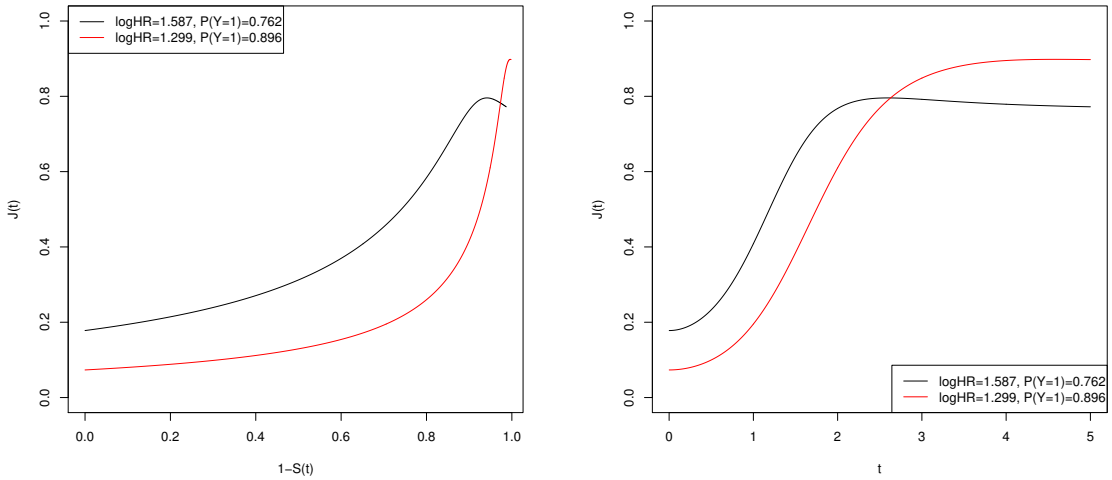


(a) TPR/FPR vs. $1 - S(t)$ (b) TPR/FPR vs. t

Figure 2.7: Variations of TPR and FPR for the 2 binary biomarkers in the example, negative p16 and negative HPV. Negative p16 protein expression (black) has a log hazard ratio of $\beta_1 = \log(1.587)$, and marker positivity rate $P(Y_1 = 1) = 0.762$. Negative HPV status (red) has a log hazard ratio of $\beta_2 = \log(1.299)$ (red), and marker positivity rate of $P(Y_2 = 1) = 0.896$. The plot on the left has $1 - S(t)$ on the horizontal scale, while the one on the right is actual time scale t.

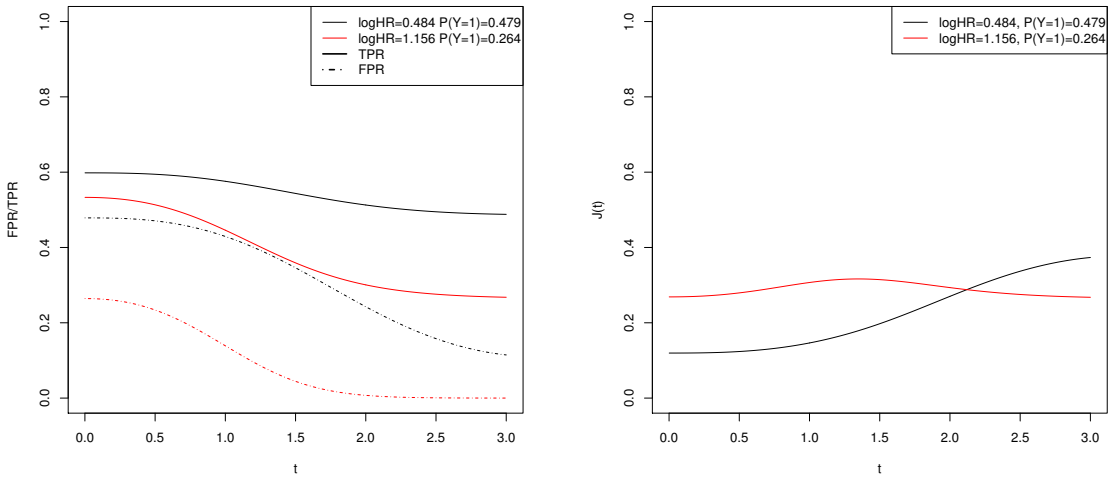
value is lower than 34 points, and high otherwise. The reported hazard ratio for self-reported fatigue was 1.622, or $\beta_1 = 0.484$, and a marker positivity rate of 0.479. The reported hazard ratio using IPSS risk categorization was 3.178, $\beta_2 = 1.156$, with a marker positivity rate of 0.264. Similar to the previous example, here we also observe from Figure 2.9 that IPSS has a higher hazard ratio than self-reported fatigue score, and better prognostic performance up to approximately 2 years. Beyond that, self-reported fatigue performs better. In the case where having high sensitivity is of utmost importance, self-reported fatigue is a better prognostic biomarker for myelodysplastic syndromes.

The authors also reported the hazard ratio associated with a continuous self-reported fatigue score. In this case, the fatigue scores ranged between 0 and 100, with higher scores indicating higher levels of fatigue. The reported hazard ratio was 1.130 for



(a) J vs. $1 - S(t)$ (b) J vs. t

Figure 2.8: Variations of the Youden Index J for the 2 binary biomarkers in the non-OPSCC example, negative p16 and negative HPV. Negative p16 protein expression (black) has a log hazard ratio of $\beta_1 = \log(1.587)$, and marker positivity rate of $P(Y_1 = 1) = 0.762$. Negative HPV status (red) has a log hazard ratio of $\beta_2 = \log(1.299)$ (red), and marker positivity rate of $P(Y_2 = 1) = 0.896$. The plot on the left has $1 - S(t)$ on the horizontal scale, while the one on the right is actual time scale t.



(a) TPR/FPR vs. t (b) J vs. t

Figure 2.9: Variation in the prognostic performance for the 2 binary biomarkers in the self-reported fatigue example. Self-reported fatigue (black) has a log hazard ratio of $\beta_1 = 0.484$, and marker positivity rate of $P(Y_1 = 1) = 0.479$. IPSS (red) has a log hazard ratio of $\beta_2 = 1.156$ (red), and marker positivity rate of $P(Y_2 = 1) = 0.264$. Both plots are based on actual time scale t.

every 10 point shift difference on the scale. However no further details regarding the distribution of the continuous self-reported fatigue score and the baseline hazard function of the baseline reference group were provided. It is not clear from the paper if the baseline reference group refers to those with zero fatigue score, or those with mean score or median score if centering was performed. In this case, the reported hazard ratio of 1.130 is uninformative with regards to the prognostic capability of self-reported fatigue.

2.4.3 Prognostic value of quantitative metabolic volumetric measurement on 18F-FDG PET/CT

The third example is based on the study by Liao et al. (2012), in which metabolic tumor volume (MTV), total lesion glycolysis (TLG), and maximum standardized uptake values (SUV) of whole body tumors were measured in nonsurgical patients with Stage IV non-small cell lung cancer for assessment of their prognostic values. These measurements were log transformed to reduce the skewness in the original scale. No further details on the distributions of the transformed biomarker measurements were provided. Survival functions were provided for dichotomized biomarkers based on the median values. The reported hazard ratios for the continuous log transformed biomarkers were 1.48, 1.37 and 1.27 respectively. The authors concluded that that MTV and TLG measurements were better prognostic measures than SUV measurements. However as shown earlier, only when all the biomarkers have the same distribution and the same baseline hazard rate, then it is possible to conclude that one would have better prognostic performance than the other on the basis of a higher hazard ratio. Comparison of the biomarkers' prognostic performance can only be made if their measurement distributions and the reference baseline hazard functions are provided. Hazard ratio alone is not sufficient to inform readers about the comparative

prognostic performance of each continuous biomarker, and may even be potentially misleading.

2.5 Conclusion

We have examined the implications of a reported hazard ratio in the context of prognostic performance for both binary and continuous biomarker measurements under the assumption that proportional hazard is valid. In the presence of valid proportional hazard assumption, we have shown that the same hazard ratio can result in very different prognostic performance under different marker positivity rates and baseline hazard functions. In some cases, a much stronger association with the disease outcome than typically observed in primary research is needed to provide a clinically usable prognostic capability. For example, a biomarker with a hazard ratio of 5.75 and marker positivity rate of 0.1 will at best, only be able to correctly predict approximately 40% of patient who will have the clinical outcomes, and wrongly predict approximately 10% of patients who will not have the clinical outcomes.

We have also demonstrated instances where the relative prognostic performance of two binary biomarkers with different hazard ratios and marker positivity rates can switch with time, thus illustrating that comparison of prognostic capabilities based on hazard ratio alone can be misleading. In the case of continuous biomarkers, we have found that the time-dependent ROC curves representing different hazard ratios coalesce with time, thus minimizing any differences in prognostic performance due to differences in hazard ratios. The rate at which the ROC curves coalesce depends strongly on the baseline hazard functions. We have shown at a particular point in time, that a biomarker with a constant baseline hazard rate and a hazard ratio of 1.65 could have a greater AUC than a biomarker with a different baseline hazard function

and a hazard ratio of 4.48.

It is clear that hazard ratio alone is inadequate in conveying the information about the prognostic performance of a biomarker. A large hazard ratio does not necessarily imply better prognostic performance. In the examples selected from primary literature, we have shown the importance in presenting details on the distribution of the biomarker, and the corresponding baseline hazard function or baseline survival function, to allow readers to properly assess the prognostic performance of the biomarker. Missing any one of this information will make assessment of the prognostic performance of the biomarker impossible.

In this study, we have only considered performance summary using Youden's Index and AUC. While these summary statistics have their own assumptions and limitations, they remain widely used in current literature and better appreciated by most consumers of such information. We also did not explicitly take censoring into account, but implicitly assume that the use of Cox model has appropriately addressed censoring.

When the proportional hazard assumption is not valid, use of hazard ratio is not indicated. However this has not prevented studies from reporting hazard ratios as the prognostic capabilities of biomarkers. Two common cases when proportional hazard assumption is not valid are when the survival curves cross each other, and when the survival curves do not cross, but the proportionality constant varies with time. The former case implies that the probability of survival of the patient subgroup with larger values of biomarker measurement (e.g. biomarker positive group), becomes higher than the lower valued biomarker subgroup (e.g. biomarker negative group) as time progresses. This is a violation of the assumption that larger values of biomarker measurement are more indicative of disease severity, or likelihood of event occurrences. This will lead to very different behaviors in the time dependent TPR and FPR, with

FPR possibly approaching 1 as time increases. In the other case where proportional hazard assumption is not valid, the trends in TPR and FPR will not change from what we have shown in this chapter since the method of time dependent ROC does not make any assumption regarding proportionality in the hazard functions. However the relationship between the different survival functions are no longer as well defined as $S_{T|Y=y}(t) = \left\{S_{T|Y=0}(t)\right\}^{\exp(\beta y)}$. In order to make any assessment on the prognostic capabilities, one would need the survival functions for all subgroups, and the reported hazard ratio is of no use at all.

For a biomarker to be associated with the outcome, the distributions of the biomarker in the biomarker positive and negative groups have to be different. Even when the distributions have significant overlap, association measure will still be statistically significant if the sample size is sufficiently large. In order to perform well as a prognostic biomarker, the distributions of the biomarker in the two groups must be sufficiently well separated at the time point of interest to allow discrimination. In conclusion, hazard ratio on its own is not a good summary measure of the prognostic performance of a biomarker. It is a summary measure that provides an overall association, but does not adequately represent prognostic performance of a biomarker with time.

Bibliography

D. G. Altman, L. M. McShane, W. Sauerbrei, and S. E. Taube. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. *PLoS medicine*, 9(5):e1001216, jan 2012. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001216.

J. Atkinson A.J., W. a. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing, D. F. Hoth, J. a. Oates, C. C. Peck, R. T. Schooley, B. a. Spilker, J. Woodcock, and S. L. Zeger. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics*, 69(3):89–95, 2001. ISSN 00099236. doi: 10.1067/mcp.2001.113989.

K. V. Ballman. Biomarker: Predictive or Prognostic? *Journal of Clinical Oncology*, sep 2015. doi: 10.1200/JCO.2015.63.3651.

K. B. Blagoev, J. Wilkerson, and T. Fojo. Hazard ratios in cancer clinical trials - a primer. *Nat Rev Clin Oncol*, 9(3):178–183, mar 2012. ISSN 1759-4774.

C. H. Chung, Q. Zhang, C. S. Kong, J. Harris, E. J. Fertig, P. M. Harari, D. Wang, K. P. Redmond, G. Shenouda, A. Trotti, D. Raben, M. L. Gillison, R. C. Jordan, and Q.-T. Le. p16 Protein Expression and Human Papillomavirus Status As Prognostic Biomarkers of Nonoropharyngeal Head and Neck Squamous Cell Carcinoma. *Journal of Clinical Oncology*, 32(35):3930–3938, sep 2014. ISSN 0732-183X. doi: 10.1200/JCO.2013.54.5228.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34:187–220, 1972. ISSN 00359246. doi: 10.2307/2985181.

F. Efficace, G. Gaidano, M. Breccia, M. T. Voso, F. Cottone, E. Angelucci, G. Caocci, R. Stauder, D. Selleslag, M. Sprangers, U. Platzbecker, A. Ricco, G. Sanpaolo, O. Beyne-Rauzy, F. Buccisano, G. A. Palumbo, D. Bowen, K. Nguyen, P. Niscola, M. Vignetti, and F. Mandelli. Prognostic value of self-reported fatigue on overall survival in patients with myelodysplastic syndromes: a multicentre, prospective, observational, cohort study. *The Lancet Oncology*, 16(15):1506–1514, nov 2015. doi: 10.1016/S1470-2045(15)00206-5.

P. J. Heagerty and Y. Zheng. Survival model predictive accuracy and ROC curves 1. *Biometrics*, 61(1):92–105, 2005.

P. J. Heagerty, T. Lumley, and M. S. Pepe. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000. ISSN 0006-341X. doi: 10.1111/j.0006-341X.2000.00337.x.

M. A. Hernán. The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)*, 21 (1):13–15, 2010. ISSN 1044-3983. doi: 10.1097/EDE.0b013e3181c1ea43.

J. A. Huwaldt and S. Steinhorst. Plot Digitizer. URL `http://plotdigitizer.sourceforge.net/`.

J. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, second edition, 2003.

S. Liao, B. C. Penney, H. Zhang, K. Suzuki, and Y. Pu. Prognostic Value of the Quantitative Metabolic Volumetric Measurement on 18F-FDG PET/CT in Stage IV Nonsurgical Small-cell Lung Cancer. *Academic Radiology*, 19(1):69–77, 2012. ISSN 10766332. doi: 10.1016/j.acra.2011.08.020.

A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. a. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Annals of Internal Medicine*, 151 (4), 2009. ISSN 00034819. doi: 10.1371/journal.pmed.1000100.

S. Mallett, A. Timmer, W. Sauerbrei, and D. G. Altman. Reporting of prognostic studies of tumour markers: a review of published articles in relation to REMARK guidelines. *British journal of cancer*, 102(1):173–180, 2010. ISSN 1532-1827. doi: 10.1038/sj.bjc.6605462.

D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman. CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *International Journal of Surgery*, 10(1):28–55, 2012. ISSN 17439191. doi: 10.1016/j.ijsu.2011.10.001.

G. S. Mudholkar, D. K. Srivastava, and G. D. Kollia. A Generalization of the Weibull Distribution with Application to the Analysis of Survival Data. *Journal of the American Statistical Association*, 91(436):1575, 1996. ISSN 01621459. doi: 10.2307/2291583.

R Core Team. R: A Language and Environment for Statistical Computing, 2015. ISSN 16000706.

T. S. Rector, B. C. Taylor, and T. J. Wilt. Systematic review of prognostic tests. In *Journal of General Internal Medicine*, volume 27, chapter 12. 2012.

D. J. Sargent and S. J. Mandrekar. Statistical issues in the validation of prognostic, predictive, and surrogate biomarkers. *Clinical trials (London, England)*, 10(5): 647–52, 2013. ISSN 1740-7753. doi: 10.1177/1740774513497125.

D. J. Sargent, B. a. Conley, C. Allegra, and L. Collette. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*, 23(9):2020–2027, 2005. ISSN 0732183X. doi: 10.1200/JCO.2005.01.112.

The Cochrane Collaboration. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]*. 2011. ISBN 9780470057964. doi: Available from www.cochrane-handbook.org. URL `www.cochrane-handbook.org`.

H. Uno, B. Claggett, L. Tian, E. Inoue, P. Gallo, T. Miyata, D. Schrag, M. Takeuchi, Y. Uyama, L. Zhao, H. Skali, S. Solomon, S. Jacobus, M. Hughes, M. Packer, and L. J. Wei. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32(22):2380–2385, 2014. ISSN 15277755. doi: 10.1200/JCO.2014.55.2208.

H. Uno, J. Wittes, H. Fu, S. D. Solomon, B. Claggett, L. Tian, T. Cai, M. A. Pfeffer, S. R. Evans, and L.-J. Wei. Alternatives to Hazard Ratios for Comparing the Efficacy or Safety of Therapies in Noninferiority StudiesAlternatives to Hazard Ratios. *Annals of Internal Medicine*, 163(2):127–134, jul 2015. ISSN 0003-4819.

J. H. Ware. The Limitations of Risk Factors as Prognostic Tools. *New England Journal of Medicine*, 355(25):2615–2617, dec 2006. ISSN 0028-4793. doi: 10.1056/NEJMp068249.

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. ISSN 0008-543X.

Chapter 3

Causal inference in studies comparing diagnostic test outcomes

Abstract

Most studies of diagnostic tests are designed to assess only the diagnostic and/or predictive performances of tests. In order to evaluate patient outcomes of diagnostic tests, randomized studies (Diagnostic Randomized Controlled Trials, DRCTs) have been utilized. A DRCT evaluates the impact of tests on patient outcomes by coupling randomization with diagnostic tests and therapeutic interventions in a single study. Randomization is used to eliminate bias arising from selection of patients into a treatment strategy, and thus providing a valid estimate of the causal effect of the treatment strategy. However this estimate of the causal effect can become biased in the presence of noncompliance to the assignments of either test, therapeutic intervention, or both. When test results dictate specific treatment recommendations, the test can be seen as a randomizer to treatment, conditional on true disease status. Based

on this insight, we used methods from the sequential randomization literature to adjust for selection bias arising from noncompliance in DRCT studies. In particular, we adapted the Structural Nested Mean Model (SNMM) for use in DRCT designs to estimate the causal effect of a test strategy had, contrary to fact, all subjects remained on protocol. The performances of the causal estimate obtained using SNMM were evaluated via simulations. Simulations also revealed the deficiencies of the commonly used intention to treat (ITT) approach in the presence of noncompliance in DRCT studies. We applied the SNMM to a subset of the National Lung Screening Trial data to illustrate its use.

3.1 Introduction

The key objective of clinical studies of diagnostic tests is typically the assessment of diagnostic and/or predictive performance under specific clinical settings. However when the research question of interest is the effect of tests on patient outcomes, the diagnostic/predictive performances of tests alone are insufficient to answer this question. An alternative class of study designs for this purpose is the Diagnostic Randomized Controlled Trial (DRCT) (De Graaff et al., 2004, Bossuyt et al., 2000). DRCTs take into consideration the impact of tests on patient outcomes, as mediated by the therapeutic interventions. Here, the combination of a test and therapeutic interventions is referred to as a treatment strategy, where the assignment of a specific therapeutic intervention is guided by the result of the test, and the entire approach is evaluated as a package. The key advantage of coupling randomization with diagnostic tests and therapeutic interventions in a single study is that it allows one to eliminate bias arising from selection of patients into a *treatment strategy*, thus providing a valid estimate of the causal effect of the *treatment strategy*. DRCT designs have been examined in Lijmer and Bossuyt (2002), Lu and Gatsonis (2013) and Hooper et al.

(2013). Examples of DRCT designs in practice can be found in De Graaff et al. (2003), Bogaerts et al. (2006), and also in the evaluation of approaches to screening for cancer, such as the PLCO Cancer Screening Trial (Prorok et al., 2000) and the National Lung Screening Trial (Gatsonis et al., 2011).

A basic DRCT design is the two-arm design (Lu and Gatsonis, 2013) shown in Figure 3.1. This design comprises two stages, namely the testing stage and the therapeutic intervention stage. In the first stage, patients are randomized to one of two diagnostic tests, and patients are then assigned to a therapeutic intervention in the second stage according to the test outcome. Unlike a typical diagnostic test accuracy study where the test result is compared against the "gold" standard reference test, carrying out the reference test is not necessary in a DRCT to meet the objective of comparing the effectiveness of two tests. However considering the resources required to conduct a DRCT, it is reasonable to assume that both the diagnostic tests and therapeutic interventions under study have been shown to provide promising results, or are standard of care approaches. In a variation to the two-arm DRCT, one of the arms can be designated as the control arm. In the control arm, all patients are either assigned one of the treatments, or randomized based on a known random allocation to different treatments. Such a design is also known as a marker-based strategy design, and it can be used for direct assessment of clinical usefulness of prognostic factors (Sargent et al., 2005).

The commonly used approach for statistical analysis of a two-arm DRCT is the intention to treat (ITT) approach. Randomization allows an investigator to make causal statements regarding the effect of a test strategy on patient outcomes. However even for a well designed and executed trial, it is likely that some patients may change their prescribed intervention after consultation with their health care provider, or choose not to follow the assigned intervention, thus introducing selection bias. In the pres-

ence of noncompliance, ITT analysis will only estimate the effect of randomization, not the causal effect of one treatment strategy compared to the other on patient outcomes. It has been shown that ITT analysis tends to bias the treatment effect towards the null (White, 2005). In screening trials, which typically have large sample size and low disease prevalence, even a small proportion of noncompliers will suffice in leading to inaccurate or even misleading estimates of screening exam efficacy (Gareen, 2007).

In order to estimate efficacy of treatment strategies when noncompliance is present, ITT analysis is unsuitable. Modern methods in causal inference have been used to address noncompliance in typical randomized clinical trials. Randomization-based efficacy estimators (White, 2005), or the instrumental variable approach (Imbens and Angrist, 1994, Angrist et al., 1996, Frangakis and Rubin, 2002), have been used to estimate the complier average causal effect or local average treatment effect. More sophisticated methods like the family of "g-methods" of Robins and his co-workers are also available to handle generalized treatment regimes g under complex longitudinal settings with time-varying treatments (Robins, 1986, 1987, 1989, 1992, 1994, 2000). One such method is the Structural Nested Mean Model (SNMM), which has been used for compliance adjustment in randomized studies as described in Robins (1994, 1998), Robins and Rotnitzky (2004), Vansteelandt and Goetghebeur (2003). The use of SNMM for compliance analysis under the scenario of exposure measurement errors was also examined by Goetghebeur and Stijn (2005). A good overview of the structural models is given in Vansteelandt and Joffe (2014).

In this chapter, we address the estimation of the causal effect of a test strategy on patient outcomes in the presence of noncompliance in a DRCT, i.e. the causal question of what would have been observed had, contrary to fact, all subjects remained on protocol. For the purpose of this study, we first focus on the two-arm design shown in Figure 3.1, and develop the approach to address the problem. We assume here

that noncompliance occurs in both stages of the 2-arm design, and the type of non-compliance actions considered here is switching of tests or therapeutic interventions. The performance of the causal estimates will be evaluated using various simulation scenarios. Based on this approach, we then extend it to a discordant pair design, Figure 3.3. The approach described here will be applied to a subset of the National Lung Screening Trial data (Aberle et al., 2011) to illustrate its use in estimating the causal effect of the test strategy in a DRCT design.

3.2 National Lung Screening Trial (NLST)

The NLST is a randomized multicenter study comparing low dose helical computed tomography (LDCT) with chest radiography (CXR) in the screening of current and former heavy smokers for early detection of lung cancer. The 53,454 participants enrolled were 55 to 74 years old, and had a history of at least 30 pack-years of smoking. Participants were randomly assigned to undergo annual screening using either LDCT (26,722 participants) or CXR (26,732 participants) for 3 years. The primary endpoint of the study was lung cancer mortality. At each screening examination, participants with positive screening results received follow-up recommendations for diagnostic evaluation, and information on the diagnostic evaluation performed was collected. A strict algorithm was applied to ascertain whether lung cancer was present at the time of screening. This algorithm was described in the supplementary appendix of Church et al. (2013). An endpoint verification process was used to verify deaths from lung cancer. The estimated sensitivity and specificity of the different modalities were 93.8% and 73.4% for LDCT, and 73.5% and 91.3% for CXR, respectively. Actual screening compliance at each of the three scheduled screens ranged from 98.5% to 92.9% in the LDCT arm, and 97.5% to 89.5% in the CXR arm. The compliance with diagnostic evaluation following a positive screen was not reported. The reported

median duration of follow-up was 6.5 years, and the maximum duration was 7.4 years in each group. Aberle et al. (2011) reported a relative reduction in mortality rate from lung cancer with LDCT screening of 20% (95% CI, 6.8 to 26.7; P = 0.004) at the completion of the study.

3.3 Notation

We will use the potential outcome framework for causal inference (Rubin, 1974) in the analysis. In this framework, let Y^a denote the potential outcome of a patient that would have been observed if the patient had received treatment a. Following the convention in statistics, upper case letters refer to random variables and lower case letters refer to the values realized by the random variables. For a dichotomous treatment, all patients will have 2 potential outcomes each, namely Y^0 and Y^1. However only one of the potential outcomes can be observed for each patient, and this problem is also known as the fundamental problem of causal inference (Holland, 1986). The observed outcome is denoted as Y. The causal effects are then defined as comparisons of potential outcomes Y^1 and Y^0 for the same patient. At a population level, the average causal effect is then defined as the expected value of the difference in potential outcomes over the population of interest, i.e. $E[Y^1 - Y^0]$ when Y is a continuous outcome.

With time varying or sequential treatments, we assume that measurements are collected at fixed time points $t_0, t_1, \ldots, t_K$. Let L_m and A_m denote respectively the auxiliary covariates measured and treatment received at time t_m for $m = 0, \ldots, K$, and not defined at t_{-1}. We further assume that at a particular time point, t_m, the variables are ordered temporally such that auxiliary covariate measurements (L_m) will precede treatment (A_m). We use overbar to denote the history of a variable, and underbar for

the future of a variable. As an example, the history of variable X up to time point t_m is $\bar{X}_m = \{X_0, X_1, \ldots, X_m\}$ for time points $m = 0, \ldots, K$, and the entire history of X is represented when subscript is omitted, i.e. $\bar{X} = \{X_0, X_1, \ldots, X_K\}$. Similarly, the future of X from t_m onward is $\underline{X}_m = \{X_m, X_{m+1}, \ldots, X_K\}$ for $m = 0, \ldots, K$. In the time varying or sequential treatments scenarios, the potential outcomes are denoted by $Y^{\bar{a}}$ where $\bar{a} = \{a_0, a_1, \ldots, a_K\}$. For ease of notation, the potential outcome with treatments set at 0 from time t_m onward $Y^{\bar{a}_{m-1}, \underline{a}_m = 0}$ will be written as $Y^{\bar{a}_{m-1}, \underline{0}}$ in the rest of the chapter.

The notation used in the two-arm DRCT design is shown in Figure 3.1. In the first stage, patients are randomized to one of two diagnostic tests ($Z = 0$ or 1), and the actual test performed on the patient is A_0. If patients adhere to the test assignment, then $A_0 = Z$. The test outcome (L_1) based on test A_0 is assumed to be binary, with values of 1 or 0. Typically in a diagnostic test, 1 is used to denote a "positive" test result (target condition present), and 0 to denote a "negative" test result (target condition absent). In the rest of the chapter, we will use the shorthand disease/non-disease to denote the presence/absence of the target condition. True binary disease status is D, with 1 being diseased and 0 being non-diseased. We will first assume that the target condition is ascertained, and then the case when the target condition is not ascertained is considered in Section 3.6. Therapeutic interventions are coded as 0 or 1, and patients are assigned to one of them based on the outcome of the test. Assuming that therapeutic intervention 1 is a more aggressive treatment, then patients who are indicated by the test to be diseased will be assigned therapeutic intervention 1, and 0 otherwise. By this design, the test outcome is indicative of therapeutic intervention assignment, and hence there is no further need to define a new variable for therapeutic intervention assignment. Intervention actually received by the patient is A_1, and under full compliance of intervention assignment, $A_1 = L_1$. We further denote patient outcome as Y, and the effect of a test strategy is essentially

a form of aggregation of the patient outcomes under the assigned test arm.

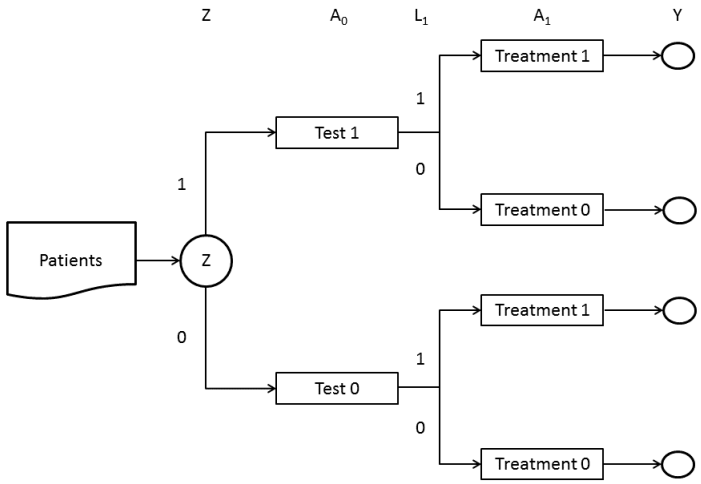


Figure 3.1: DRCT two-arm design

The connections between the notation in the two-arm DRCT design and the notation for potential outcomes in sequential treatments can be made when one considers $K = 1$, i.e. there are only two time points, t_0 and t_1, in the sequential treatments' notation. The auxiliary covariates at t_0 that constitute L_0 are Z and D in the two-arm DRCT design. At t_1, L_1 is the only auxiliary covariate in both the notation for DRCT and sequential treatments. The sequential treatments A_0 and A_1 are respectively test received, and therapeutic intervention received in the DRCT. Y^{a_0,a_1} will then denote the potential outcome that would be seen were the patient to receive test a_0, and therapeutic intervention a_1. For dichotomous test and therapeutic intervention options, each patient will have four potential outcomes. Recall that in the DRCT scenario that we are considering here, the patient's therapeutic intervention is based on the test received, and the corresponding test outcome. Therapeutic intervention is not intervened independently of the test. To avoid confusion with the conventional notion in potential outcome framework that treatment interventions can be manipulated independently, we use $Y^{a_0,a_1=L_1(a_0)}$ to remind ourselves that the therapeutic

intervention is based on the test, and the corresponding test outcome. Although the potential outcomes can also be represented as Y^{a_0} *under full compliance* to therapeutic intervention assignment made by the test result, we feel that this notation does not convey the information that the effect of the test is mediated by the therapeutic intervention received.

Thus in this analysis, the causal estimand that we are interested in for a continuous Y is represented by

$$\Delta = E\left[Y^{a_0=1,a_1=L_1(a_0=1)} - Y^{a_0=0,a_1=L_1(a_0=0)}\right]$$

3.4 Structural Nested Mean Model (SNMM)

In the two-arm DRCT design, patients are effectively randomized twice when the tests are not perfect. The first randomization occurs at the test assignment phase with known probability of assignment as defined in the trial design, i.e. a marginal randomization process. The second randomization is a conditional randomization that occurs with the result of the test. The test outcome can be viewed as a randomization process conditioned on the disease status of the patient. Diseased patients are allocated to test outcome $L_1 = 1$ with probability equal to the test sensitivity, and outcome $L_1 = 0$ with probability of $1 -$ sensitivity. Likewise, non-diseased patients are randomized to $L_1 = 0$ and $L_1 = 1$ with probability equal to the test specificity and $1 -$ specificity, respectively. From this vantage point, the two-arm DRCT design has strong resemblance to a sequential randomization trial, and this provides a starting point for addressing the question in this chapter.

In a Structural Nested Mean Model (Robins, 1994), the effect of a treatment at t_m on the subsequent outcome mean when holding all future treatments fixed at their

reference level 0 is

$$E[Y^{\bar{a}_{m-1},a_m,\underline{0}} - Y^{\bar{a}_{m-1},\underline{0}}|\bar{L}_m = \bar{\ell}_m, \bar{A}_m = \bar{a}_m] = \gamma_m(\bar{\ell}_m, \bar{a}_m; \psi) = A_m X^T \psi$$

By convention, the parameterization of $\gamma_m(\bar{\ell}_m, \bar{a}_m; \psi)$ is such that $\gamma_m(\bar{\ell}_m, \bar{a}_m; 0) = 0$ represents the case of no treatment effect. It is further assumed here that the effect is linear in ψ, and X is the covariate vector consisting of $\bar{L}_m$ and $\bar{A}_{m-1}$. The following assumptions are needed for estimation:

1. Stable unit treatment value assumption (SUTVA) (Rubin, 1980). SUTVA is the apriori assumption that the value of Y for patient i when exposed to treatment a will be the same regardless of the mechanism used to assign the treatment, and regardless of the treatments other patients received.

2. Consistency. For a given patient with treatment history $\bar{A}$, then $Y^{\bar{a}=\bar{A}} = Y$ for that patient.

3. Sequential ignorability: $A_m \perp\!\!\!\perp Y^{\bar{a}_{m-1},\underline{0}}|\bar{L}_m,\ \bar{A}_{m-1} = \bar{a}_{m-1}$ for $m = 0, \ldots, K$. This assumption implies that at each time t_m, the observed history of covariates L_m and treatments A_{m-1} includes all risk factors of A_m that are associated with the outcome, which has all future treatments from t_m onward held at the reference level 0.

We further define

$$H_m(\psi) = Y - \sum_{l=m}^{K} \gamma_l(\bar{\ell}_l, \bar{a}_l; \psi)$$

such that the expectation of $H_m(\psi)$ equals the expected outcome that would have been observed if treatments were suspended from t_m onward, i.e. $E[H_m(\psi)|\bar{L}_m, \bar{A}_{m-1}] = E[Y^{\bar{a}_{m-1},\underline{0}}|\bar{L}_m, \bar{A}_{m-1}]$ (See Appendix 3.A for the derivation).

From the sequential ignorability assumption, an estimator for ψ can be obtained by

solving the estimating equations

$$\sum_{i=1}^{n}\sum_{m=0}^{K}\begin{bmatrix}\left(\upsilon_{i,m}-E[\upsilon_{i,m}|\bar{L}_{i,m},\bar{A}_{i,m-1}]\right)\left(H_{i,m}-E[H_{i,m}|\bar{L}_{i,m},\bar{A}_{i,m-1}]\right)\\ \Upsilon_{i,m}\left(H_{i,m}-E[H_{i,m}|\bar{L}_{i,m},\bar{A}_{i,m-1}]\right)\end{bmatrix}=\begin{pmatrix}0\\0\end{pmatrix}$$

where

$$E[H_{i,m}|\bar{L}_{i,m},\bar{A}_{i,m-1}]=\Upsilon_{i,m}^{T}\xi$$
$$\upsilon_{i,m}(\bar{L}_{i,m},\bar{A}_{i,m})=E\left[\frac{\partial H_{i,m}(\psi)}{\partial\psi}\middle|\,\bar{L}_{i,m},\bar{A}_{i,m}\right]$$

for n patients in the study. In the above formulation, $E[H_{i,m}|\bar{L}_{i,m},\bar{A}_{i,m-1}]$ is linear in ξ, and $\Upsilon_{i,m}$ is a vector of covariates for subject i at t_m. When the previous outcome is included in the confounder history, and the conditional variance of $H_{i,m}$ given $\bar{L}_{i,m},\bar{A}_{i,m}$ is constant, local semiparametric efficiency is attained using the above definition for $\upsilon_{i,m}$ (Vansteelandt and Joffe, 2014). This estimator also has the property of being doubly robust, in the sense that as long as at least one of the specified models for $E[A_m|\bar{L}_m,\bar{A}_{m-1}]$ or $E[H_m|\bar{L}_m,\bar{A}_{m-1}]$ is correct, the first row of the estimating equation has mean 0 at $\hat{\psi}=\psi$. Solving these estimating equations result in the following expression for the estimators of ψ and ξ:

$$\begin{bmatrix}\hat{\psi}\\ \hat{\xi}\end{bmatrix}=\left\{\sum_{i=1}^{n}\sum_{m=0}^{K}\begin{bmatrix}(A_{i,m}-E[A_{i,m}|\bar{L}_{i,m},\bar{A}_{i,m-1}])X_{i,m}\\ \Upsilon_{i,m}\end{bmatrix}\begin{bmatrix}\sum_{j=m}^{1}A_{i,j}X_{i,j}^{T} & \Upsilon_{i,m}^{T}\end{bmatrix}\right\}^{-1}$$
$$\left\{\sum_{i=1}^{N}\sum_{m=0}^{K}Y_i\begin{bmatrix}(A_{i,m}-E[A_{i,m}|\bar{L}_{i,m},\bar{A}_{i,m-1}])X_{i,m}\\ \Upsilon_{i,m}\end{bmatrix}\right\}$$

Variance of the estimators can be estimated using bootstrap samples. With the esti-

mator $\hat{\psi}$, the estimate for $E[Y^{\bar{a}}]$ can then be obtained using a Monte Carlo algorithm to capture the dynamic nature of the assignment process.

So far we have used the identity link SNMM. More generally, the SNMM can be written as $g\{E[Y^{\bar{a}_{m-1},a_m,\underline{0}}|\bar{L}_m = \ell_m, \bar{A}_m = \bar{a}_m]\} - g\{E[Y^{\bar{a}_{m-1},\underline{0}}|\bar{L}_m = \ell_m, \bar{A}_m = \bar{a}_m]\} = \gamma_m(\bar{\ell}_m, \bar{a}_m; \psi)$, where $g(\cdot)$ is a known link function like identity, log, logit, or probit.

Unlike additive and multiplicative structural mean models for continuous and count outcomes, solving logit or probit link SNMM for a dichotomous outcome is known to be problematic, and it has been shown that no unbiased estimating equations exist (Robins and Rotnitzky, 2004). When the outcome is dichotomous, the log link or identity link functions can possibly be used, but keeping in mind that these models will not guarantee that the predicted response probabilities are within the interval of $[0, 1]$.

3.5 Simulation

3.5.1 Setup & Analysis

In this section, we describe the scenarios used to simulate a two-arm DRCT design as shown in Figure 3.1. The two hypothetical tests considered are Test 1 vs. Test 0, followed by the appropriate therapeutic interventions, which in this case are Treatment 1 if tested positive, or Treatment 0 if tested negative.

Disease status is assumed to be the only confounder, and no other covariates are simulated here for simplicity. Baseline covariate L_0 consists of only D and Z, where D is the disease status, Z is the test assignment, and the only measurement at t_1 is L_1, the test outcome. A_0 is the actual test received and A_1 is the actual therapeutic intervention received. Y is a continuous variable representing the log survival time of

the patient.

The specific details of the simulation are given in Appendix 3.B.1. Here, we give a brief outline of the simulation setup. Each patient's potential outcomes are generated from a bivariate normal distribution with mean given by $\mu = E[Y^{a_1}|D] = \log 5 - 0.4D - 0.1a_1 + 0.6Da_1$, depending on the disease condition of the patient. This design also implies no direct effect of test received (a_0) on the patient's outcome, i.e. all effects are mediated by the therapeutic intervention.

Disease prevalence, $\pi = Pr(D = 1)$, is assumed to be 0.3. Two different scenarios are considered for the diagnostic performances of the two tests, Table 3.1. In the first scenario, Test 1 is better in both sensitivity and specificity. This presents a straight forward scenario where one modality is clearly better than the other in terms of performance. For the second scenario, the sensitivity of the two tests are the same, but the specificity of Test 1 is slightly better than Test 0. This presents a more challenging scenario where the separation between the performance of the 2 modalities is a lot closer. These scenarios are based on the study performed by Deserno et al. (2004) on preoperative nodal staging of patients with urinary bladder cancer using enhanced vs. conventional MRI. Additional scenarios are included in Table 3.9 in Appendix 3.B.2, including the scenario that uses the diagnostic performance of the two tests as reported by Deserno et al. (2004).

Table 3.1: Diagnostic performance scenarios

Scenario	Test 1		Test 0	
	$Sens_1$	$Spec_1$	$Sens_0$	$Spec_0$
1	0.96	0.95	0.86	0.89
2	0.96	0.99	0.96	0.95

The causal effect of treatment strategy, Δ, can be computed by defining

$$\begin{aligned} EY^{a_0=1,a_1=L_1(a_0=1)} &= \mu_1 Sens_1 \pi + \mu_2(1 - Sens_1)\pi \\ &\quad + \mu_3(1 - Spec_1)(1 - \pi) + \mu_4 Spec_1(1 - \pi) \\ EY^{a_0=0,a_1=L_1(a_0=0)} &= \mu_1 Sens_0 \pi + \mu_2(1 - Sens_0)\pi \\ &\quad + \mu_3(1 - Spec_0)(1 - \pi) + \mu_4 Spec_0(1 - \pi) \end{aligned}$$

where $\mu_1 = E[Y^{a_1=1}|D = 1]$, $\mu_2 = E[Y^{a_1=0}|D = 1]$, $\mu_3 = E[Y^{a_1=1}|D = 0]$, and $\mu_4 = E[Y^{a_1=0}|D = 0]$.

Probabilities of compliance to test and treatment assignments are modeled using a logit model, with assignment and disease condition as covariates. Compliance probability arising from this model is as high as 95% for patients who are diseased and assigned to either Test 1 or Treatment 1, and as low as 50% for diseased patients assigned to either Test 0 or Treatment 0. Unless stated otherwise, 1000 simulations were performed with each simulation involving 1000 bootstrap iterations and $n = 10000$ simulated patients. Coverage is ascertained using the bootstrap percentile intervals method.

In this two-arm DRCT design, the index for the time points are $m = 0, 1$ with $K = 1$. For the analysis, models for $\gamma_m(D, \bar{A}_m; \psi)$, $E[H_m(\psi)|\bar{L}_m, \bar{A}_{m-1}]$, and $E[A_m|\bar{L}_m, \bar{A}_{m-1}]$ are specified as shown in Appendix 3.B.2. Estimation of $E[Y^{a_0,a_1=L_1(a_0)}]$ is obtained via Monte Carlo simulation for $a_0 = \{0, 1\}$. Estimates for the variance of $\hat{\Delta}$, i.e. $\hat{\sigma}^2_{\hat{\Delta}}$, is obtained from the bootstrap samples.

For dichotomous outcome (5-year mortality risk), the corresponding causal estimand of interest is

$$\Delta_{bin} = \frac{E[Y_{bin}^{a_0=1,a_1=L_1(a_0=1)}]}{E[Y_{bin}^{a_0=0,a_1=L_1(a_0=0)}]}$$

We use the same simulation setup as before, and define the dichotomous outcome as

$Y_{bin} = 1$ if $Y \leq \log 5$ and $Y_{bin} = 0$ otherwise. The identity link SNMM is used to estimate $E[Y^{a_0=1,a_1=L_1(a_0=1)}]$ and $E[Y^{a_0=0,a_1=L_1(a_0=0)}]$. An alternative approach is to use log link SNMM (Picciotto et al., 2012). The implementation details for log link SNMM, and its simulation results are included in Appendix 3.C.

3.5.2 Results

The results from the simulation for continuous outcome are given in Table 3.2. Results based on the ITT approach are also included for comparison purposes.

Table 3.2: Results from simulation scenarios. $Bias = \hat{\Delta} - \Delta$, "MSE" is the mean squared error, and "Cvg" refers to the coverage of the 95% Confidence Interval.

		SNMM				ITT			
SN	Δ	$\hat{\Delta}_{SNMM}$	Bias	MSE	Cvg	$\hat{\Delta}_{ITT}$	Bias	MSE	Cvg
1	0.0192	0.0191	-0.0001	0.0001	0.954	0.0048	-0.0144	0.0002	0.213
2	0.0028	0.0025	-0.0003	0.0001	0.953	0.0015	-0.0013	0.0001	0.936

As expected, the presence of noncompliance leads to biased estimates when using the ITT approach, but the estimates from SNMM remain unbiased and have smaller MSE with reasonable 95% Confidence Interval coverage. The coverage of the ITT's 95% Confidence Interval in Scenario 1 is very poor (21.3%). However when the actual causal effect is close to null (Scenario 2), ITT's coverage recovered to a reasonable level.

Table 3.3: Results from simulation scenarios. $Bias = \hat{\Delta}_{bin} - \Delta_{bin}$, "MSE" is the mean squared error, and "Cvg" refers to the coverage of the 95% Confidence Interval.

		SNMM				ITT			
SN	Δ_{bin}	$\hat{\Delta}_{SNMM}$	Bias	MSE	Cvg	$\hat{\Delta}_{ITT}$	Bias	MSE	Cvg
1	0.9421	0.9410	-0.0011	0.0016	0.933	0.9880	0.0459	0.0023	0.111
2	0.9883	0.9903	0.0021	0.0021	0.927	0.9928	0.0046	0.0001	1.000

For the simulation scenarios with binary outcomes, the results are tabulated in Table 3.3. Similar to the previous scenarios, the estimates from SNMM are unbiased

while those from ITT are biased towards the null. The SNMM's coverage of the 95% Confidence Interval remains reasonable for the binary outcome cases. In contrast, the ITT's coverage degrades rapidly when the true effect moves further away from the null.

As mentioned in the previous section, results for the alternative log link SNMM scenarios are given in Appendix 3.C.

3.6 Disease Condition Not Ascertained in Study

In this section, we consider the situation when the DRCT study did not perform "gold" standard reference test to ascertain the disease condition D. As we have seen so far, D is an important confounder and it is needed to provide an unbiased causal estimate. To address this, we propose a multiple imputation procedure to impute the missing disease condition based on the posterior distribution of D. We use the same simulation scenario described in Section 3.5.1, but with disease condition (D) not ascertained, to assess the performance of this multiple imputation procedure.

The approach to this problem is to make use of the test outcome, L_1, which is a measure of the patient's disease status, but with an error or misclassification rate determined by the accuracy of the test. Similar to Ogburn and VanderWeele (2012), we assume that $L_1 \perp\!\!\!\perp A_1, Y^{a_0,a_1} | D, A_0$. One could then impute the missing disease status based on an appropriate imputation model. While test accuracies are assumed to be known in the model below, suitable priors can be incorporated to allow for estimated test accuracies. C_0 and C_1 are the compliance indicators at the test and therapeutic intervention stages respectively, as defined in Appendix 3.B.1.

The imputation approach is based on the posterior distribution of D given the ob-

served variables,

$$p(D|Y, A_1, C_1, L_1, A_0, C_0, Z) \propto p(Y|A_1, D)p(C_1|L_1, D)p(L_1|A_0, D)p(C_0|Z, D)p(D)$$

$$\begin{aligned}
Y_i|A_{i,1}, D &\sim N(\mu_{y_i}, \sigma_y^2) \\
\mu_{y_i} &= \beta_1 + \beta_2 A_{i,1} + \beta_3 D_i + \beta_4 A_{i,1} D_i \\
C_{i,1}|L_1, D &\sim Bernoulli(\mu_{c_{i,1}}) \\
\text{logit}(\mu_{c_{i,1}}) &= \alpha_1 + \alpha_2 D_i + \alpha_3 L_{i,1} + \alpha_4 D_i L_{i,1} \\
L_{i,1}|A_0, D &\sim Bernoulli(\mu_{l_i}) \\
\mu_{l_i} &= (1 - A_{i,0})\{Sens_0 D_i + (1 - Spec_0)(1 - D_i)\} \\
&\quad + A_{i,0}\{Sens_1 D_i + (1 - Spec_1) * (1 - D_i)\} \\
C_{i,0}|Z, D &\sim Bernoulli(\mu_{c_{i,0}}) \\
\text{logit}(\mu_{c_{i,0}}) &= \alpha_1 + \alpha_2 D_i + \alpha_3 Z_i + \alpha_4 D_i Z_i \\
D_i &\sim Bernoulli(\mu_d)
\end{aligned}$$

Priors specified are

$$\begin{aligned}
&\mu_d \sim Unif(0, 1),\ \sigma_y \sim Unif(0, 5) \\
&\alpha_1 \sim N(0, 100^2),\ \alpha_2 \sim N(0, 100^2),\ \alpha_3 \sim N(0, 100^2),\ \alpha_4 \sim N(0, 100^2) \\
&\beta_1 \sim N(0, 100^2),\ \beta_2 \sim N(0, 100^2),\ \beta_3 \sim N(0, 100^2),\ \beta_4 \sim N(0, 100^2)
\end{aligned}$$

The procedure that will provide unbiased estimate and with appropriate coverage is illustrated in Figure 3.2. Coverage is assessed using bootstrap percentile intervals.

The posterior distributions of the parameters for a typical simulation sample are summarized in Table 3.4 for the scenario where $Sens_1 = 0.96$, $Spec_1 = 0.95$, $Sens_0 = 0.86$,

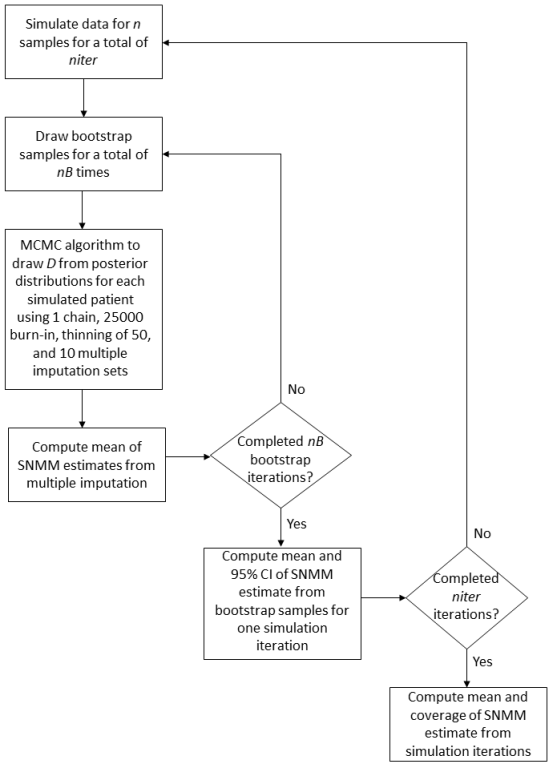


Figure 3.2: Flow chart for multiple imputation procedure

and $Spec_0 = 0.89$. From the simulations, the estimate from SNMM for this scenario is 0.0202, and the corresponding bias and MSE are 0.001 and 0.0001 respectively. The coverage attained is 0.93. The performance of ITT for this scenario will be similar to the results shown in Table 3.2. The results are based on 100 simulation iterations ($niter$), of which each iteration involves 200 bootstrap samples (nB), and 10 multiple imputation sets within each bootstrap sample.

Table 3.4: Summary Statistics for Posterior Distributions in the Bayesian approach

	Actual	Lower95	Median	Upper95	Mean	SD	psrf
μ_d	0.300	0.291	0.302	0.311	0.302	0.005	1.000
σ_y	0.200	0.198	0.201	0.204	0.201	0.002	1.001
α_1	1.700	1.602	1.658	1.713	1.658	0.028	1.002
α_2	-1.700	-1.726	-1.603	-1.481	-1.603	0.063	1.003
α_3	0.400	0.279	0.410	0.535	0.410	0.066	1.002
α_4	-2.500	-2.709	-2.507	-2.295	-2.507	0.106	1.004
β_1	1.609	1.604	1.610	1.615	1.610	0.003	1.001
β_2	-0.100	-0.112	-0.100	-0.087	-0.100	0.006	1.001
β_3	-0.400	-0.410	-0.400	-0.390	-0.400	0.005	1.001
β_4	0.600	0.562	0.592	0.625	0.592	0.016	1.001

3.7 Extension to Discordant Pair Design

From a statistical perspective, a more efficient DRCT design is the discordant pair design (Lu and Gatsonis, 2013). This is the design adopted in the MINDACT trial (Bogaerts et al., 2006). In the discordant pair design, Figure 3.3, all patients undergo both tests, i.e. $A_0 = 0$ and $A_0 = 1$. When the results of both tests agree, the patient undergoes Treatment 1 ($A_1 = 1$) if the tests are positive ($L_1 = 1$ for both tests), or Treatment 0 ($A_1 = 0$) if the tests are negative ($L_1 = 0$ for both tests). If the results disagree, the patient is randomized to follow the result of either Test 1 or 0.

With full compliance to the test and treatment assignments, the causal estimate of the effect of one test strategy versus another, Δ, can be obtained by $\Delta = \Delta_{disc} f$, where Δ_{disc} is the difference in effect between those who followed Test 1 vs. those who followed Test 0 in the discordant population, and $f = Pr(discordant)$ is the discordant rate.

Under the discordant design, every patient will undergo both tests and full compliance is assumed at the *test* assignment/receipt stage. We assume that noncompliance occurs only in the therapeutic intervention (treatment) assignment/receipt stage. The objective of the analysis is still to estimate Δ, the causal effect of treatment strategy

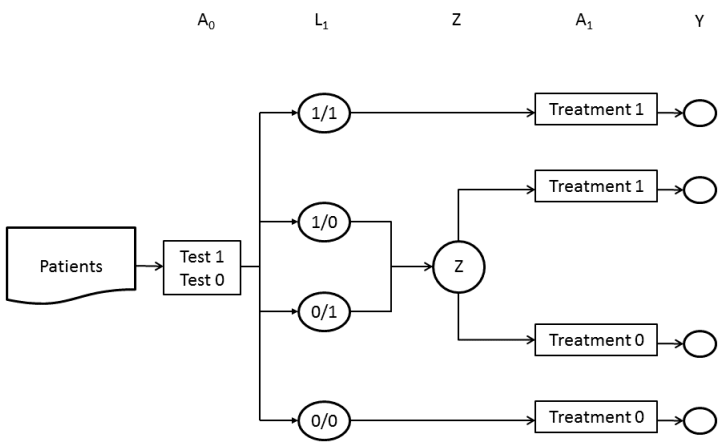


Figure 3.3: DRCT discordant design

based on test $A_0 = 1$ versus test $A_0 = 0$.

The simulation setup from the previous section is used for the discordant pair design. In this simulation, we consider the scenarios where outcome Y is continuous. We further assume for the simulation that the patients are blinded to the results of both tests, and are only informed of the therapeutic intervention assignment. In this case, the SNMM approach for the discordant pair design is then reduced to the simpler case of estimating the effect under different strata of A_0 separately. Instead of the two stages considered in the 2-arm design, the discordant design will only require modeling of a single stage, which is the therapeutic intervention assignment/receipt stage.

The results from the simulation of the discordant design, based on similar simulation set-up as in the 2-arm design, are given in Table 3.5. Here we observe that the performance of SNMM is better than in the 2-arm design (Table 3.2). This is likely due to the fact that only a single stage of noncompliance adjustment is required in the discordant pair design. Similar to the two-arm design, the performance of the ITT estimate is worse when the true effect is away from the null.

Table 3.5: Results from simulation scenarios for discordant pair design. $Bias = \hat{\Delta} - \Delta$, "MSE" is the mean squared error, and "Cvg" refers to the coverage of the 95% Confidence Interval. Note that MSE=0.0000 refers to $< 1 \times 10^{-4}$.

		SNMM				ITT			
SN	Δ_{bin}	$\hat{\Delta}_{SNMM}$	Bias	MSE	Cvg	$\hat{\Delta}_{ITT}$	Bias	MSE	Cvg
1	0.0192	0.0192	0.0000	0.0000	0.945	-0.0026	-0.0218	0.0005	0.000
2	0.0028	0.0028	0.0000	0.0000	0.945	0.0021	-0.0007	0.0000	0.915

3.8 Illustrative Example

In this section, we apply the approach for the two-arm DRCT described in this chapter to the NLST data, as described earlier in Section 3.2. We will only use data from the first screen and ignore the subsequent screens. In this way, the NLST design resembles that of a two-arm DRCT design. Since disease conditions of the patients prior to the baseline screen were ascertained in the NLST data, disease condition is included as one of the baseline covariates in the SNMM.

Data from a subset of the NLST participants were extracted, and used for the analysis here. This subset consists of 18314 participants from the ACRIN Centers that participated in this study. The summary statistics for this subset are tabulated in Table 3.6.

Table 3.6: Summary statistics based on subset of 18314 participants, and actual test received.

Modality	Sensitivity	Specificity	Prevalence	5-year Mortality
CXR	0.697	0.922	0.008	0.0453
LDCT	0.908	0.770	0.012	0.0417

For the purpose of this illustrative example, compliance to test assignment is defined as following randomization assignment to either LDCT or CXR. Compliance to therapeutic intervention is defined as following recommended guidelines according to test result, as reported by participants during follow-up over a period of one year from screening date. For participants with positive screening results, compliance to

therapeutic intervention refers to adhering to follow-up recommendations for diagnostic evaluation while noncompliance refers to not pursuing any follow-ups related to lung cancer. On the other hand, compliance to therapeutic intervention for participants with negative screening results refers to not pursuing any follow-ups related to lung cancer. Follow-up recommendations for diagnostic evaluation can entail multiple versions of treatment. For the intent and purpose here, we assume treatment variation irrelevance (VanderWeele, 2009) for the follow-up treatments on lung cancer, i.e. the different versions of treatment have the same effect. Among participants who had lung cancer at screening, compliance to test and therapeutic intervention assignments based on the definitions used in this example were 100% and 50.3% respectively. For the remaining participants, compliance to test and therapeutic intervention assignments were 99.8% and 82.7% respectively.

In this example, we are interested in estimating the causal effect of a test strategy based on LDCT on patient outcome (5-year mortality from lung cancer) if, contrary to fact, all subjects had remained on protocol. In mathematical notation, this causal effect is defined as $\Delta_{bin} = E[Y_{bin}^{a_0=1,a_1=L_1(a_0=1)}]/E[Y_{bin}^{a_0=0,a_1=L_1(a_0=0)}]$. Using the models described in Appendix 3.B.2 for the two-arm DRCT, the estimates and corresponding 95% Confidence Interval based on SNMM approach with 100000 bootstrap iterations are given in Table 3.7. For comparison purposes, the results from ITT and PP (Per Protocol) analyses are also included in the table. In the PP analysis, only participants who complied with both test and therapeutic intervention assignments are included.

Table 3.7: Analyses results for the causal estimate of the relative risk of 5-year mortality for a test strategy based on LDCT.

Methods	$\hat{\Delta}$	95% CI
SNMM	0.794	[0.566, 1.108]
ITT	0.923	[0.801, 1.057]
PP	0.948	[0.812, 1.108]

In the presence of noncompliance, the ITT estimator is an estimate of the effect of

randomization while the PP estimator is relevant to a subpopulation that cannot be identified a priori. From the behavior of the ITT estimator observed in the simulation, the estimate is likely to be biased towards the null, and the coverage of the 95% Confidence Interval is likely to be very poor. The PP estimator is also a known biased estimator of the causal estimand of interest, and here it showed an effect even closer to the null. On the other hand, the SNMM estimator shows a stronger effect in the LDCT strategy compared to CXR strategy. This result is not unreasonable considering that the final reported result based on the entire study population of 53,454 participants was a relative reduction in mortality rate of 20% (Aberle et al., 2011). The estimate from SNMM is also consistent with the findings from the systematic review reported in Humphrey et al. (2013). While the SNMM has a lower point estimate for the relative risk, the SNMM's confidence interval is wider than the other 2 approaches. The wider confidence interval is expected considering the uncertainty involved in estimating the potential outcomes. The confidence interval for SNMM is computed using bootstrap percentile intervals method. More accurate bootstrap confidence intervals, as well as larger bootstrap samples, can also be used to improve the confidence interval estimation.

3.9 Discussion

In this chapter, we set out to address the question of obtaining a causal estimate in a DRCT design when noncompliance is present, i.e. the causal question of what would have been observed had, contrary to fact, all subjects remained on protocol. The insight obtained from noting that the test is in effect a conditional randomizer, allowed us to use methods from the sequential randomization literature, specifically SNMM. The SNMM methodology has the flexibility to be adapted to the different DRCT designs that exist in the literature today. In this chapter, we have laid out

the general approach for using SNMM in a two-arm design, and in a discordant pair design. Through simulations, we have established the viability of SNMM in providing an unbiased estimate of the causal estimand of interest in DRCT designs under a set of assumptions. While this has provided the statistician with a tool to potentially overcome noncompliance in DRCT, it does not replace the need to have good clinical trial planning, and execution. The simulations also highlighted the deficiencies in the commonly used ITT approach in the presence of noncompliance in DRCT studies. Not only were the ITT estimates biased, the 95% Confidence Interval coverage was found to be as low as 11.1%.

As in most causal inference problems, the untestable assumption of ignorability requires great care and subject matter knowledge. Sensitivity analysis should be performed to determine how sensitive the analyses are to unmeasured confounder. Major limitations of the SNMM are the difficulties associated with logit SNMM, and the computationally intensive process for SNMM that are not based on a linear model with identity link.

Various extensions to the current work would be important to pursue. The two-arm DRCT design can be extended to include multiple stages, similar to the design used in the NLST. However in this case, one would need to consider the effects of prior tests and therapeutic interventions on later tests and interventions. When trials are conducted over a prolonged period of time, one would also need to consider the possibility that disease condition may change over time, and account for such a time varying confounder in SNMM. When the patient outcome of interest is quality of life related, then one can no longer make the assumption that the test has no direct effect on the patient's outcome as it may have an effect on the patient's perceived state of well-being. Similarly if the test is a prognostic/predictive biomarker, direct effect should be included. Last but not least, we have assumed in the discordant pair

design that patients are blinded to the results from both tests. This assumption can be relaxed to allow generalizability.

The emphasis on patient centered outcomes research in recent years has expanded the scope of the study of diagnostic tests to include the impact on subsequent care and patient outcomes. However, the practical challenges in carrying out randomized studies of diagnostic tests in this context and the emphasis on real world settings have made evaluation of diagnostic tests using observational data increasingly attractive. There is a strong similarity in the analytical approach between observational data and randomized trial with noncompliance. The results from this study will therefore contribute to the evaluation of diagnostic tests using observational data, and the needs of comparative effectiveness research.

3.A Derivation of $E[H_m(\psi)|\bar{L}_m, \bar{A}_{m-1}]$

Without loss of generality, assume that $K = 1$, then

$$\begin{aligned}
H_1(\psi) &= Y - \gamma_1(\bar{\ell}_1, \bar{a}_1; \psi) \\
E[H_1(\psi)|\bar{L}_1, \bar{A}_1] &= E[Y|\bar{L}_1, \bar{A}_1] - \gamma_1(\bar{\ell}_1, \bar{a}_1; \psi) \\
&= E[Y^{\bar{a}}|\bar{L}_1, \bar{A}_1] - \gamma_1(\bar{\ell}_1, \bar{a}_1; \psi) \\
&= E[Y^{a_0,0}|\bar{L}_1, \bar{A}_1] \\
&= E[Y^{a_0,0}|\bar{L}_1, A_0] \\
&= E[H_1(\psi)|\bar{L}_1, A_0]
\end{aligned}$$

$$
\begin{aligned}
H_0(\psi) &= Y - \gamma_1(\bar{\ell}_1, \bar{a}_1; \psi) - \gamma_0(\ell_0, a_0; \psi) \\
E[H_0(\psi)|L_0, A_0] &= E[Y - \gamma_1(\bar{\ell}_1, \bar{a}_1; \psi) - \gamma_0(\ell_0, a_0; \psi)|L_0, A_0] \\
&= E[Y - \gamma_1(\bar{\ell}_1, \bar{a}_1; \psi)|L_0, A_0] - \gamma_0(\ell_0, a_0; \psi) \\
&= E\left[E[Y - \gamma_1(\bar{\ell}_1, \bar{a}_1; \psi)|\bar{L}_1, \bar{A}_1]|L_0, A_0\right] - \gamma_0(\ell_0, a_0; \psi) \\
&= E[Y^{a_0,0}|L_0, A_0] - \gamma_0(\ell_0, a_0; \psi) \\
&= E[Y^{0,0}|L_0, A_0] \\
&= E[Y^{0,0}|L_0] \\
&= E[H_0(\psi)|L_0]
\end{aligned}
$$

3.B Main Simulation

3.B.1 Setup

Specific details for the simulation are as follows:

- Patient's potential outcomes, $E[Y^{a_1}|D] = \log 5 - 2\sigma D - 0.5\sigma a_1 + 3\sigma D a_1$ with $\sigma = 0.2$. Here we have assumed that test has no direct effect on the outcome. Outcomes are simulated as bivariate normal given disease condition with means,

$$\mu_D = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} E[Y^{a_1=1}|D=1] \\ E[Y^{a_1=0}|D=1] \end{pmatrix} = \begin{pmatrix} 1.71 \\ 1.21 \end{pmatrix}$$

and

$$\mu_{\bar{D}} = \begin{pmatrix} \mu_3 \\ \mu_4 \end{pmatrix} = \begin{pmatrix} E[Y^{a_1=1}|D=0] \\ E[Y^{a_1=0}|D=0] \end{pmatrix} = \begin{pmatrix} 1.51 \\ 1.61 \end{pmatrix}$$

and covariance matrices

$$\Sigma_D = \begin{bmatrix} \sigma^2 & -0.5\sigma^2 \\ -0.5\sigma^2 & \sigma^2 \end{bmatrix}$$

$$\Sigma_{\bar{D}} = \begin{bmatrix} \sigma^2 & 0.8\sigma^2 \\ 0.8\sigma^2 & \sigma^2 \end{bmatrix}$$

- Probability of Disease: $\pi = Pr(D = 1) = 0.3$
- Causal Effect of Strategies, $\Delta = EY^{a_0=1,a_1=L_1(a_0=1)} - EY^{a_0=0,a_1=L_1(a_0=0)}$, where

$$EY^{1,\ell_1} = \mu_1 Sens_1 \pi + \mu_2(1 - Sens_1)\pi + \mu_3(1 - Spec_1)(1 - \pi) + \mu_4 Spec_1(1 - \pi)$$

$$EY^{0,\ell_1} = \mu_1 Sens_0 \pi + \mu_2(1 - Sens_0)\pi + \mu_3(1 - Spec_0)(1 - \pi) + \mu_4 Spec_0(1 - \pi)$$

- The models for probability of compliance are:

$$\text{logit}[Pr(C_0 = 1|D, Z)] = 1.7 - 1.7D + 0.4Z + 2.5ZD$$

$$\text{logit}[Pr(C_1 = 1|D, L_1)] = 1.7 - 1.7D + 0.4L_1 + 2.5L_1D$$

The above model will result in the compliance distribution shown in Table 3.8.

Table 3.8: Simulated compliance distribution

	$Z = 0$	$Z = 1$
$D = 0$	0.846	0.891
$D = 1$	0.5	0.948

- Number of subjects, $n = 10,000$
- Number of bootstrap, $B = 1000$
- Number of simulations, $Nsim = 1000$

3.B.2 SNMM

In this two-arm DRCT design, the index for the time points are $m = 0, 1$ with $K = 1$. The models used in the SNMM approach are described below with the subscript i for the covariates dropped for ease of notation.

- At $m = 1$,

$$\gamma_1(D, A_1, A_0; \psi) = \psi_0 A_1 + \psi_1 A_1 D + \psi_2 A_1 A_0 + \psi_3 A_1 D A_0$$

and at $m = 0$,

$$\gamma_0(D, A_0; \psi) = \psi_4 A_0 + \psi_5 A_0 D$$

The coefficients can be written as

$$\psi = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \\ \psi_5 \end{pmatrix}$$

- $E[H_m|\bar{L}_m, \bar{A}_{m-1}] = \xi_0 + \xi_1 D + \xi_2 A_0 + \xi_3 D A_0 = \Upsilon\xi$
- $E[A_m|\bar{L}_m, \bar{A}_{m-1}]$ are modeled using parametric models

$$\text{logit}Pr(A_0|D, Z) = \omega_{00} + \omega_{01} D + \omega_{02} Z + \omega_{03} DZ$$
$$\text{logit}Pr(A_1|D, L_1) = \omega_{10} + \omega_{11} D + \omega_{12} L_1 + \omega_{13} DL_1$$

The parameters ω are estimated using maximum likelihood estimators.

- Estimation of $E[Y^{a_0, a_1 = L_1(a_0)}]$ is obtained via Monte Carlo simulation.
  1. Compute $\hat{E}[Y^{0,0}] = \frac{1}{n} \sum_{i=1}^{n} \{Y - \sum_{l=0}^{1} \gamma_l(\bar{\ell}_l, \bar{a}_l; \hat{\psi})\}$
  2. $\hat{\mu}_D = \frac{1}{n} \sum_{i=1}^{n} D_i$
  3. $\text{logit}(p_{\ell_1}) = \omega_{l_1,0} + \omega_{l_1,1} D + \omega_{l_1,2} A_0 + \omega_{l_1,3} DA_0$
  4. For iterations $v = 1, \ldots, V$ and a_0,
     (a) Draw d_v from $Bernoulli(\hat{\mu}_D)$
     (b) Set $a_{v,0} = a_0$
     (c) Draw $\ell_{v,1}$ from $Bernoulli(p_{\ell_1}; d_v, a_{v,0})$
     (d) Set $a_{v,1} = \ell_{v,1}$

(e) $\hat{\delta}_v = \hat{E}[Y^{0,0}] + \sum_{l=0}^{1} \gamma_l(d_v, a_{v,0}, a_{v,1}; \hat{\psi})$

5. $\hat{E}[Y^{a_0, a_1 = L_1(a_0)}] = \frac{1}{V} \sum_{v=1}^{V} \hat{\delta}_v$

Here we have used $V = 10^6$.

- Finally $\hat{\Delta} = \hat{E}[Y^{a_0=1, a_1=L_1(a_0=1)}] - \hat{E}[Y^{a_0=0, a_1=L_1(a_0=0)}]$, and estimate for the variance of $\hat{\Delta}$, $\hat{\sigma}^2_{\hat{\Delta}}$, is obtained from B bootstrap samples.

3.B.3 Simulation results

Results from different scenarios are given in Table 3.9.

The simulation results show that under full compliance, both ITT and SNMM provide unbiased estimates. As expected, noncompliance leads to biased estimates when using ITT approach, but the estimates from SNMM remain unbiased and have smaller MSE with reasonable 95% coverage.

Table 3.9: Results from additional simulation scenarios. ($Bias = \hat{\theta} - \theta$). Coverage for ITT not included. Note that MSE=0.0000 refers to $< 1 \times 10^{-4}$.

S/N	Test1		Test0		Δ	SNMM				ITT		
	Sens	Spec	Sens	Spec		Est	Bias	MSE	95% Cvg	Est	Bias	MSE
1	0.96	0.95	0.86	0.89	0.0192	0.0191	-0.0001	0.0001	0.954	0.0048	-0.0144	0.0003
2	0.96	0.95	0.96	0.99	-0.0028	-0.0030	-0.0002	0.0001	0.930	-0.0017	0.0011	0.0001
3	0.96	0.95	0.76	0.99	0.0272	0.0273	0.0001	0.0000	0.940	0.0033	-0.0239	0.0006

3.C Binary outcome

When Y_{bin} is a binary outcome, and the objective of the analysis is to estimate the causal estimand

$$\Delta_{bin} = \frac{E[Y_{bin}^{a_0=1,a_1=L_1(a_0=1)}]}{E[Y_{bin}^{a_0=0,a_1=L_1(a_0=0)}]}$$

had, contrary to fact, all subjects remained on protocol, then the SNMM can be defined as

$$\frac{E[Y_{bin}^{\bar{a}_m,\underline{0}}|\bar{L}_m = \bar{\ell}_m, \bar{A}_m = \bar{a}_m]}{E[Y_{bin}^{\bar{a}_{m-1},\underline{0}}|\bar{L}_m = \bar{\ell}_m, \bar{A}_m = \bar{a}_m]} = \exp\{\gamma_m(\bar{\ell}_m, \bar{a}_m; \psi)\}$$

Defining $H_m(\psi) = Y_{bin} \exp\Big\{-\sum_{l=m}^{K} \gamma_l(\bar{L}_l, \bar{A}_l; \psi)\Big\}$, then

$$\begin{aligned}\Rightarrow E[H_m(\psi)|\bar{L}_m = \bar{\ell}_m, \bar{A}_m = \bar{a}_m] &= E[Y_{bin}|\bar{L}_m = \bar{\ell}_m, \bar{A}_m = \bar{a}_m] \exp\left\{-\sum_{l=m}^{K} \gamma_l(\bar{\ell}_l, \bar{a}_l; \psi)\right\} \\ &= E[Y_{bin}^{\bar{a}_{m-1},\underline{0}}|\bar{L}_m = \bar{\ell}_m, \bar{A}_m = \bar{a}_m]\end{aligned}$$

For the 2-arm DRCT design that we are considering, the estimating equation for each individual i is

$$U_i(\psi, \xi; \Upsilon, J) = \begin{bmatrix} \sum_{m=0}^{1}\big(A_{m,i} - E[A_{m,i}|\bar{A}_{m-1,i}, \bar{L}_{m,i}; \hat{\alpha}_m]\big) J_{m,i}\big[Y_{bin,i} \exp\big\{-\sum_{j=m}^{1} \gamma_j(\bar{L}_{j,i}, \bar{A}_{j,i}; \psi)\big\} - \exp\big\{\Upsilon_m(\bar{A}_{m-1,i}, \bar{L}_{m,i}; \xi)\big\}\big] \\ \sum_{m=0}^{1} Q_{m,i}\big[Y_{bin,i} \exp\big\{-\sum_{j=m}^{1} \gamma_j(\bar{L}_{j,i}, \bar{A}_{j,i}; \psi)\big\} - \exp\big\{\Upsilon_m(\bar{A}_{m-1,i}, \bar{L}_{m,i}; \xi)\big\}\big] \end{bmatrix}$$

This estimating equation does not have a closed form solution, but it can be solved

as a minimization problem.

$$\begin{aligned} \underset{\psi,\xi}{\text{minimize}} \quad & \tilde{U}^T S_u^{-1} \tilde{U} \\ \text{subject to} \quad & \gamma_0(\psi) + \gamma_1(\psi) + \Upsilon_0(\xi) \leq 0 \\ & \gamma_1(\psi) + \Upsilon_1(\xi) \leq 0 \end{aligned}$$

where

$$\begin{aligned} \tilde{U} &= \sum_{i=1}^{n} U_i \\ S_u &= \sum_{i=1}^{n} U_i U_i^T \\ \gamma_1(\bar{L}_{1,i}, \bar{A}_{1,i}; \psi) &= (\psi_1 + \psi_2 D_i + \psi_3 A_{0,i} + \psi_4 D_i A_{0,i}) A_{1,i} \\ \gamma_0(\bar{L}_{0,i}, \bar{A}_{0,i}; \psi) &= (\psi_5 + \psi_6 D_i) A_{0,i} \\ \Upsilon_1 = \Upsilon_0 &= \xi_1 + \xi_2 D_i + \xi_3 A_{0,i} + \xi_4 D_i A_{0,i} \\ J_{1,i} &= \begin{bmatrix} 1 & D_i & A_{0,i} & D_i A_{0,i} & 0 & 0 \end{bmatrix}^T \\ J_{0,i} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & D_i \end{bmatrix}^T \\ Q_{1,i} = Q_{0,i} &= \begin{bmatrix} 1 & D_i & A_{0,i} & D_i A_{0,i} \end{bmatrix}^T \end{aligned}$$

The constraints are derived from

$$\begin{aligned} \exp\{\gamma_m(\bar{\ell}_m, \bar{a}_m; \psi)\} &= \frac{E[Y_{bin}^{a_{m-1}, \underline{a}_m} | \bar{L}_m = \bar{\ell}_m, \bar{A}_m = \bar{a}_m]}{E[Y_{bin}^{\bar{a}_{m-1}, \underline{0}} | \bar{L}_m = \bar{\ell}_m, \bar{A}_m = \bar{a}_m]} \\ \exp\{\gamma_m(\bar{\ell}_m, \bar{a}_m; \psi)\} &\leq \frac{1}{E[Y_{bin}^{\bar{a}_{m-1}, \underline{0}} | \bar{L}_m = \bar{\ell}_m, \bar{A}_m = \bar{a}_m]} = \frac{1}{E[H_m(\psi) | \bar{L}_m = \bar{\ell}_m, \bar{A}_m = \bar{a}_m]} \end{aligned}$$

at $m = 1$,

$$\exp\{\gamma_1(\bar{\ell}_1, \bar{a}_1; \psi)\} = \frac{E[Y_{bin}^{a_0,a_1} | \bar{L}_1 = \bar{\ell}_1, \bar{A}_1 = \bar{a}_1]}{E[Y_{bin}^{a_0,0} | \bar{L}_1 = \bar{\ell}_1, \bar{A}_1 = \bar{a}_1]}$$

$$\gamma_1(\bar{\ell}_1, \bar{a}_1; \psi) + \Upsilon_1(\bar{\ell}_1, \bar{a}_1; \xi) \leq 0$$

at $m = 0$,

$$\exp\{\gamma_1(\bar{\ell}_1, \bar{a}_1; \psi)\} \exp\{\gamma_0(\ell_0, a_0; \psi)\} = \frac{E[Y_{bin}^{a_0,a_1} | \bar{L}_1 = \bar{\ell}_1, \bar{A}_1 = \bar{a}_1]}{E[Y_{bin}^{0,0} | L_1 = \ell_1, A_1 = a_1]}$$

$$\gamma_1(\bar{\ell}_1, \bar{a}_1; \psi) + \gamma_0(\ell_0, a_0; \psi) + \Upsilon_0(\ell_0, a_0; \xi) \leq 0$$

A consistent estimate of $E[Y_{bin}^{0,0}]$ can be obtained by computing

$$\frac{1}{n} \sum_{i=1}^{n} Y_{bin,i} \prod_{m=0}^{1} \exp[-\gamma_m(\bar{L}_{m,i}, \bar{A}_{m,i}; \hat{\psi})]$$

and this estimate can then be used to generate other average potential outcomes using the same Monte Carlo approach as the continuous outcome case to derive the average causal risk ratio $E[Y^{a_0=1,a_1=L_1(a_0=1)}]/E[Y^{a_0=0,a_1=L_1(a_0=0)}]$.

Simulation results for 5-year survival are shown in Table 3.10. The results are based on 19 and 12 simulations for scenarios 1 and 2 respectively, and 500 bootstrap samples each. The constraint optimization algorithm `constrOptim` in `R` was used.

Table 3.10: Simulation results for binary response

S/N	Test1		Test0		Δ	SNMM			ITT		
	Sens	Spec	Sens	Spec		Est	Bias	MSE	Est	Bias	MSE
1	0.96	0.95	0.86	0.89	0.942	0.945	0.003	0.00047	0.992	0.052	0.00011
2	0.96	0.99	0.96	0.95	0.988	0.993	0.004	0.00514	0.997	0.008	0.00032

Bibliography

D. R. Aberle, A. M. Adams, C. D. Berg, W. C. Black, J. D. Clapp, R. M. Fagerstrom, I. F. Gareen, C. Gatsonis, P. M. Marcus, and J. D. Sicks. Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England journal of medicine*, 365(5):395–409, 2011. ISSN 1533-4406. doi: 10.1056/NEJMoa1102873.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996. ISSN 0162-1459. doi: 10.1080/01621459.1996.10476902.

J. Bogaerts, F. Cardoso, M. Buyse, S. Braga, S. Loi, J. a. Harrison, J. Bines, S. Mook, N. Decker, P. Ravdin, P. Therasse, E. Rutgers, L. J. van 't Veer, and M. Piccart. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. *Nature clinical practice. Oncology*, 3(10):540–551, 2006. ISSN 1743-4254. doi: 10.1038/ncponc0591.

P. M. Bossuyt, J. G. Lijmer, and B. W. Mol. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet*, 356(9244):1844–1847, 2000. ISSN 01406736. doi: 10.1016/S0140-6736(00)03246-3.

T. R. Church, W. C. Black, D. R. Aberle, C. D. Berg, K. L. Clingan, F. Duan, R. M. Fagerstrom, I. F. Gareen, D. S. Gierada, G. C. Jones, I. Mahon, P. M. Marcus, J. D. Sicks, A. Jain, and S. Baum. Results of initial low-dose computed tomographic

screening for lung cancer. *The New England journal of medicine*, 368(21):1980–91, 2013. ISSN 1533-4406. doi: 10.1056/NEJMoa1209120.

J. C. De Graaff, D. T. Ubbink, D. A. Legemate, J. G. P. Tijssen, and M. J. H. M. Jacobs. Evaluation of toe pressure and transcutaneous oxygen measurements in management of chronic critical leg ischemia: A diagnostic randomized clinical trial. *Journal of Vascular Surgery*, 38(3):528–534, 2003. ISSN 07415214. doi: 10.1016/S0741-5214(03)00414-2.

J. C. De Graaff, D. T. Ubbink, J. G. P. Tijssen, and D. A. Legemate. The diagnostic randomized clinical trial is the best solution for management issues in critical limb ischemia. *Journal of Clinical Epidemiology*, 57(11):1111–1118, 2004. ISSN 08954356. doi: 10.1016/j.jclinepi.2004.02.020.

W. M. L. L. G. Deserno, M. G. Harisinghani, M. Taupitz, G. J. Jager, J. A. Witjes, P. F. Mulders, C. A. Hulsbergen van de Kaa, D. Kaufmann, and J. O. Barentsz. Urinary bladder cancer: preoperative nodal staging with ferumoxtran-10-enhanced MR imaging. *Radiology*, 233(2):449–56, 2004.

C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002. ISSN 0006-341X. doi: 10.2307/3068286.

I. F. Gareen. Noncompliance in cancer screening trials. *Clinical trials (London, England)*, 4(4):341–349, 2007. ISSN 1740-7745. doi: 10.1177/1740774507081341.

C. Gatsonis, D. R. Aberle, C. D. Berg, W. C. Black, T. R. Church, R. M. Fagerstrom, B. Galen, I. F. Gareen, J. Goldin, J. K. Gohagan, B. Hillman, C. Jaffe, B. S. Kramer, D. Lynch, P. M. Marcus, M. Schnall, D. C. Sullivan, D. Sullivan, and C. J. Zylak. The National Lung Screening Trial: overview and study design. *Radiology*, 258(1):243–253, 2011. ISSN 0033-8419. doi: 10.1148/radiol.10091808.

E. Goetghebeur and V. Stijn. Structural mean models for compliance analysis in

randomized clinical trials and the impact of errors on measures of exposure. *Statistical methods in medical research*, 14(4):397–415, 2005. ISSN 0962-2802. doi: 10.1191/0962280205sm407oa.

P. W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 0162-1459. doi: 10.2307/2289064.

R. Hooper, K. Díaz-Ordaz, A. Takeda, and K. Khan. Comparing diagnostic tests: Trials in people with discordant test results. *Statistics in Medicine*, 32(14):2443–2456, 2013. ISSN 02776715. doi: 10.1002/sim.5676.

L. Humphrey, M. Deffebach, M. Pappas, C. Baumann, K. Artis, J. Priest Mitchell, B. Zakher, R. Fu, and C. Slatore. Screening for lung cancer with low-dose computed tomography: A systematic review to update the u.s. preventive services task force recommendation. *Annals of Internal Medicine*, 159(6):411–420, sep 2013. ISSN 0003-4819.

G. W. Imbens and J. D. Angrist. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–75, 1994. ISSN 00129682. doi: 10.2307/2951620.

J. G. Lijmer and P. M. Bossuyt. Diagnostic testing and prognosis: the randomised controlled trial in diagnostic research. In J. A. Knottnerus, editor, *Evidence Base of Clinical Diagnosis*, chapter 4, pages 61–80. BMJ Books, 2002.

B. Lu and C. Gatsonis. Efficiency of study designs in diagnostic randomized clinical trials. *Statistics in Medicine*, 32(9):1451–1466, 2013. ISSN 02776715. doi: 10.1002/sim.5655.

E. L. Ogburn and T. J. VanderWeele. On the Nondifferential Misclassification of a Binary Confounder. *Epidemiology*, 23(3):433–439, 2012. ISSN 1044-3983. doi: 10.1097/EDE.0b013e31824d1f63.

S. Picciotto, M. A. Hernán, J. H. Page, J. G. Young, and J. M. Robins. Structural Nested Cumulative Failure Time Models to Estimate the Effects of Interventions. *Journal of the American Statistical Association*, 107(499):886–900, sep 2012. ISSN 0162-1459. doi: 10.1080/01621459.2012.682532.

P. C. Prorok, G. L. Andriole, R. S. Bresalier, S. S. Buys, D. Chia, E. David Crawford, R. Fogel, E. P. Gelmann, F. Gilbert, M. A. Hasson, R. B. Hayes, C. C. Johnson, J. S. Mandel, A. Oberman, B. O'Brien, M. M. Oken, S. Rafla, D. Reding, W. Rutt, J. L. Weissfeld, L. Yokochi, and J. K. Gohagan. Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Controlled Clinical Trials*, 21(6):273S–309S, 2000. ISSN 0197-2456. doi: http://dx.doi.org/10.1016/S0197-2456(00)00098-2.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986. ISSN 02700255. doi: 10.1016/0270-0255(86)90088-6.

J. Robins. Addendum to a new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Computers & Mathematics with Applications*, 14(9-12):923–945, 1987. ISSN 08981221. doi: 10.1016/0898-1221(87)90238-0.

J. Robins. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79(2):321–334, 1992. ISSN 00063444. doi: 10.1093/biomet/79.2.321.

J. Robins and A. Rotnitzky. Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika*, 91(4):763–783, 2004. ISSN 00063444. doi: 10.1093/biomet/91.4.763.

J. M. Robins. The analysis of randomized and non- randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. Freeman, and A. Mulley, editors, *Health Service Research Methodology: A Focus on AIDS*. U.S. Public Health Service, Washington, DC., 1989.

J. M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and Methods*, 23(8): 2379–2412, 1994. ISSN 0361-0926. doi: 10.1080/03610929408831393.

J. M. Robins. Correction for non-compliance in equivalence trials. *Statistics in Medicine*, 17(3):269–302, 1998. ISSN 02776715.

J. M. Robins. Marginal Structural Models versus Structural nested Models as Tools for Causal inference. In M. E. Halloran and D. Berry, editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, volume 116, pages 95–133. Springer New York, 2000. ISBN 978-1-4612-7078-2. doi: 10.1007/978-1-4612-1284-3_2.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. ISSN 0022-0663. doi: 10.1037/h0037350.

D. B. Rubin. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75(371): 591–593, sep 1980. ISSN 01621459. doi: 10.2307/2287653.

D. J. Sargent, B. a. Conley, C. Allegra, and L. Collette. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*, 23(9):2020–2027, 2005. ISSN 0732183X. doi: 10.1200/JCO.2005.01.112.

T. J. VanderWeele. Concerning the Consistency Assumption in Causal Inference. *Epidemiology*, 20(6), 2009. ISSN 1044-3983.

S. Vansteelandt and E. Goetghebeur. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 65(4):817–835, 2003. ISSN 13697412. doi: 10.1046/j.1369-7412.2003.00417.x.

S. Vansteelandt and M. Joffe. Structural Nested Models and G-estimation: The Partially Realized Promise. *Statistical Science*, 29(4):707–731, 2014. ISSN 0883-4237. doi: 10.1214/14-STS493.

I. R. White. Uses and limitations of randomization-based efficacy estimators. *Statistical Methods in Medical Research*, 14:327–347, 2005.