Simulations and Case Studies Concerning the Relative Impact Curve

by

Lisa Leung

B.Sc., Simon Fraser University, 2014

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Sciences

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Statistics)

The University of British Columbia
(Vancouver)

April 2018

Abstract

Relative Impact Characteristic (RIC) curve is a new statistical methodology to assist physicians on treatment decisions based on defined biomarkers. Similar to Receiving Operating Characteristic (ROC) curve, the results provide a graphical curve that uses the same measure, area under the curve, to evaluate the model. This project provides the motivation behind the development on RIC, the functions and its definitions related to RIC, a simulation study on the sampling variabilities, and a case study to implement the RIC model. As a result, the simulation study show that there are differences if linear regression model is used as the prediction model, but not much differences when a poisson model is used. The case study shows that different biomarker models produce different RIC curves when the same curve uses the same benefit function.

Table of Contents

List of Tables

List of Figures

Acknowledgments

Firstly, I'd like to thank Dr. Paul Gustafson from UBC for his guidance on my work for this project. He has provided great support during my work on this research project. I'd also like to thank the entire Statistics department at UBC for their help during my entire Master's program.

Chapter 1

Introduction

The Receiving Operating Characteristic (ROC) curve has been a widely used statistical tool to determine how well a model fits in predicting a binary outcome. The field of biostatistics has been relying on this tool for researchers to assess how well a model predicts a binary disease outcome. Obtaining a statistical model that best predicts a disease status based on data collected beforehand is quite essential to the industry, and what is equally important is a model that best predicts the effects of a treatment using data beforehand as well. Utilizing the idea of an ROC curve, a new statistical tool, namely the Relative Impact Characteristic (RIC) curve, is created to address this problem (Sadatsafavi et al., In Review).

Biomarkers are often used as an outcome measurement for disease assessements. Researchers and physicians use various biomarker thresholds as indicators to determine diagnosis on individuals of a certain disease. For example, researchers and specialists in COPD, the third leading cause of death (Obeidat et al., 2015) uses Forced Expiratory Volume (FEV) and Forced Vital Capacity (FVC) measurements as biomarkers to determine whether a patient has mild or severe COPD. On the other hand, diseases such as Huntington's disease (Disatnik et al., 2016) and rheumatic diseases (Robinson et al., 2016) have recently been researched to determine relevant biomarkers. These research studies are done so that physicians can utilize these measures to process a more accurate and possibly faster diagnosis. To assess the best performing biomarkers on detecting diseases, an ROC curve is usually determined. Based on the growing importance and popularity of

using biomarkers and other multi-omics data, a similar tool to assess the effect of treatment on disease is needed to further improve the field of biomedical science.

In a similar fashion, biomarkers can also be used to predict the treatment effects of a diseased population. The use of biomarkers can be seen in action with the RIC method. For the purposes of this methodology, the definition of biomarkers mentioned in this paper would be described as any possible patient characteristics that are recorded prior to treatment or at baseline. Further details of the definition are given in the next section.

As mentioned, the methodology of RIC is inspired by a very commonly used graph, namely the ROC curve. The statistical methodology could eventually help physicians decide whether patients should receive treatment or not in the future based on the added benefit to the overall population. As a result, a visualization will act as a tool for physicians to assess the how well the biomarker-based treatment decisions will benefit the population. In application, RIC takes in a data set that contains results from either clinical trials or observational studies. The data provide biomarker values as an indicator for treatment thresholds, the benefit of treatment based on improvement of the disease, a response variable to measure the outcome of the disease, and risk factors that are associated with the beneficial measurements or the outcome variable. Note that it is possible that the biomarker function and benefit function could be equivalent which will be discussed below.

1.1 Definition and Assumption

Unlike the ROC method of assessing a model that uses biomarkers to predict a binary disease outcome, the RIC method takes the ROC curve in a different direction and assesses the treatment effect (measured in a continuous or a high-level nominal variable) at a population level using biomarkers. Ideally, the visualization is designed so that physicians and researchers can easily assess the beneficial effect on a particular disease or disorder based on the biomarker various threshold points. A beneficial effect can be interpreted in many ways; it could be a single variable that the researcher chooses to capture the end effects of the treatment, or it could be researcher-defined variable that is a combination of multiple effects of variables post treatment. The RIC method consists of a graph between biomarker

values and relative treatment benefit. Both of these functions are defined based on the researcher's subject knowledge of the disease. The benefit of treatment should return a non-negative scalar value that measures health improvements in which a higher value represents a positive result.

RIC is defined as the relative benefit outcome on a treated proportion of the diseased population against the threshold of a biomarker to treat. The relative benefit outcome is illustrated as a health measure that sums up the outcomes from a treated proportion over the whole population. In practice, and in many study designs with treatment randomization, the benefit function is assumed to have an equal impact on each patient; hence, randomized clinical trials are conducted by looking at the overall outcomes of treatment effects (ie. the benefit function). With RCTs, researchers automatically assume that the effects of treatments are consistent for each individual. However, in reality, treatment effects may vary due to specific biomarker(s) of a patient. Hence, we would like to look at what happens to the treatment effects if patients are assigned only by a certain threshold of biomarkers, defined by researchers. The RIC curve will present whether or not the treatment effect of the whole population differ from a subgroup of patients who received the treatment by a biomarker's threshold.

There are three main types of variables in a data set used to produce on RIC curve:

- **Treatment.** Treatment variable is labeled as X in which X is a binary treatment decision: $X = 1$ if treatment is assigned and $X = 0$ if treatment is not assigned.

- **Patient Characteristics.** Patient characteristics is a set of variables which will be denoted as C. The variables can take in any type of data in which describes the clinical measures of a patient at baseline. For the purposes of definition described below, we will denote C_1 as baseline characteristics measured routinely, and C_2 as baseline characteristics measured only within the research context and are difficult to measure. Patient characteristics could be age, gender, weight, height, blood pressure levels, obesity levels, and so on.

- **Patient Outcome.** Patient outcome is denoted as Y in which Y could be any type of variable that best measures the outcome after treatment. Outcomes could include any measurements of disease progression such as Forced Expiratory Volume (FEV), survival times, and so on. For illustration in this paper, we will assume that higher values of Y correspond to worse health outcomes.

With the types of variables defined, two main measures are used for the RIC curve. These measures take in the functions defined as:

- **Benefit Function.** Benefit function is denoted as $b()$, which is a function of the baseline variables (ie. covariates of patient characteristics). Here, the baseline variables is C which could be any sensible set of patient characteristics available in the data set. We define benefit function as the difference of expected outcome measurements with and without treatment. Note that C include both baseline characteristics measured routinely, and baseline characteristics that are only obtained in the research context prior to treatment. The benefit function is defined as:

$$b(c) = E(Y|X = 0, C) - E(Y|X = 1, C)$$

- **Biomarker Function.** Biomarker function is denoted as $m()$ in which $m()$ is a function of C_1, a set of covariates describing patient characteristics. Note that C_1 is used because the biomarker function is designed so that physicians can easily compute these values prior to assigning a treatment to patients. When constructing an RIC curve, the biomarker function is the variable that determines the treated proportion of the population and vice versa. A subject is treated if their biomarker function passes a certain threshold. For example, if the population's biomarker ranges continuously from 0 to 10 and the median value is 3, then the threshold for treating half the population corresponds to subjects with biomarker level at 3 and above. A biomarker function could also be a linear combination of age, gender, weight, and blood pressure levels at the beginning of the study.

Based on the description, some examples of biomarker function is defined as:
$m(C) = E(Y|X = 0, C_1)$ where C_1 is a subset of C as explained above, and X is the assignment of treatment. In this specific example, we assume the expected outcome measurements given all subjects have not received treatment.

There are cases in which we would like to evaluate biomarkers by the beneficial value. Alternatively, the goal of researchers using this model is attempting to assign treatment for those that get the most added benefit. If that is so, the biomarker could then defined as:
$m(C) = b(C) = E(Y|X = 0, C_1) - E(Y|X = 1, C_1)$

With the description explained previously and the definitions given, there are three assumptions made according to Sadatsafavi's definition of RIC curve.

- **Biomarker values do not differ through time.** In order to use the biomarker function that could be computed at any given time, biomarker values are assumed not to differ through time.
- **The marker positivity rule.** In many cases, physicians follow a simple guideline that gives informative threshold points to determine the severity of diseases. For instance, a common practice is to use BMI scores to determine whether a patient is underweight, normal, overweight, or obese. For the purpose of illustration, we will assume that individuals will receive treatment if the patient passes a certain threshold.
- **Individuals are independent and identically distributed.** The treatment effect of any individual is assumed to not affect another individual's treatment outcome. Although it is possible to further implement the correlation in future work, contagious diseases such as a cold flu are not applicable for this demonstration. Nontransmissable diseases such as emphysema, which is affected by a patient's genetics and environmental factors, are considered valid for this definition.

1.2 Interpretation

The purpose of an ROC curve is to assess a statistical model that predicts a binary outcome measure. Typically, researchers use the area under the curve of the ROC to tell us how the model can effectively predict true positive rates on a given probability threshold with a trade off of predicting false positives of a model. If the ROC curve is closer to a straight diagonal line (an AUC of 0.5), then it tells the researcher that the statistical model is no better than predicting the outcome by random chance. On the other hand, the interpretation of RIC curve is quite different - rather than assessing the model we use, we instead assess whether a treatment decision based on the biomarker would be effective or not.

Besides the axes labels on an RIC curve, the interpretation of the RIC curve highly resembles ROC. Area under the curve is measured, but AUC in this case has two different implications. The first implication of AUC is whether or not the biomarker is useful to look at as a physician when treating patients, and secondly, whether this treatment is effective based on one's biomarker level. Note that RIC does not measure whether the treatment is effective or not, but instead is the interpretation of whether the treatment is associated with the different levels of biomarker based on the population.

Chapter 2

Understanding the Sampling Variability Among Estimated and True Functions

The data sets simulated in the following examples are designed to have certain behaviours of the biomarker and benefit functions. Risk factors are generated to be positive values so that a large positive value is associated with a high value on outcome measurements (eg. outcome measurements would be the risks of disease or a worsen progression of the disease). Hence, the benefit values of a subject is calculated so they are all positive as well.

When analyzing an RIC curve on a data set, researchers assume that the data set used consists of results from a randomized control trial or an observational study and that the baseline characteristics are available for use. This chapter will look at two different statistical models: Linear Regression and Poisson Regression. In addition, we will look at two differently defined biomarkers in which one will include all variables in the prediction model, and another will include only some of the variables in the prediction model.

For the purposes of simulation, the data are simulated laboratory results that are baseline measures of phenotypes and simulated outcome measurements on the progression of disease. To estimate a patient's biomarker prior to treatment, we will need to establish a well estimated biomarker function based on the data available.

Consequently, we propose that the biomarker function is obtained by fitting the best prediction model on outcome measurements with a subset of risk factors and the treatment variable. Treatment variable will be included in the prediction model regardless of its significance due to the backgrond assumption that the treatment variable will affect benefit values. In our simulation study, we obtain an estimation of pre-treatment biomarkers by setting treatment value as 0 on the biomarker function.

To this date, standard errors from estimating the AUC under the RIC method have yet to be obtained. Hence, to understand the estimation errors of AUC under a RIC method, we will perform simulations under four different scenarios defined below.

2.1 Procedure on RIC function

- **Case 1: Estimated biomarker values and estimated benefit values.** Pre-treatment biomarker values are obtained from the estimated biomarker function. The benefit values are obtained by taking the difference of estimated outcome with treatment and estimated outcome without treatment. Note that the pre-treatment biomarker and benefit values in the RIC models are both estimated from a fitted prediction model.

- **Case 2: Estimated biomarker values and true benefit values.** The estimated biomarker values are obtained similar to Case 1. However, the true benefit values are obtained by taking a difference of true expected outcome with treatment and the true expected outcome without treatment.

- **Case 3: True biomarker values and estimated benefit values.** The true pre-treatment biomarker values are obtained by setting treatment as 0 for all subjects in the true biomarker function which are calculated without errors. Estimated benefit values are obtained as in Case 1.

- **Case 4: True biomarker values and true benefit values.** True pre-treatment biomarker values and true benefit values are obtained as in Case 2 and Case 3. True biomarker values used a true biomarker function with treatment as 0,

and true benefit values are obtained by taking the difference of true expected outcomes with and without treatment.

As mentioned, simulation is conducted in a situation such that previous clinical trials have been conducted. With this in mind, we will assume that risk factors are coded such that larger values correspond to the progression of disease and the decline in health benefit values. A treatment variable interacting with the risk factors will eventually decrease the progression of disease and increase in health benefit values.

2.2 Linear Regression Model

The data set simulated will consist of 100 sets of samples each with a sample size of 50 observations. Parameters in the generated data set are chosen arbitrarily. In this particular simulation, we assume that the linear regression model estimating outcome measurement incorporates into our definition of the biomarker and benefit values as well.

2.2.1 Steps of Data Generation

Step 1: Choose a set of risk factors in which a subset of them are within the biomarker and the benefit function. For simplicity, all parameters on the distributions are chosen to be positive. In this case, a data set of risk factors C_1 and C_2 are generated such that

$C_1 \sim U(min, max)$ where $min = 0, max = 10$

$C_2 \sim N(\mu, \sigma)$ where $\mu = 10, \sigma = 6$

Step 2: Generate a vector of treatment, x, which contains binary values of 0 and 1. To mimick a randomized trial data, treatments are randomly assigned from a binomial distribution. Note that X is independent of C_1 and C_2

$X \sim Bin(n, p)$ where $n = 1, p = 0.5$

Step 3: Using the generated samples from Steps 1 and 2, define a linear regression model for the outcome measurement, y. Below is the model for this particular simulation. Note that the model satisfies our assumption that risk factors will lower outcome values and treatment will increase benefit values.

$$E(Y|X,C_1,C_2) = C_1 + 0.5 * C_2 - 0.25 * X - 0.5 * X * C_1 - 0.25 * X * C_2$$

Step 4: Generate the samples of outcome measurements based on Step 3. The set of values are generated with errors such that

$$Y_i = C_{1i} + 0.5 * C_{2i} - 0.25 * X_i - 0.5 * X_i * C_{1i} - 0.25 * X_i * C_{2i} + \varepsilon_i$$

where

$\varepsilon_i \sim N(\mu, \sigma)$ where $\mu = 0, \sigma = 5$

Step 5: Generate the estimated and true biomarkers using the biomarker function. To generate the estimated biomarkers, we first fit a linear regression prediction model which takes in the risk factors and treatment variables as covariates and outcome measurement on response variable. This regression model is then used to predict the subjects' outcome measurement when all subjects receive no treatment.

The linear regression on outcome measurement is modeled as:

$$\widehat{y_i} = \widehat{b}_1 * C_{1i} + \widehat{b}_2 * C_{2i} + \widehat{b}_3 * X_i + \widehat{b}_4 * X_i * C_{1i} + \widehat{b}_5 * X_i * C_{2i}$$

Estimated biomarker values are fitted as:

$$\widehat{biomarker_i} = \widehat{b}_1 * C_{1i} + \widehat{b}_2 * C_{2i}$$

True biomarker values are calculated as:

$$biomarker_i = C_{1i} + 0.5 * C_{2i}$$

Step 6: Generate the estimated and true benefit values using the biomarker function. For the purposes of simulation, the benefit function is a function of the linear regression model. Hence, estimated benefit values are generated as:

$$\widehat{benefit}_i = \widehat{Y}_{i|X_i=0} - \widehat{Y}_{i|X_i=1} = -\widehat{b}_3 - \widehat{b}_4 * C_{1i} - \widehat{b}_5 * C_{2i}$$

The true benefit values are calculated as:

$$benefit_i = Y_{i|X_i=0} - Y_{i|X_i=1} = 0.25 + 0.5 * C_{1i} + 0.25 * C_{2i}$$

2.2.2 Simulation Results

Area Under the Curve (AUC) of the RIC curves are calculated using the AUC() function in DescTools R-package. Table 2.1 shows summary statistics on the four different cases. Histograms of AUC are plotted in the figures below. Note that the x-axes and y-axes are fixed within the same window frame at 0.4 to 0.9, and 0 to 45 respectively, to allow a better visual comparison between the graphs.

In summary, we expect the range to be the highest in Case 1 and lowest in Case 2. This is in correspondence with the biomarkers and benefit values we have to estimate from our simulated data. Overall, the results of these graphs correspond with our expectations.

It is interesting to observe similarities between Case 2 and Case 4. The variables computed between Case 2 and Case 4 differ by the true and estimated biomarker values, although they have provided very similar results of the AUC values. It seems that the errors from the estimated biomarker values do not affect the sampling variations produced in AUC; rather, the errors from the estimated benefit values seem to affect a lot more on the sampling variations surrounding AUC values.

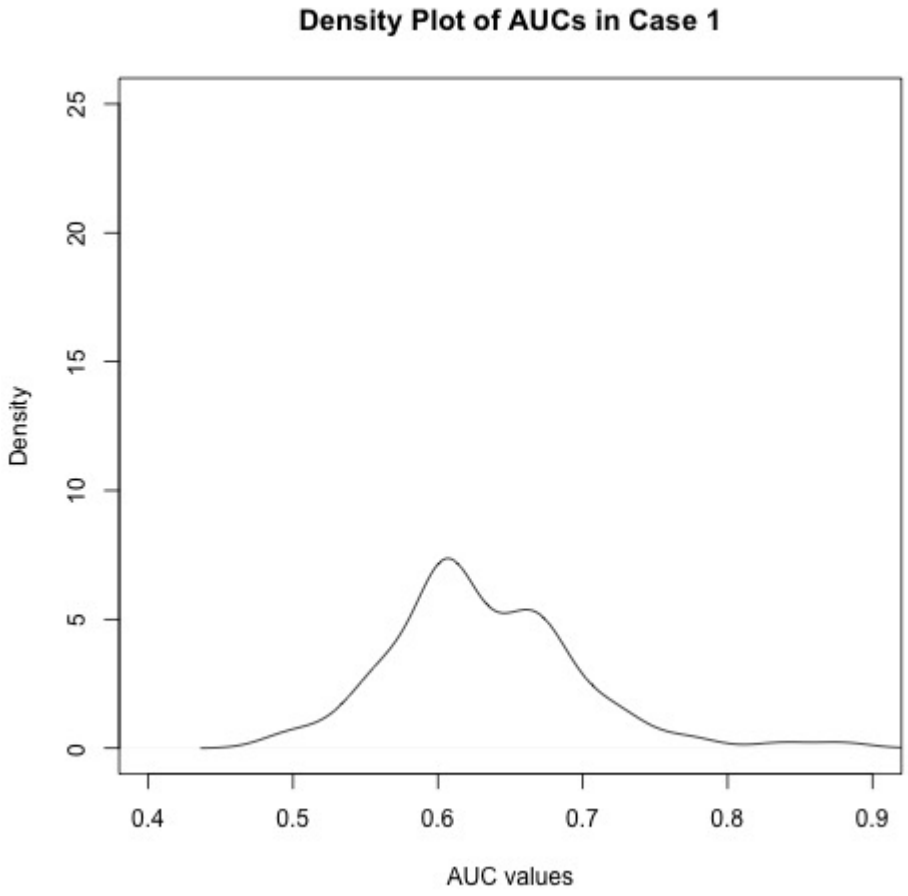

Figure 2.1: *Case 1 - Estimated biomarker values and estimated benefit values.* The AUC values for Case 1 ranges from 0.49 to 0.88. Notice that the distribution is normally distributed with a slight right-tail which could be negligible. The median of the AUC value is 0.62 and the mean of the AUC is 0.63. The middle half of the AUC values (ie. the range from 25th percentile to 75th percentile) lies at approximately 0.60 to 0.67.

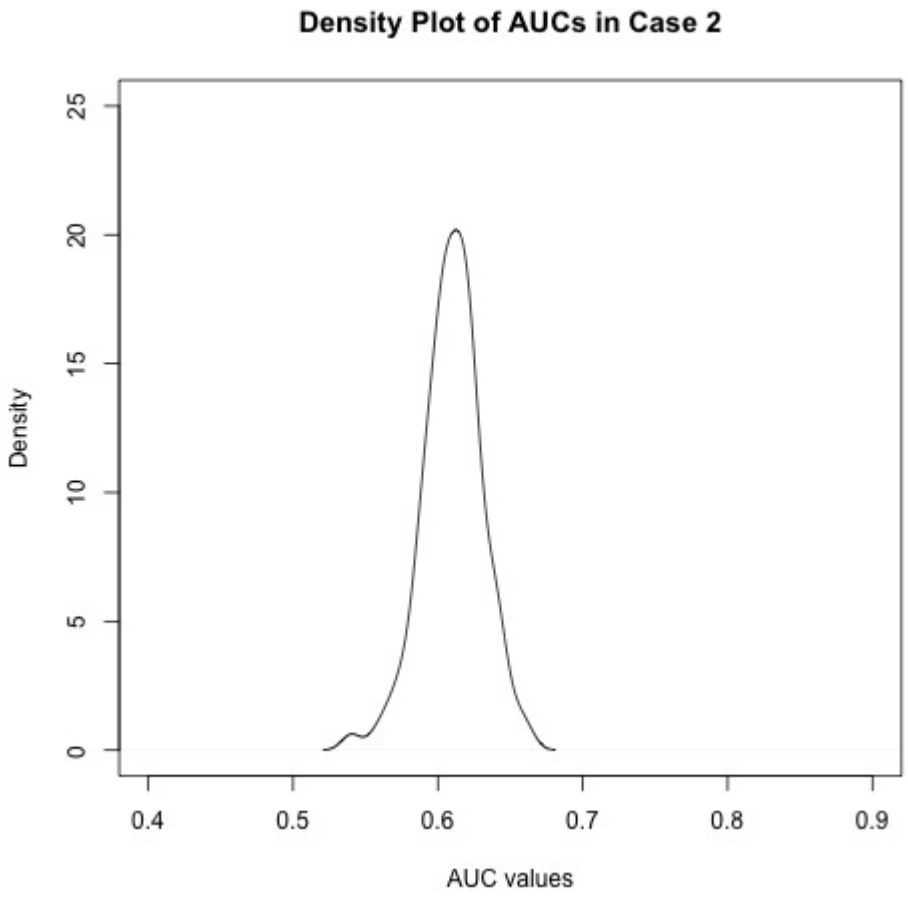

Figure 2.2: *Case 2 - Estimated biomarker values and true benefit values.* The AUC values for Case 2 ranges from 0.54 to 0.66. Notice that the distribution is normally distributed and more symmetrical than Case 1. The median of the AUC value is 0.61 and the mean of the AUC is 0.61, which seem to show a symmetrical evidence of the distributions. The middle half of the AUC values (ie. the range from 25th percentile to 75th percentile) lies at approximately 0.60 to 0.62, which is a relatively a narrow range when comparing to Case 1. These results are due to the true benefit values.

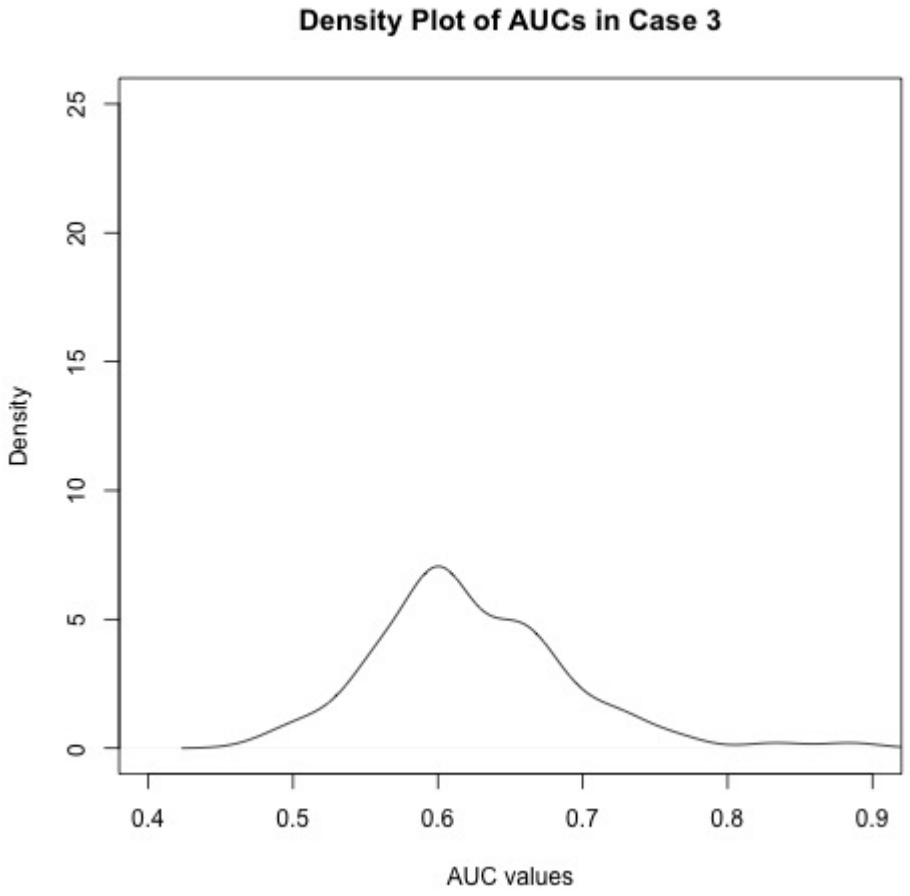

Figure 2.3: *Case 3 - True biomarker values and estimated benefit values.* The AUC values for Case 3 ranges from 0.48 to 0.89. In contrast to the observed distributions in Case 1 and Case 2, we see that the distribution in Case 3 seem to be highly skewed to the right. The median of the AUC value is 0.61 and the mean of the AUC is 0.62. The middle half of the AUC values (ie. the range from 25th percentile to 75th percentile) lies at approximately 0.59 to 0.66, which is quite similar to the range in Case 1. Comparing these numbers to Case 2 in which both cases are looking at one estimated variable, the graph from Case 3 seem to suggest that errors in estimating the benefit values are much bigger than the errors in estimating the biomarker values.

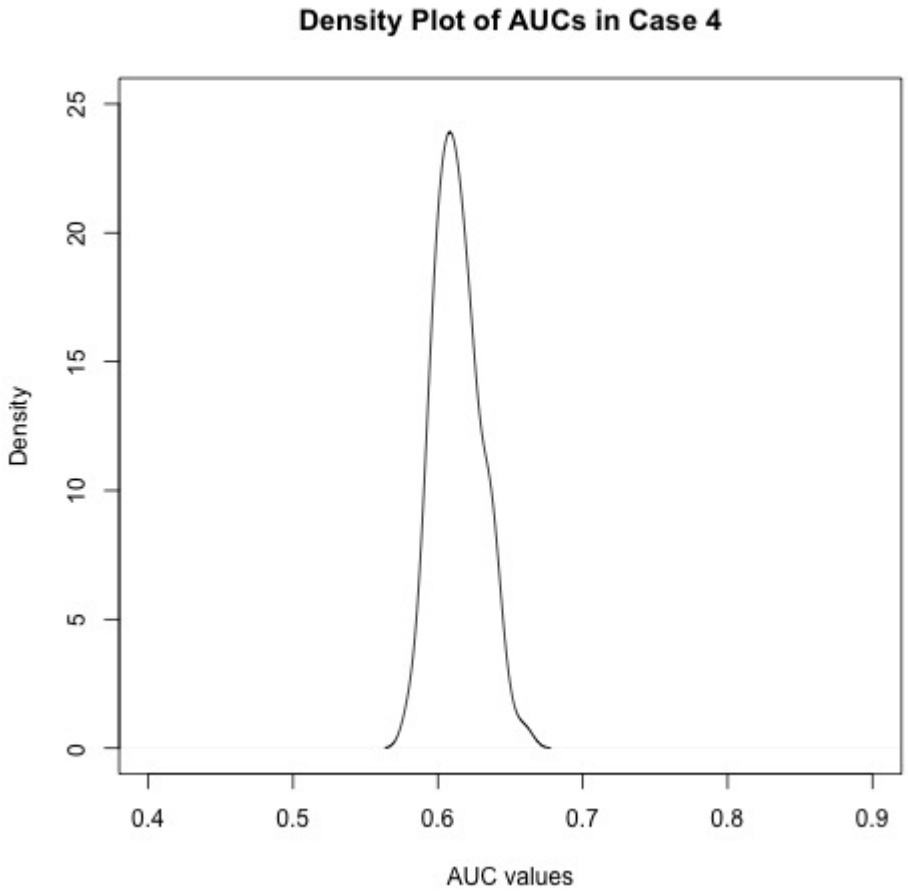

Figure 2.4: *Case 4 - True biomarker values and true benefit values.* The AUC values for Case 4 ranges from 0.58 to 0.66. As expected, the range is the smallest amongst all the four cases defined. The median and the mean of the AUC is 0.61 which is similar to Case 2. The middle half of the AUC values (ie. the range from 25th percentile to 75th percentile) lies at approximately 0.60 to 0.62, which is a very similar range in Case 2. Overall, Case 4 gave us the smallest ranges

Table 2.1: Summary statistics on the four cases

Case	Mean of AUC	Median of AUC	Standard Deviation of AUC
1	0.632	0.619	0.065
2	0.501	0.502	0.018
3	0.500	0.500	0.025
4	0.614	0.612	0.016

2.3 Poisson Regression Model

The data simulated under the Poisson regression model will consist of 100 sets of samples each with a sample size of 20 observations. The sample size is reduced to differentiate the results among the four cases. In addition to the change in sample size, there are also changes in the biomarker and benefit functions as well. Instead of using all covariates available in the biomarker and benefit function, we will choose a subset of these covariates.

2.3.1 Steps of Data Generation

Step 1: A set of risk factors are chosen denoted as C_1 and C_2 such that:

$C_1 \sim U(min, max)$ where $min = 0, max = 10$

$C_2 \sim N(\mu, \sigma)$ where $\mu = 10, \sigma = 6$

Step 2: A vector of treatment denoted as X contains binary values of 0 and 1 such that:

$X \sim Bin(n, p)$ where $n = 1, p = 0.5$

Step 3: Using the generated vectors from Steps 1 and 2, we define a Poisson regression model for the outcome measurement, $log(E(Y))$.

$$log(E(Y)) = 0.2 + 0.1 * C_1 + 0.25 * C_2 - 0.25 * X - 0.05 * X * C_1 - 0.125 * X * C_2$$

Step 4: Generate the samples of outcome measurements based on Step 3. The set of values are generated such that

$$Y_i \sim Poisson(e^{0.2+0.1*C_{1i}+0.25*C_{2i}-0.25*X_i-0.05*X_i*C_{1i}-0.125*X_i*C_{2i}})$$

Note that $E(Y) = Var(Y) = \lambda$ when $Y \sim Poisson(\lambda)$

Step 5: To generate the estimated biomarkers, we fit a Poisson prediction model which uses the risk factors and treatment variables as covariates. To further differentiate the results between the four cases, the predicted model will use a subset of variables from step 3. We then replace the treatment = 0 and compute the estimated biomarkers.

The biomarker function is modeled as:

$$\widehat{Y_i} = e^{\widehat{b}_1*C_{1i}+\widehat{b}_2*C_{2i}+\widehat{b}_3*X_i*C_{1i}}$$

Estimated biomarker values are fitted as:

$$\widehat{biomarker_i} = e^{\widehat{b}_1*C_{1i}+\widehat{b}_2*C_{2i}}$$

True biomarker values are calculated as:

$$biomarker_i = e^{C_{1i}+0.25*C_{2i}}$$

Step 6: Generate the estimated and true benefit values using the biomarker function. For the purposes of simulation, the benefit function is based on the prediction model under Step 5. Hence, estimated benefit values are generated as:

$$\widehat{benefit_i} = e^{\widehat{Y}_{i|X_i=0}} - e^{\widehat{Y}_{i|X_i=1}} = e^{-\widehat{b}_3}$$

The true benefit values are calculated as:

Table 2.2: Summary statistics on the four cases

Case	Mean of AUC	Median of AUC	Standard Deviation of AUC
1	0.842	0.841	0.050
2	0.841	0.841	0.046
3	0.832	0.832	0.056
4	0.840	0.840	0.047

$$benefit_i = Y_{i|X_i=0} - Y_{i|X_i=1} = e^{0.25+0.05*X_{1i}}$$

2.3.2 Simulation Results

Area Under the Curve (AUC) of the RIC curves are calculated using the AUC() function in DescTools R-package. Table 2.2 summarizes numerically of the variabilities among the four cases. Histograms of AUC are plotted in the figures below. Note that the x-axes and y-axes are fixed within the same window frame at 0.5 to 19, and 0 to 10 respectively, to allow a better visual comparison between the graphs.

Interestingly, the variabilities of AUCs in all four cases are very similar under the Poisson regression which contrasts the results we have compared under the Linear regression simulation. As a reminder, the simulation under the Linear regression takes biomarker values as a function of all the covariates in the prediction model, whereas the Poisson regression takes biomarker values as a function of a few covariates in the prediction model. Even with the distinction, the Poisson model seem to give more consistent results than the Linear model itself.

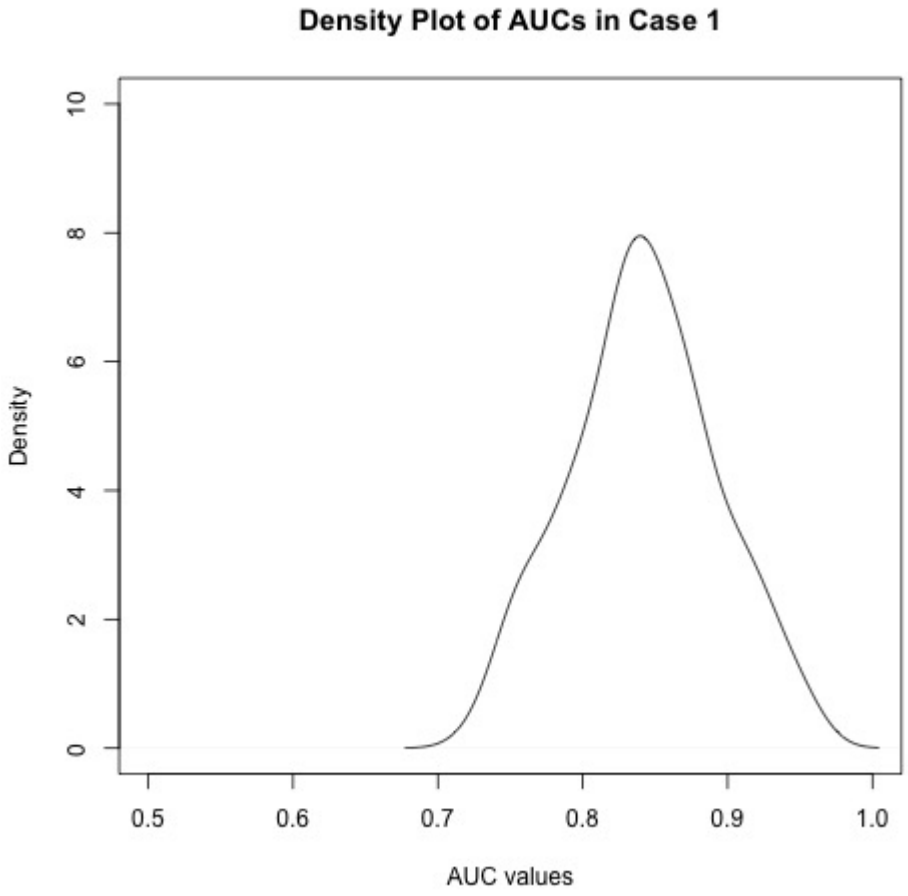

Figure 2.5: *Case 1 - Estimated biomarker values and estimated benefit values.* The mean and median of Case 1 are 0.84. The range of AUCs are between 0.73 to 0.95 with an interquartile range of 0.81 to 0.87. The spread appears to be normally distributed.

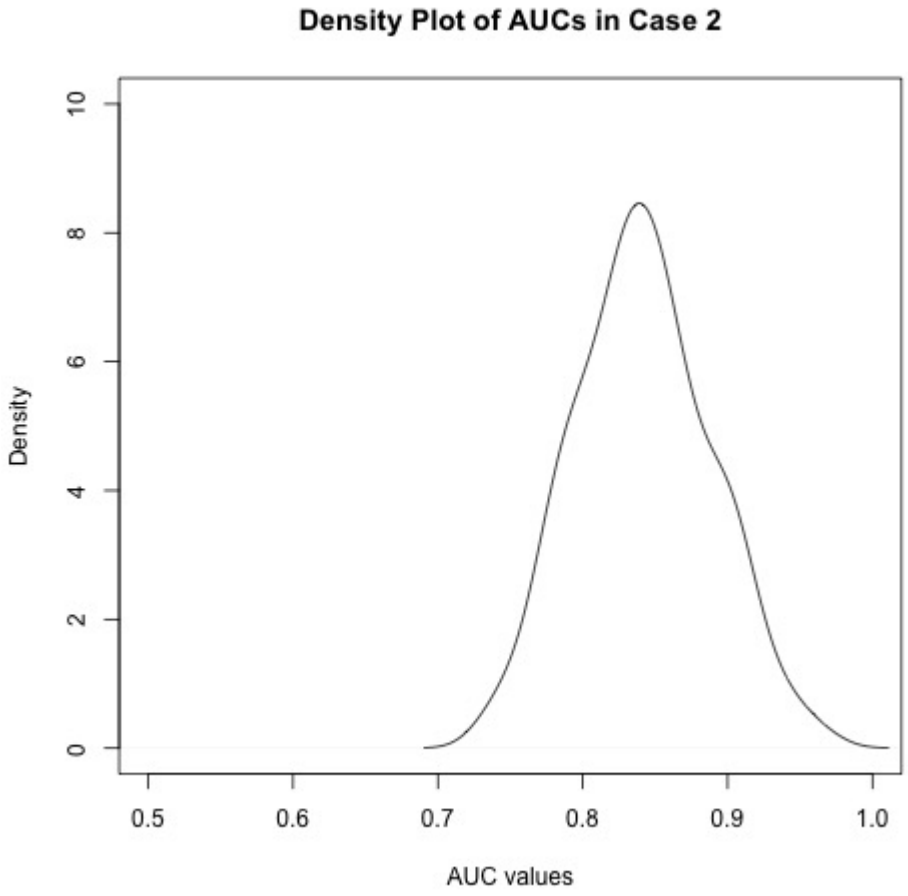

Figure 2.6: *Case 2 - Estimated biomarker values and true benefit values.* The mean and median of AUC in Case 2 are 0.84. The range of AUCs are between 0.74 to 0.96 with an interquartile range of 0.81 to 0.87. The spread appears to be normally distributed.

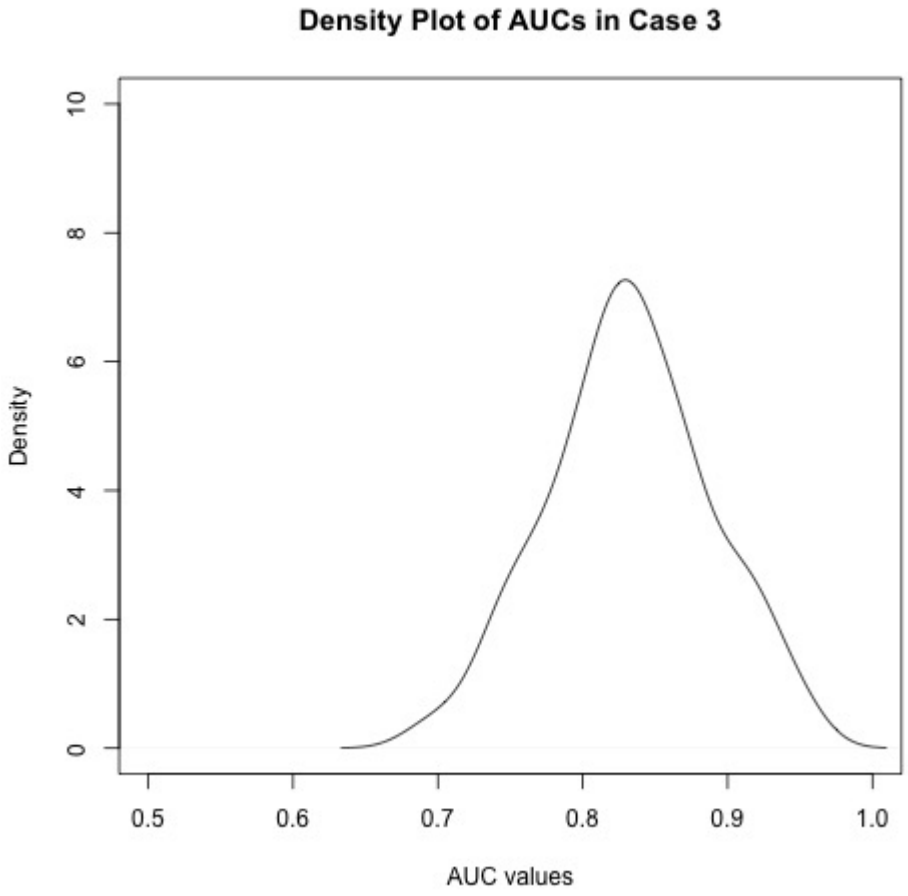

Figure 2.7: *Case 3 - True biomarker values and estimated benefit values.* The mean and median of AUCs are 0.83. The range of AUCs are between 0.70 to 0.96 with an interquartile range of 0.80 to 0.87. The spread appears to be normally distributed.

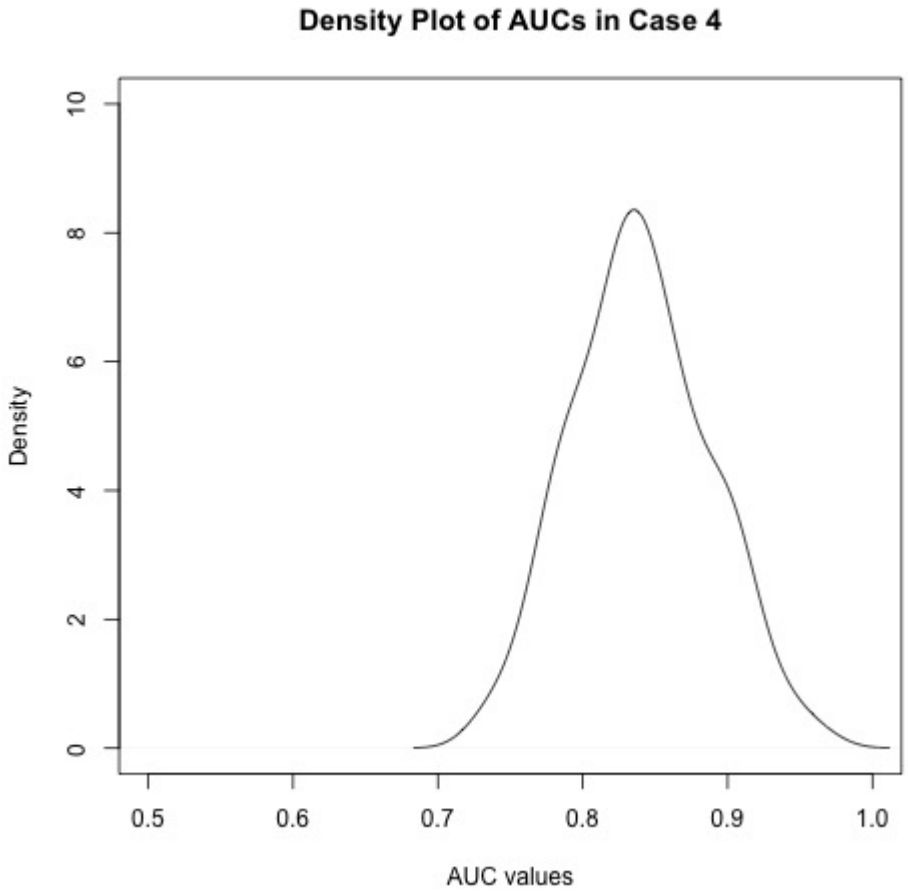

Figure 2.8: *Case 4 - True biomarker values and true benefit values.* The mean and median of AUC are 0.84. The range of AUCs are between 0.73 to 0.96 with an interquartile range of 0.81 to 0.87. The spread appears to be normally distributed.

Chapter 3

Case Study

To further apply this RIC methodology to a real data set, we will use the Primary Biliary Cirrhosis (PBC) Data (Murtaugh et al., 1994)

As an introduction, these data arose from a randomized clinical trial (RCT) conducted at the Mayo Clinic on January 1974 to May 1984 (Murtaugh et al., 1994). PBC is a chronic liver disease in which the cause of disease still remains unknown. Fortunately, there are various treatments to relieve the symptoms and the progression can be controlled. As such, the researchers of the PBC study studied the effect of a drug, D-penicillamine, on patients with treatment compared to placebo using different statistical models.

For our purposes of illustrating the RIC curve, we will not cover the research and studies that have been analyzed to study the effects of the treatment. Furthermore, we will not be assessing the performance of these prediction models. Instead, our goal is to analyze RIC curve develop biomarker and benefit functions to illustrate the usage of RIC on a publicly accessible data set. We will then compare the changes of an RIC curve in respect to different biomarker functions to a specified benefit function.

3.1 Description of Data

The original data set consists of 424 subjects in which 312 subjects participated in the RCT and 112 subjects were followed but were not randomized after the trial.

Table 3.1: *Variable Descriptions in PBC Data.* This table consists of variables that we are interested in. Note that there are other variables in the data set, but are omitted in this table.

Variable Name	Description
id	Subject ID
status	Results at the end of the study: 0 = alive, 1 = undergone a liver transplant, 2 = dead
trt	The treatment received: 1 = D-penicillamine, 2 = placebo
age	Age in days
sex	Sex: 0 = male, 1 = female
edema	Whether or not patient has edema: 0 = no edema and no diuretic therapy for edema, 0.5 = edema present without diuretics, or edema resolved by diuretics, 1 = edema despite diuretic therapy
bili	Bilirubin serum level (mg/dl)
albumin	Albumin level (mg/dl)
protime	Prothrombin time (seconds)
stage	Histologic stage of disease
time	number of days between registration and the earlier of death, transplantion, or study analysis in July, 1986

Due to the inconsistency and the missing information on the treatment variable, we will not be utilizing these additional subjects. Subjects who participated in this trial were recruited from January 1974 to May 1984.

In previous studies, researchers have identified a few prognostic variables for PBC. These variables include: age, bilirubin value, albumin value, prothrombin time and edema. There are also studies suggesting that laboratory measures including serum bilirubin levels, enlargement of the liver (hepatomegaly), accumulation of water in the legs (edema), and visible veins in the chest and shoulders (spiders) are signs of liver damage (Therneau and Grambsch, 2000). All these variables are described within the Variable Description Table below.

3.2 Benefit Function

An accelerated failure time (AFT) model is computed as our prediction model in which we use it for our benefit function. Briefly speaking, the AFT model is used to predict the whether the effect of a risk factor will accelerate or decelerate the survival time. Hence, our benefit values are equivalent to the predicted values from this AFT model. We believe that a prediction model on survival time based on known risk factors is suitable as a benefit function in this case study.

The benefit model used is determine by a few variables that are not affected by the disease but are known to be potentional risk factors to the disease. In this case, we find that age and sex are significant predictors of disease progression. The treatment variable is needed despite its lack of significance. In addition, we include an interaction effect of treatment and sex.

The fitted AFT model is shown below. We have used survreg() function in the Rpackage: survival.

$$\widehat{log(SurvivalTimeinDays)} = 8.875 + 0.649 * trt + 1.450 * sex - 0.033 * age - 0.794 * trt * sex$$

where sex is an indicator function such that 0 = male and 1 = female

Notice that through this model, the treatment increases the survival time in men. However, the model also suggests that the treatment decreases the survival time in women. For patients who do not receive treatment, the model suggests that women have a longer survival time than men adjusting for age. In addition, we see that a yearly increase in age will decrease in survival time as well. In summary, our risk factors for this case study are age and sex, in which males are of higher risk than females.

Using the fitted AFT model, we compute the benefit values by this definition:

$$benefit_i = \widehat{log(SurvivalTimeinDays)}_{i|trt_i=0} - \widehat{log(SurvivalTime)}_{i|trt_i=1}$$

$benefit_i = 0.649 - 0.794 * Sex_i$ where *Sex* is an indicator for Female patients

3.3 Biomarker Model

Similar to constructing the benefit model, a prediction model is used to as a function to compute the biomarker values. Due to the nature of the study, our values in the data set are baseline clinical values which are collected in the pre-treatment phase.

There are three biomarker functions we are comparing to the benefit model we have defined in the previous section. This is to demonstrate the different RIC curves produced given the same benefit model.

- **Biomarker 1:** Biomarker 1 is chosen such that variables are highly associated with the survival time and PBC. Through preliminary analysis, we find that Sex and Age fulfill the criteria. The two variables are also easily obtained without any intensive or invasive process of clinical work. Biomarker 1 is then defined as: $\widehat{biomarker}_i = log(\widehat{SurvivalTimeinDays})_i = 9.782 + 0.315 * Sex_i - 0.033 * Age_i$ where *Sex* is an indicator for Female patients

- **Biomarker 2:** To compare Biomarker 1, we chose to include a clinical variable in addition to Biomarker 1. For Biomarker 2, we will add in the measures of Bilirubin through blood draws. Biomarker is defined as: $\widehat{biomarker}_i = log(\widehat{SurvivalTimeinDays})_i = 9.653 + 0.318 * Sex_i - 0.026 * Age_i - 0.102 * Bilirubin_i$ where *Sex* is an indicator for Female patients

- **Biomarker 3:** To look into the differences of how the first two biomarkers differ from a random variable that is irrelevant to the outcome measurements, we chose time as our biomarker. Time corresponds to the number of days between registration to the earliest of death, transplantation or end of study. Because subjects are recruited throughout the study, the shorter number of days do not correspond to the severity of outcomes. Biomarker 3 is then defined as: $\widehat{biomarker}_i = Time_i$

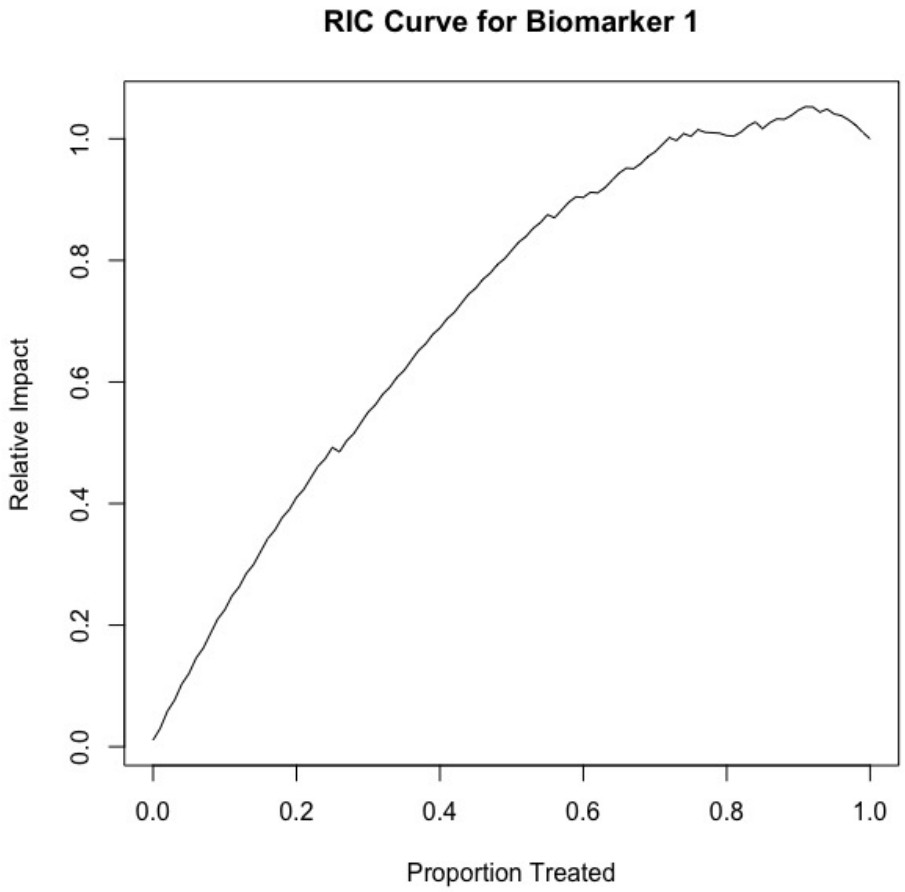

Figure 3.1: . The RIC curve under Biomarker 1 shows a curvature indicating that the biomarker chosen seem to be better than random assignment. In this case, we see that at a proportion of approximately 0.8 treated, the relative impact is similar to the entire population treated. In a similar fashion, treating the top 60% of patients in the population will compute a relative impact of approximately 0.8.

3.4 Results

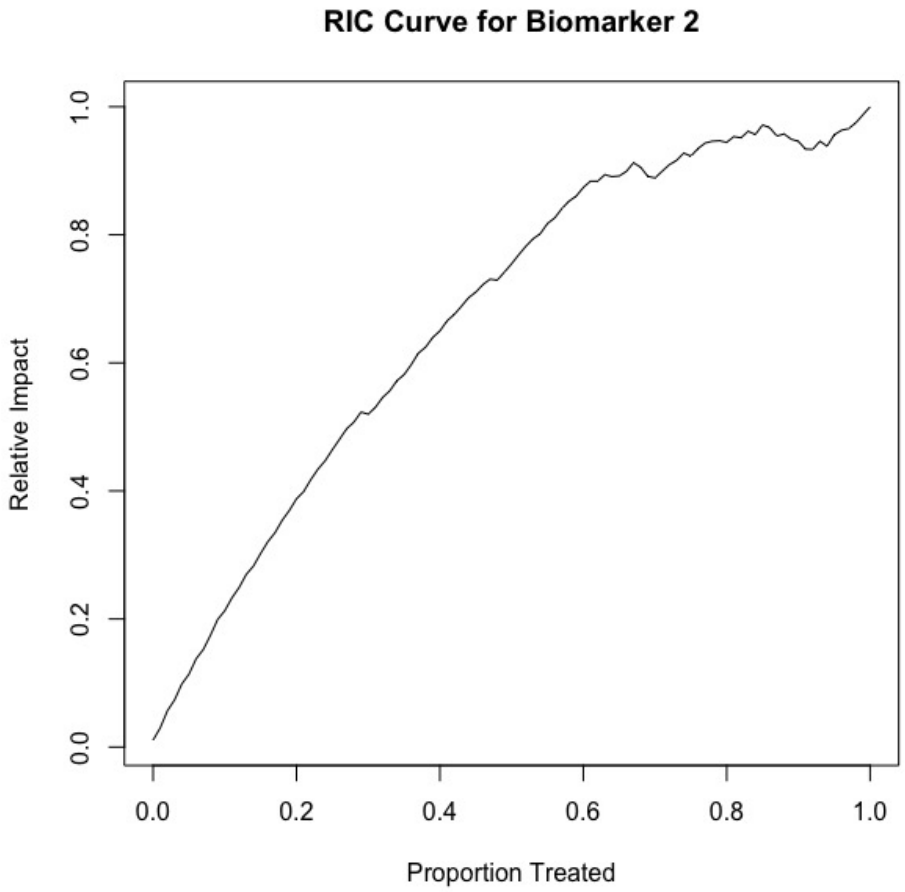

Figure 3.2: . From what we observe, the RIC seem to show a similar curve between Biomarker 1 and 2. Due to the similarity and the ease to obtain Biomarker 1 values, it seems that Biomarker 1 is the preferred biomarker in comparison to the two.

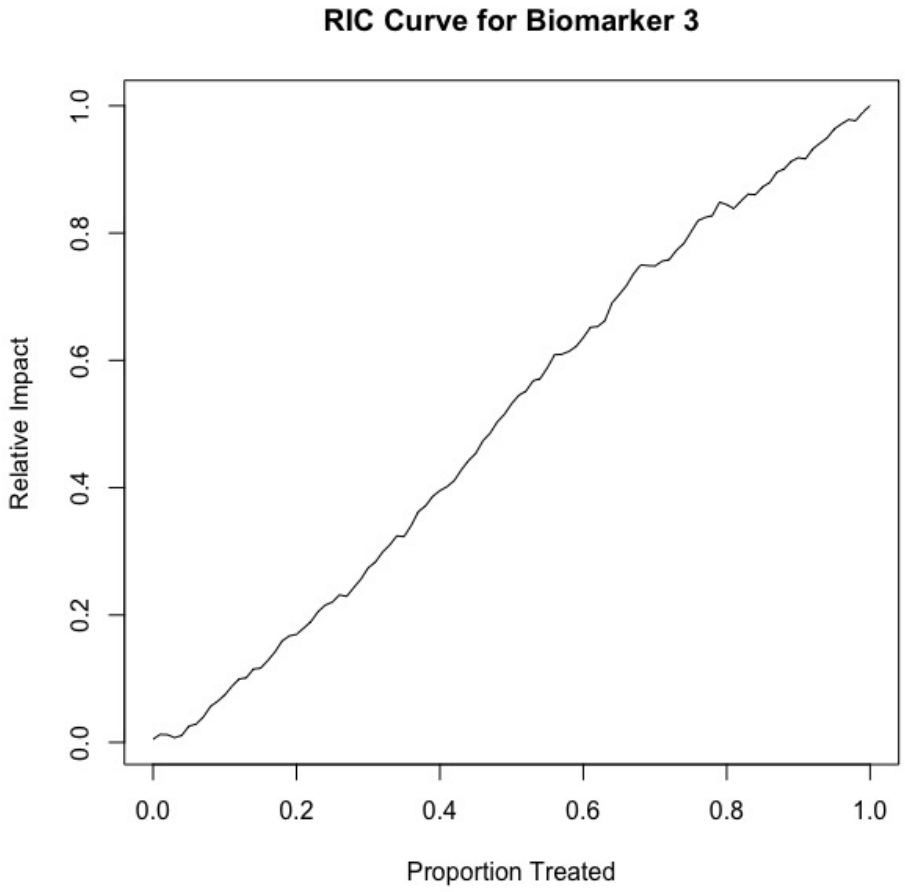

Figure 3.3: . From what we observe, the RIC curve for Biomarker 3 behaves differently from Biomarkers 1 and 2. Instead, we see a diagonal line that shows no curvature, indicating that assigning treatment based on this biomarker is no better than assigning treatment at random. This observation seem to agree with the assumption that the time between registration until the end of death, liver implantation, or end of study is not the best factor to indicate which proportion of patients should be treated.

Chapter 4

Discussion and Conclusion

In summary, the main goal of this project is to discover and understand how various components in model assumptions and functions affect the variabilities on RIC results. Using simulation study, we observed how distributions of AUC curves differ among each other solely by setting biomarker and benefit functions to be true or estimated functions. With the results in the simulation study, we find that the two regression assumptions, linear and poisson regression, produced different variabilities among the different settings of the AUCs.

Similarly, the case study that uses PBC data (Murtaugh et al., 1994) have observed different RIC curves on the three unique biomarker functions. This ultimately led to our conclusion that a biomarker should be chosen sensibly and not by random.

There are definitely a few advantages and disadvantages in this report, including the simulation data set and data set that was used in the case study. The simulation data set was limited by the assumptions of linear regression model and poisson model as prediction models. Although these two models are straightforward and commonly used in the clinical health setting, it will very helpful to understand where the variability comes in on other types of prediction models. Consequently, re-running the simulation study with a data set that follows a different model assumption is suggested.

A case study that would be even more interesting to work on may involve more variety of variables to choose from. Since this case study uses one single benefit

function, it would be interested to implement multiple benefit functions using a single biomarker function. Results of RIC curves may or may not be similar - the comparison would be an interesting find. One could also possibly take this analysis further and apply cross validations to further understand the behaviour of RIC curves.

Another interesting analysis would be using a data set from a different field. Although RIC method was inpsired in a clinical health setting, it could also be extended to other areas of the field. Further discussion and background expertise are needed in order to implement this method.

Bibliography

[1] Primary biliary cholangitis (primary biliary cirrhosis). *National Institute of Diabetes and Digestive and Kidney Diseases*.

[2] E. R. Dickson, P. M. Grambsch, T. R. Fleming, L. D. Fisher, and A. Langworthy. Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, 10(1):1–7, 1989.

[3] Marie-Hlne Disatnik, Amit U. Joshi, Nay L. Saw, Mehrdad Shamloo, Blair R. Leavitt, Xin Qi, and Daria Mochly-Rosen. Potential biomarkers to follow the progression and treatment response of huntingtons disease. *The Journal of Experimental Medicine*, 213(12):26552669, Jul 2016.

[4] Thomas R. Fleming and David P. Harrington. *Counting Processes and Survival Analysis*. John Wiley & Sons, 1991.

[5] P. A. Murtaugh, E. R. Dickson, G. M. Van Dam, M. Malinchoc, P. M. Grambsch, A. L. Langworthy, and C. H. Gips. Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*, 20(1 Pt 1):126–134, Jul 1994.

[6] Maen Obeidat, Ke Hao, Yohan Boss, David C Nickle, Yunlong Nie, Dirkje S Postma, Michel Laviolette, Andrew J Sandford, Denise D Daley, James C Hogg, and et al. Molecular mechanisms underlying variations in lung function: a systems genetics analysis. *The Lancet Respiratory Medicine*, 3(10):782795, 2015.

[7] William H. Robinson and Rong Mao. Biomarkers to guide clinical therapeutics in rheumatology? *Current Opinion in Rheumatology*, 28(2):168175, 2016.

[8] Mohsen Sadatsafavi, Don D. Sin, Paul Gustafson, and Zafar Zafari. Relative impact characteristic (ric) curve: a graphical tool to visualize and quantify the population-level consequences of implementing markers. *In Review*.

[9] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.

[10] Terry M Therneau. *A Package for Survival Analysis in S*, 2015. version 2.38.

Appendix A

Supporting Materials

A.1 R Code on Simulation Study

```
########################
####### RIC_exp function
########################

library(DescTools)
library(data.table)

#RIC_exp is a function used to plot the RIC curve on simulated dataset

RIC_exp <- function(proportion, biomarker, benefit){

  #Grabbing biomarker threshold such that it matches the proportion, in this case,
  #our biomarker is y_no_trt
  k <- quantile(biomarker, probs = 1 - proportion)

  #Creating treatment vectors such that patients are assigned to treatment if biomarker
  #is greater than threshold
  new_trt <- sapply(k, function (x) as.numeric(biomarker >= x))

  #relative benefit
  b1 <- apply(new_trt, 2, function(x) sum((benefit)*x)/
              sum(benefit))

  #return relative benefit
  b1
```

```
}

################################################
###### Simulation for Multiple Linear Regression
################################################

#Simulation on one sample data for biomarker function data with n subjects and two risk factors.

n <- 50
int <- 0
b1 <- 1
b2 <- 0.5
b3 <- -0.25
b4 <- -0.5
b5 <- -0.25

#INITIATE a vector of aucs
auc_vec_lm_1<-c()
auc_vec_lm_2<-c()
auc_vec_lm_3<-c()
auc_vec_lm_4<-c()

for (i in 1: 100){

set.seed(i)
x1 <- runif(n, 0, 10)
x2 <- rnorm(n, 10, sd=6)
e <- rnorm(n,sd=5)

#Treatment was randomly assigned
trt <- rbinom(n, size=1, prob=0.5)

#if biomarker and benefit is calculated  with errors
y_err <- int + b1*x1 + b2*x2 + b3*trt + I(b4*trt*x1) + I(b4*trt*x2) + e

#Make a dataframe out of the whole data
mat_err = data.frame(x1=x1, x2=x2, trt=trt, y=y_err)

#Prediction model
lm_err <- lm(y ~ x1 + x2 + trt + I(trt*x1) + I(trt*x2), data=mat_err)

#NOTE: Marker values were calculated as the predicted number of exacerbations
#without treatment by setting,
#for all individuals, the treatment variable to 0. (Sadatsafavi, In Review)

#Creating data set that has all trt variable as 0
no_trt_data <- as.data.frame(cbind(mat_err[,!names(mat_err) %in% "trt"], trt = 0))
```

```
#Creating data set that has all trt variable as 1
all_trt_data <- as.data.frame(cbind(mat_err[,!names(mat_err) %in% "trt"], trt = 1))

#Biomarker is the predicted y values when treatment is 0
y_hat_no_trt <- predict(lm_err, newdata = no_trt_data)

#Predict outcome when all trt variable is 1
y_hat_all_trt <- predict(lm_err, newdata = all_trt_data)

#Predict Benefit
benefit_predict <- y_hat_no_trt - y_hat_all_trt

#True y when treatment is 0 (true biomarker)
y_true_no_trt <- int + b1*x1 + b2*x2
y_true_all_trt <- int + b1*x1 + b2*x2 + b3 + I(b4*x1) + I(b4*x2)
benefit_true <- y_true_no_trt - y_true_all_trt

#Case 1: both biomarker and benefits are predicted

#CALCULATING AUC
ric_x <- seq(0, 1, by = 0.01)
ric_y_1 <- RIC_exp(proportion=ric_x, biomarker = y_hat_no_trt, benefit = benefit_predict)
auc_vec_lm_1 <- append(auc_vec_lm_1, AUC(ric_x,ric_y_1))

#Case 2: predicted biomarker and true benefit

#CALCULATING AUC
ric_y_2 <- RIC_exp(proportion=ric_x, biomarker = y_hat_no_trt, benefit = benefit_true)
auc_vec_lm_2 <- append(auc_vec_lm_2, AUC(ric_x,ric_y_2))

#Case 3: true biomarker and predicted benefit

#CALCULATING AUC
ric_y_3 <- RIC_exp(proportion=ric_x, biomarker = y_true_no_trt, benefit = benefit_predict)
auc_vec_lm_3 <- append(auc_vec_lm_3, AUC(ric_x,ric_y_3))

#Case 4: both biomarker and benefits is true

#CALCULATING AUC
ric_y_4 <- RIC_exp(proportion=ric_x, biomarker = y_true_no_trt, benefit = benefit_true)
auc_vec_lm_4 <- append(auc_vec_lm_4, AUC(ric_x,ric_y_4))

cat("Done:", i,"\n")

}
```

```
#Summary of AUC values on simulated data
cat("Summary statistics on AUC values for Case 1 data \n",
    "Mean AUC: ", mean(auc_vec_lm_1), "\n",
    "Median AUC: ", median(auc_vec_lm_1), "\n",
    "SD on AUC: ",sd(auc_vec_lm_1), "\n",
    "50% range: ", quantile(auc_vec_lm_1,0.25), " to ",quantile(auc_vec_lm_1,0.75),"\n",
    "Range on AUC", min(auc_vec_lm_1), " to ", max(auc_vec_lm_1))

jpeg("Rplot_LM_1.jpeg")
plot(density(auc_vec_lm_1), main = "Density Plot of AUCs in Case 1", xlab="AUC values",
  xlim=c(0.4,0.9), ylim=c(0,25))
dev.off()

cat("Summary statistics on AUC values for Case 2 data \n",
    "Mean AUC: ", mean(auc_vec_lm_2), "\n",
    "Median AUC: ", median(auc_vec_lm_2), "\n",
    "SD on AUC: ",sd(auc_vec_lm_2),
    "50% range: ", quantile(auc_vec_lm_2,0.25), " to ",quantile(auc_vec_lm_2,0.75),"\n",
    "Range on AUC", min(auc_vec_lm_2), " to ", max(auc_vec_lm_2))

jpeg("Rplot_LM_2.jpeg")
plot(density(auc_vec_lm_2), main = "Density Plot of AUCs in Case 2", xlab="AUC values",
  xlim=c(0.4,0.9), ylim=c(0,25))
dev.off()

cat("Summary statistics on AUC values for Case 3 data \n",
    "Mean AUC: ", mean(auc_vec_lm_3), "\n",
    "Median AUC: ", median(auc_vec_lm_3), "\n",
    "SD on AUC: ",sd(auc_vec_lm_3),
    "50% range: ", quantile(auc_vec_lm_3,0.25), " to ",quantile(auc_vec_lm_3,0.75),"\n",
    "Range on AUC", min(auc_vec_lm_3), " to ", max(auc_vec_lm_3))

jpeg("Rplot_LM_3.jpeg")
plot(density(auc_vec_lm_3), main = "Density Plot of AUCs in Case 3", xlab="AUC values",
  xlim=c(0.4,0.9), ylim=c(0,25))
dev.off()

cat("Summary statistics on AUC values for Case 4 data \n",
    "Mean AUC: ", mean(auc_vec_lm_4), "\n",
    "Median AUC: ", median(auc_vec_lm_4), "\n",
    "SD on AUC: ",sd(auc_vec_lm_4),
    "50% range: ", quantile(auc_vec_lm_4,0.25), " to ",quantile(auc_vec_lm_4,0.75),"\n",
    "Range on AUC", min(auc_vec_lm_4), " to ", max(auc_vec_lm_4))

jpeg("Rplot_LM_4.jpeg")
plot(density(auc_vec_lm_4), main = "Density Plot of AUCs in Case 4", xlab="AUC values",
  xlim=c(0.4,0.9), ylim=c(0,25))
```

```
dev.off()

################################################
###### Simulation for Poisson Regression
################################################
#Simulation on one sample data for biomarker function data with n subjects and two risk factors.

n <- 20
int <- 0.2
b1 <- 0.1
b2 <- 0.25
b3 <- -0.25
b4 <- -0.05
b5 <- -0.125


#INITIATE a vector of aucs
auc_vec_glm_1<-c()
auc_vec_glm_2<-c()
auc_vec_glm_3<-c()
auc_vec_glm_4<-c()

for (i in 1: 100){

  set.seed(i)
  x1 <- runif(n, 0, 10)
  x2 <- rnorm(n, 10, 6)

  #Treatment was randomly assigned
  trt <- rbinom(n, size=1, prob=0.5)

  #Biomarker and benefit are with errors
  y_err <- rpois(n=n, lambda = exp(int + b1*x1 + b2*x2 + b3*trt + I(b4*trt*x1) + I(b4*trt*x2)))

  #Make a dataframe out of the whole data
  mat_err = data.frame(x1=x1, x2=x2, trt=trt, y=y_err)

  #Prediction model
  glm_err <- glm(y ~ -1 + x1 + x2 + I(trt*x1) , family = "poisson", data=mat_err)

  #Creating data set that has all trt variable as 0
  no_trt_data <- as.data.frame(cbind(mat_err[,!names(mat_err) %in% "trt"], trt = 0))

  #Creating data set that has all trt variable as 1
  all_trt_data <- as.data.frame(cbind(mat_err[,!names(mat_err) %in% "trt"], trt = 1))

  #Biomarker is the predicted y values when treatment is 0
```

```
  y_hat_no_trt <- predict(glm_err, newdata = no_trt_data, type = "response")

  #Predict outcome when all trt variable is 1
  y_hat_all_trt <- predict(glm_err, newdata = all_trt_data, type = "response")

  #Predict benefit
  benefit_predict <- y_hat_no_trt - y_hat_all_trt

  #True benefit
  y_true_no_trt <- exp(b1*x1 + b2*x2)
  y_true_all_trt <- exp(b1*x1 + b2*x2 + b3 + b4*x1)
  benefit_true <- y_true_no_trt - y_true_all_trt

  #Case 1: both biomarker and benefits are predicted

  #CALCULATING AUC
  ric_x <- seq(0, 1, by = 0.01)
  ric_y_1 <- RIC_exp(proportion=ric_x, biomarker = y_hat_no_trt, benefit = benefit_predict)
  auc_vec_glm_1 <- append(auc_vec_glm_1, AUC(ric_x,ric_y_1))

  #Case 2: predicted biomarker and true benefit

  ric_y_2 <- RIC_exp(proportion=ric_x, biomarker = y_hat_no_trt, benefit = benefit_true)
  auc_vec_glm_2 <- append(auc_vec_glm_2, AUC(ric_x,ric_y_2))

  #Case 3: true biomarker and predicted benefit

  #CALCULATING AUC
  ric_y_3 <- RIC_exp(proportion=ric_x, biomarker = y_true_no_trt, benefit = benefit_predict)
  auc_vec_glm_3 <- append(auc_vec_glm_3, AUC(ric_x,ric_y_3))

  cat("Done:", i,"\n")

  #Case 4: both biomarker and benefits is true

  #CALCULATING AUC
  ric_y_4 <- RIC_exp(proportion=ric_x, biomarker = y_true_no_trt, benefit = benefit_true)
  auc_vec_glm_4 <- append(auc_vec_glm_4, AUC(ric_x,ric_y_4))

}

par(mfrow=c(2,2))

#Summary of AUC values on simulated data
cat("Summary statistics on AUC values for Case 1 data \n",
    "Mean AUC: ", mean(auc_vec_glm_1), "\n",
    "Median AUC: ", median(auc_vec_glm_1), "\n",
```

```
    "SD on AUC: ",sd(auc_vec_glm_1), "\n",
    "50% range: ", quantile(auc_vec_glm_1,0.25), " to ",quantile(auc_vec_glm_1,0.75),"\n",
    "Range on AUC", min(auc_vec_glm_1), " to ", max(auc_vec_glm_1))


jpeg("poisson_case1.jpg")
plot(density(auc_vec_glm_1), main = "Density Plot of AUCs in Case 1", xlab="AUC values",
  xlim=c(0.5,1), ylim=c(0,10))
dev.off()

cat("Summary statistics on AUC values for Case 2 data \n",
    "Mean AUC: ", mean(auc_vec_glm_2), "\n",
    "Median AUC: ", median(auc_vec_glm_2), "\n",
    "SD on AUC: ",sd(auc_vec_glm_2), "\n",
    "50% range: ", quantile(auc_vec_glm_2,0.25), " to ",quantile(auc_vec_glm_2,0.75),"\n",
    "Range on AUC", min(auc_vec_glm_2), " to ", max(auc_vec_glm_2))

jpeg("poisson_case2.jpg")
plot(density(auc_vec_glm_2), main = "Density Plot of AUCs in Case 2", xlab="AUC values",
  xlim=c(0.5,1), ylim=c(0,10))
dev.off()

cat("Summary statistics on AUC values for Case 3 data \n",
    "Mean AUC: ", mean(auc_vec_glm_3), "\n",
    "Median AUC: ", median(auc_vec_glm_3), "\n",
    "SD on AUC: ",sd(auc_vec_glm_3), "\n",
    "50% range: ", quantile(auc_vec_glm_3,0.25), " to ",quantile(auc_vec_glm_3,0.75),"\n",
    "Range on AUC", min(auc_vec_glm_3), " to ", max(auc_vec_glm_3))

jpeg("poisson_case3.jpg")
plot(density(auc_vec_glm_3), main = "Density Plot of AUCs in Case 3", xlab="AUC values",
  xlim=c(0.5,1), ylim=c(0,10))
dev.off()

cat("Summary statistics on AUC values for Case 4 data \n",
    "Mean AUC: ", mean(auc_vec_glm_4), "\n",
    "Median AUC: ", median(auc_vec_glm_4), "\n",
    "SD on AUC: ",sd(auc_vec_glm_4), "\n",
    "50% range: ", quantile(auc_vec_glm_4,0.25), " to ",quantile(auc_vec_glm_4,0.75),"\n",
    "Range on AUC", min(auc_vec_glm_4), " to ", max(auc_vec_glm_4))

jpeg("poisson_case4.jpg")
plot(density(auc_vec_glm_4), main = "Density Plot of AUCs in Case 4", xlab="AUC values",
  xlim=c(0.5,1), ylim=c(0,10))
dev.off()
```

A.2 R Code on Case Study

```
library(survival)
library(tidyverse)

dat <- na.omit(pbc[pbc$status!=1,])
dat$status[dat$status==2]<-1

#RIC_exp is a function used to plot the RIC curve on simulated dataset

RIC_exp <- function(proportion, biomarker, benefit){
  k <- quantile(biomarker, probs = 1 - proportion)
  new_trt <- sapply(k, function (x) as.numeric(biomarker >= x))
  b1 <- apply(new_trt, 2, function(x) sum((benefit)*x)/
              sum(benefit))
  b1
}

mod5 <- survreg(Surv(time, status) ~ trt * sex + age , data=dat)
summary(mod5)

mod6 <- survreg(Surv(time, status) ~ sex + age, data=dat)
summary(mod6)

mod7 <- survreg(Surv(time, status) ~ sex + age + bili , data=dat)
summary(mod7)

#Plot the AUC
curve(RIC_exp(x, bio_mod6, ben_mod5),0,1, xlab = "Proportion Treated", ylab = "Relative Impact",
  main = "RIC Curve for Biomarker 1")
curve(RIC_exp(x, bio_mod7, ben_mod5),0,1, xlab = "Proportion Treated", ylab = "Relative Impact",
  main = "RIC Curve for Biomarker 2")
curve(RIC_exp(x, dat$time, ben_mod5),0,1, xlab = "Proportion Treated", ylab = "Relative Impact",
  main = "RIC Curve for Biomarker 3")

#biomarker
bio_mod6 <- predict(mod6, type="response")
bio_mod7 <- predict(mod7, type="response")
bio_mod8 <- predict(mod8, type="response")

#E[y|x=1]
dat_all_trt <- dat[,names(dat) %in% c("time", "status", "sex", "age")]
dat_all_trt$trt <- 1
dat_all_trt$new <- predict(mod5, newdata = dat_all_trt, type="response")
```

```
#E[y|x=0]
dat_no_trt <- dat[,names(dat) %in% c("time", "status", "sex", "age")]
dat_no_trt$trt <- 0
dat_no_trt$new <- predict(mod5, newdata = dat_no_trt, type="response")

#benefit vector
ben_mod5 <- dat_no_trt$new - dat_all_trt$new
```