Strategies and Considerations for Receiver Operating Characteristic (ROC) Regression Modeling

By

Nathaniel Patrick Dowd

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

May 10, 2024

Nashville, Tennessee

Approved:

Andrew J. Spieker, Ph.D

Bryan Blette, Ph.D

ACKNOWLEDGEMENTS

I would like to thank Dr. Andrew Spieker for being my advisor on this thesis project; his patience and knowledge of statistics made it easier for me to take on a large project like this and never feel like I was out of my depth. Having Dr. Spieker as not only an advisor, but also an instructor for multiple classes throughout my time at Vanderbilt, has helped me develop a solid foundation in both practical applications and theory that have helped on this project immensely.

I would like to thank Dr. Robert Greevy for admitting me into the biostatistics program and believing in my ability to learn and grow as a student, while also giving me important guidance to me as I navigated whatever came my way. I would also like to thank all the teachers I have had in each of my courses here who have been so valuable in my development as a student, and who never hesitated to lend their knowledge to me whenever I would have a question. Additionally, I would like to thank Dr. Bryan Blette for contributions and invaluable feedback on my thesis.

Lastly, I would like to thank my family, my mom, dad, and two brothers, whose support through not only this master's program but through my entire academic career, has been instrumental in making me the student and man I am today.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction to ROC Curves

Receiver operating characteristic (ROC) curves are commonly used to evaluate the quality of diagnostic tests. The most widespread uses of ROC curves in medical scenarios involve evaluating tests based on continuous outcome measures that classify a patient as positive or negative for some disease at some sensible numerical cutoff in the range of test results (without loss of generality, we can suppose any values below such a cutoff would represent a "positive" test, while others represent a "negative" test). At any given cutoff point, we can calculate the test's sensitivity (i.e., true positive rate) and the specificity (i.e., true negative rate). Different cutoff points across the range of the test results can yield different values for the sensitivity and specificity. At a given cutoff point, the sensitivity can be estimated as the proportion of the diseased individuals in the data that are correctly categorized as positive; the specificity can be estimated as the proportion of healthy individuals correctly classified as negative. The ROC curve is derived by acknowledging all possible cutoff points, with the estimated false positive rate (FPR, which is one minus the specificity) on the *x*-axis and estimated sensitivity on the *y*-axis. There are two fundamental points that are featured on every empirical ROC curve, described as follows under the convention that lower values are less desirable: selecting a cutoff point lower than the minimum test result (i.e., classifying each result as negative), which corresponds to an estimated sensitivity of zero, and an estimated specificity of one; and selecting a cutoff point exceeding the maximum test result, which corresponds to an estimated sensitivity of one and an estimated specificity of zero. The area under the curve (AUC) provides a

convenient way of summarizing the sensitivity and specificity of a test across all possible cutoff points with a single number (that is, it is a function of the ROC curve).

1.2 ROC Regression

Regression methods for continuous outcomes (e.g. height, weight, GDP, etc.) or binary outcomes (sick/healthy, dead/alive) are widely used. If one is willing to embed certain assumptions into these models (namely, linearity structures), linear combinations of coefficients can be used to identify key contrasts such as mean differences and odds ratios). ROC regression, on the other hand, is a less frequently utilized regression method in which we can evaluate the effect of a covariate on the discriminating capacity of a diagnostic test (more broadly, stochastic ordering). There are several variants of ROC regression modeling described in the literature: for instance, parametric modeling of the ROC function or non-parametric estimation of its AUC. While the coefficients originating from such models do not possess known straightforward interpretations absent parametric assumptions, tests of these coefficients can inform us as to how useful a biomarker is as a classifier across different subgroups.

1.3 Background on SARS-CoV-2 Vaccination in Solid Organ Transplant Recipients

The COVID-19 pandemic was declared in March of 2020, and vaccines were undergoing preliminary approval later that year. COVID-19 refers to the disease caused by the SARS-CoV-2 virus, which is considered very contagious. An observational vaccine immunogenicity study was previously conducted by Yanis et. al, which enrolled solid organ transplant (SOT) recipients and healthy controls (HCs) from December 18th, 2020 to March 7th, 2021 [1]. As SOT recipients are

typically on immunosuppressive regimens, it is of interest to compare vaccine immunogenicity in this more vulnerable population to that of HCs.

Study participants were required to be eligible to receive two doses of the BNT162b2 SARS-CoV-2 vaccine at the Vanderbilt University Medical Center (VUMC). Exclusion criteria included age younger than 55 years, as well as a positive test for COVID-19 prior to the study's start.

To collect data, the research group scheduled three visits with both HCs and SOT recipients. The first visit took place 0 to 2 days before the first vaccine dose, the second visit took place 21-42 days after the first vaccine dose, and the final visit for collection took place 21-42 days after the second vaccine dose. At each visit, blood samples were taken from each participant, and immune response is measured by enzyme linked immunosorbent assay (ELISA) [8]. This technique was used to measure three forms of immune response; immunoglobulin G (IgG) to SARS-CoV-2 spike receptor-binding domain (RBD), spike extracellular domain (ECD), and nucleocapsid protein (N). We make use of a simulated version of this data set to protect patient information; the data were simulated in a way such that the overall patterns reflect the actual study results reasonably closely.

Importantly, RBD and ECD are specifically targeted by the vaccine; the hypothesis—at least among HCs—is that IgG to RBD and ECD would meaningfully rise following vaccinations, but not IgG to N [8]. The anticipated rise in IgG to RBD and ECD levels among SOT recipients is less clear. However, unlike RBD and ECD, N is not a specific target of the vaccine, so elevated

IgG to N may be indicative of SARS-CoV-2 exposure. For example, if high levels of N are present at baseline (before any vaccination), this may suggest immune response (possibly from asymptomatic infection) prior to the start of the study that is unrelated to the administration of the vaccine. To determine if this was a significant issue in this dataset, the baseline IgG to N was evaluated for the 80 patients included in the study. Figure 1.1 shows a histogram of baseline IgG to N in the simulated data set.

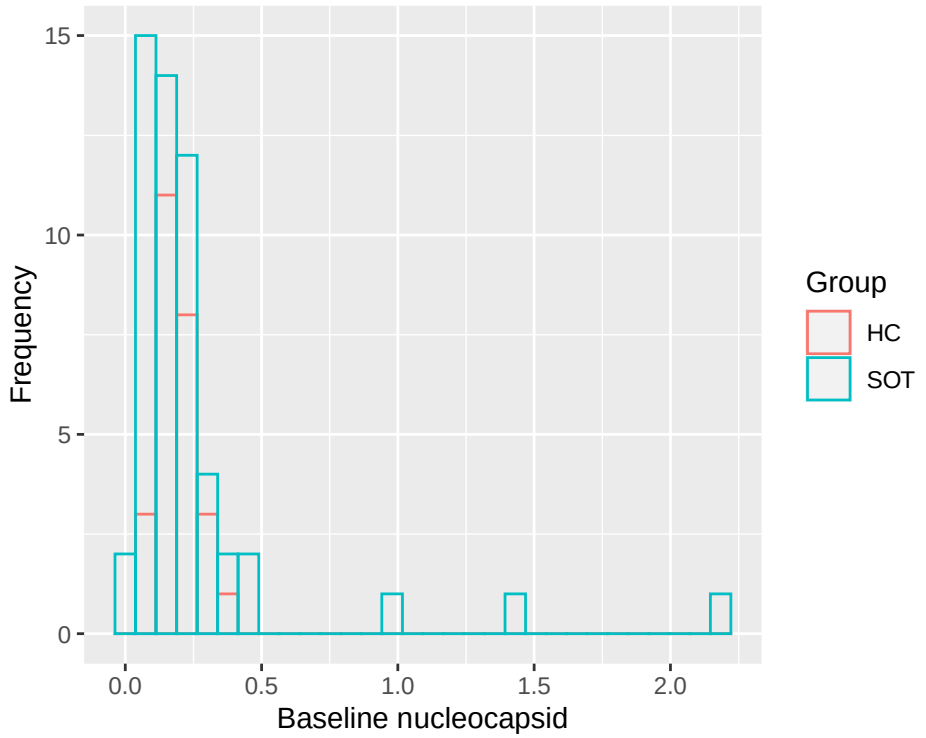

Figure 1.1: Histogram of IgG to N at baseline.

From visual evaluation of the histogram, we note three observations that have high IgG levels to nucleocapsid before vaccination. These observations were excluded from the main study but included as a sensitivity analysis in which the fundamental conclusions did not change. We elect to include these observations in our exploration of the simulated data.

A key finding of this study was that both healthy controls and the SOT recipients exhibited an increased mean humoral response after both vaccine doses. Between the two groups, SOT recipients showed lower mean antibody response compared to HCs, particularly after the first dose. Among SOT recipients, there was a significant association between the antibody response and the patient's immunosuppressant regimen.

The analyses conducted in the original study primarily involved assessment of mean differences. We seek to explore an alternative contrast—namely, one of stochastic ordering—using ROC methods as a way of investigating the degree to which the antibody responses differ between SOT recipients and HCs. We focus specifically on IgG to RBD following the second vaccine dose, which we will notate for the rest of the report as RBD_2. The distribution of RBD_2 is shown in Figure 1.2 for SOT recipients and HCs.

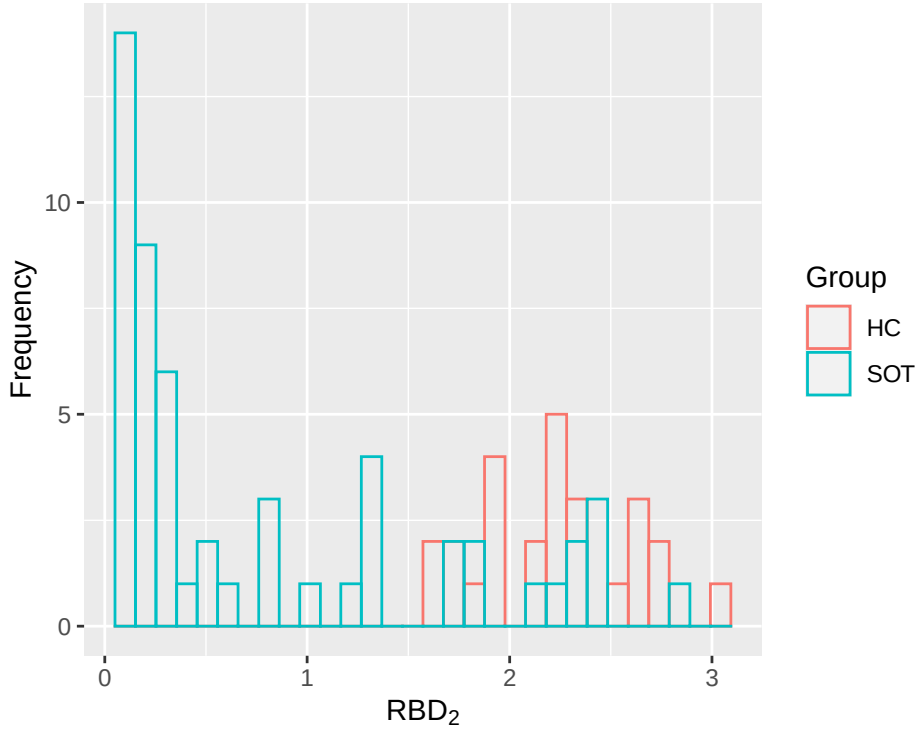

Figure 1.2: Histogram of RBD_2 among HCs and SOT recipients.

Regression of the ROC curve will allow us to quantify how effective RBD_2 is as a classifier depending, as an example, on time post-transplant. When fitting the model, note that time post-transplant is a fact that applies only to SOT recipients. Our modeling approach will take this into account.

In addition, a crucial fact that we will elaborate on later is that in many cases, higher values of a biomarker indicate a worse health outcome. Such examples include hemoglobin A1c percentage, systolic blood pressure, and prostate-specific antigen. Many existing presentations of ROC methods operate under this convention. In contrast, antibody levels are such that higher values signify stronger immune response (so that lower values are less desirable). Our presentation of the methodology follows the latter convention without loss of generality.

CHAPTER 2

ROC Estimation Techniques

2.1 The ROC Curve as a Function of the FPR

In this section, we define the mathematical framework more thoroughly, using the SOT-COVID data as an anchoring example. Note that the sensitivity of a test is a function of the FPR. Therefore, we can index the ROC curve as a function of p, the FPR, such that the corresponding sensitivity is denoted as $\mathrm{ROC}(p)$. Let $F_1(c) = P(Y_1 < c)$ denote the cumulative distribution function (CDF) associated with the SOT recipients; following the convention that higher values are more desirable, this is simply the sensitivity of a test based on a cutoff of c. Now, a cutoff point can be conceptualized as some quantile of the distribution among HCs. That is to say that the cutoff point function is given as $F_0^{-1}(p)$, where F_0^{-1} represents the inverse CDF (i.e., quantile function) for the HCs. Putting this together, the ROC function can be recognized as the following composition of functions [2]:

$$\mathrm{ROC}(p) = F_1\left(F_0^{-1}(p)\right).$$

In the sections that follow, we elaborate on different modeling strategies for the ROC curve. The simplest, and most widely used approach for estimation of the ROC curve is an empirical estimate (i.e., one that does not assume either group to conform to an underlying distribution having a specific parametric form), which is formed by computing the estimated specificity and sensitivity values at each unique cutoff point as simple proportions, and representing these values derived from the data graphically. Other approaches include parametric and semiparametric modeling, all of which will be discussed in this section.

2.2 Empirical ROC Estimation

The empirical ROC curve, as previously stated, does not invoke assumptions regarding the underlying distributions of the test results among HCs or SOT recipients; rather, it is estimated by estimating the sensitivity and specificity at each possible cutoff point. Owing to the finite sample, this curve will be reflected by a step function. The number of "upward" steps (starting from the origin) for the empirical method of estimation cannot exceed the number of SOT recipients, and the number of "rightward" steps (starting from the origin) cannot exceed the number of HCs. Figure 2.1 presents the empirical ROC curve for RBD_2 in the SOT-COVID study.

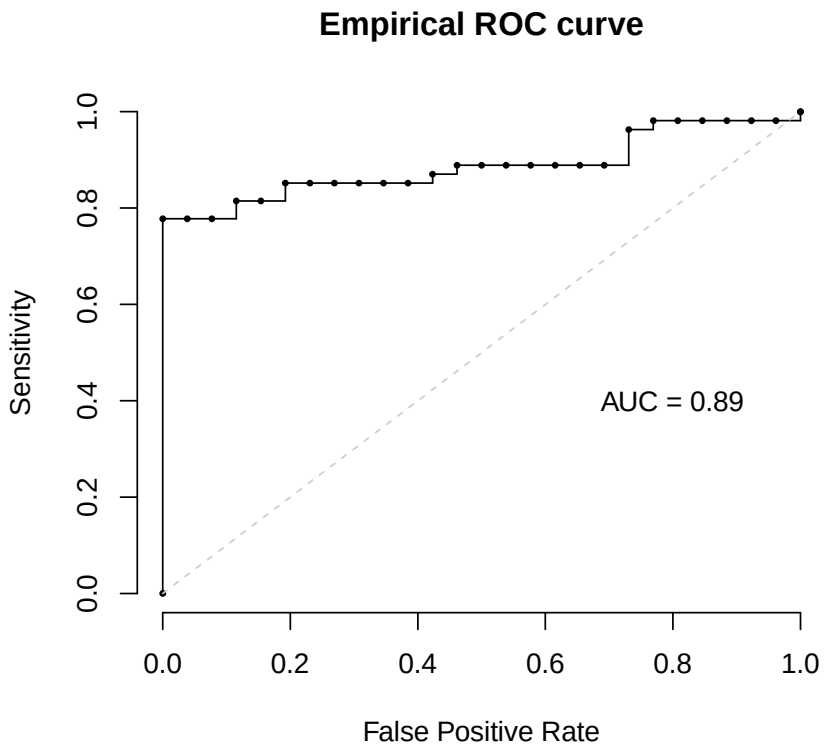

Figure 2.1: Empirical ROC curve comparing HCs to SOT recipients, the central reference line of $y=x$ represents the scenario of a so-called "useless" diagnostic test in which the sensitivity equals the false positive rate at all points (AUC=0.5).

In this empirical example, the AUC is 0.89, which is generally understood to mean that the measure is quite successful as a classifier for HCs and SOT recipients in this dataset. We note that classification isn't necessarily the goal of this work, although ROC curves are often used for classification purposes.

2.3 Parametric Methods: The Exponential ROC Curve

In this section, we will consider an exponential model, whereby we parametrically model the RBD_2 distribution in each group (i.e., the HCs and SOT recipients) as exponential. Note that this example is purely for the purposes of illustration rather than a reflection of an assumption we believe to hold. Recall that the exponential CDF is given as $F(t) = 1 - e^{-\lambda t}$ for some $\lambda > 0$. Under this model, the quantile function for the HCs is given by $F_0^{-1}(p) = {\lambda_0}^{-1}\log(1-p)$. The ROC function is therefore given by $\mathrm{ROC}(p) = F_1\left(F_0^{-1}(p)\right) = 1 - (1-p)^{\lambda_1/\lambda_0}$. When applying this function to the SOT COVID data, we can use maximum likelihood estimation to estimate both λ_0 and λ_1. In our example, the probability density function of the exponential distribution is parameterized as follows: $f(t) = \lambda e^{-\lambda t}$. Under this parameterization, it is straightforward to show that $\hat{\lambda}_0 = 1/\overline{Y}_0$ and $\hat{\lambda}_1 = 1/\overline{Y}_1$, where $\overline{Y}_0$ represents the sample mean test result from the HCs (of which there are n_0) and $\overline{Y}_1$ represents the sample mean for the SOT recipients (of which there are n_1). The derivation for the maximum likelihood estimator is shown below for a general λ estimated under a sample size of n:

Step 1: Express log-likelihood function:	$\ell(y\|\lambda) = n\log(\lambda) - \lambda\sum_{i=1}^{n} y_i$
Step 2: Determine score equation:	$\frac{d}{d\lambda}\ell(y\|\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i = 0$
Step 3: Solve score equation for λ:	$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} y_i} = \frac{1}{\overline{\overline{Y}}}$

Standard likelihood theory provides insights into the asymptotic behavior of $\hat{\lambda}_0$ and $\hat{\lambda}_1$; for $\hat{\lambda}_0$ in particular, $\sqrt{n_0}\left(\hat{\lambda}_0 - \lambda_0\right) \rightarrow_d \mathcal{N}(0, \lambda_0^2)$. Log-transformation serves as a variance-stabilizing function under this model [5]. In particular, $\sqrt{n_0}\left(\log\hat{\lambda}_0 - \log\lambda_0\right) \rightarrow_d \mathcal{N}(0,1)$, such that $\mathrm{Var}\left(\log\hat{\lambda}_1 - \log\hat{\lambda}_0\right) = n_0^{-1} + n_1^{-1}$. It is therefore readily shown that:

$$\log\left[\frac{\log\left(1 - \mathrm{ROC}(p)\right)}{\log(1-p)}\right] = \log\lambda_1 - \log\lambda_0,$$

which lends itself to the following (asymptotically justified) confidence band for $\mathrm{ROC}(p)$:

$$\mathcal{G} = \left\{\left(p, 1 - (1-p)^{\exp\left(\left(\log\hat{\lambda}_1 - \log\hat{\lambda}_0\right) \pm z_{1-\alpha/2}\sqrt{n_0^{-1}+n_1^{-1}}\right)}\right) : p \in [0,1]\right\}$$

The ROC curve with 95% confidence intervals included are shown in Figure 2.3.

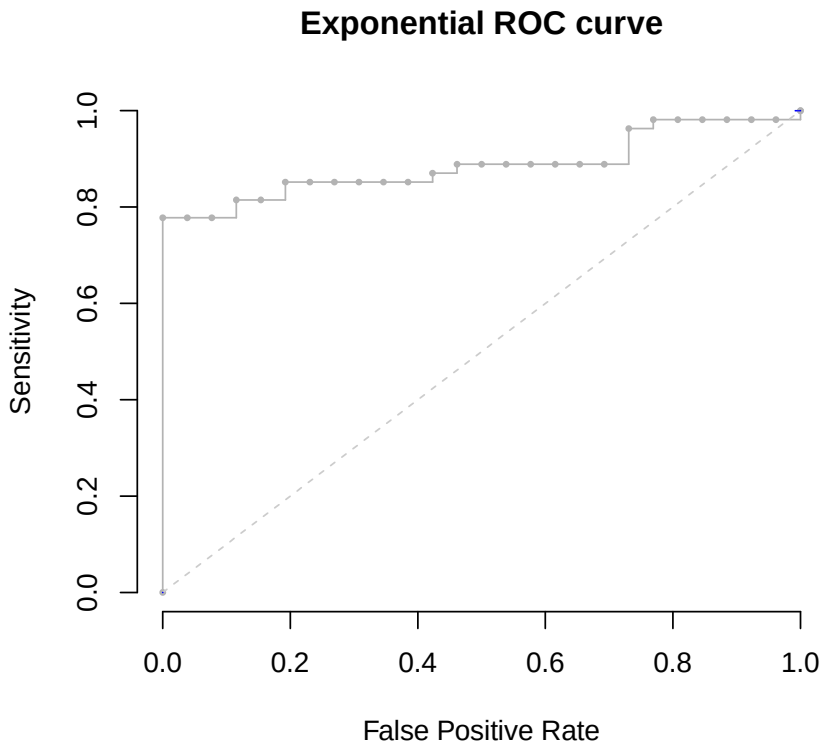

Figure 2.2: The exponential model of the ROC curve as estimated by maximum likelihood ($\hat{\lambda}_0$ = 0.453; $\hat{\lambda}_1$= 1.26), with the accompanying 95% confidence interval (using the variance stabilization method above). The dashed line serves as a reference for a "useless" test, and the empirical ROC curve is also shown for reference in gray.

The parametric fit to the ROC curve produces differs meaningfully from the empirical fit, which may be a reflection of the fact that the exponential distribution poorly describes the RBD_2 among the HCs (see Figure 1.2).

2.4 Parametric Methods: The Binormal ROC Curve

Another parametric method to estimate an ROC curve would be to assume normally distributed data in each group (i.e., $Y_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$). For ease of notation, we will denote the standard normal CDF as Φ, so that $F_1(y) = \Phi((y - \mu_1)/\sigma_1)$ represents the CDF for the SOT recipients and $F_0^{-1}(p) = \sigma_0\Phi^{-1}(p) + \mu_0$ represents the quantile function for the HCs. It follows readily that the ROC function takes the following form:

$$\mathrm{ROC}(p) = \Phi\left(\frac{\sigma_0\Phi^{-1}(p) + (\mu_0 - \mu_1)}{\sigma_1}\right)$$

Note that this parameterization of the ROC curve is referred to as the binormal form; it is so named because it arises from the setting in which both groups are presumed to adhere to a normal distribution. As we will later see, this form will be useful in semiparametric model. Parameters can be estimated using maximum likelihood, although it is typical to use the corresponding unbiased estimators for σ_0^2 and σ_1^2:

$$S_0^2 = \frac{1}{n_0 - 1}\sum_i^n \left(Y_{0i} - \overline{Y}_0\right)^2$$

The corresponding (asymptotically valid) pointwise confidence interval is given as:

$$\mathcal{G} = \left\{\left(p, \Phi\left(\frac{\left(\overline{Y}_0 - \overline{Y}_1\right) \pm t_{1-\alpha/2,\mathrm{df}}\sqrt{S_0^2 n_0^{-1} + S_1^2 n_1^{-1}} + S_0\Phi^{-1}(p)}{S_1}\right)\right) : p \in [0,1]\right\},$$

where "df" denotes the Welch-Satterthwaite approximation to degrees of freedom when the variances are not presumed equal. Figure 2.3 presents the estimated ROC curve along with corresponding 95% confidence intervals.

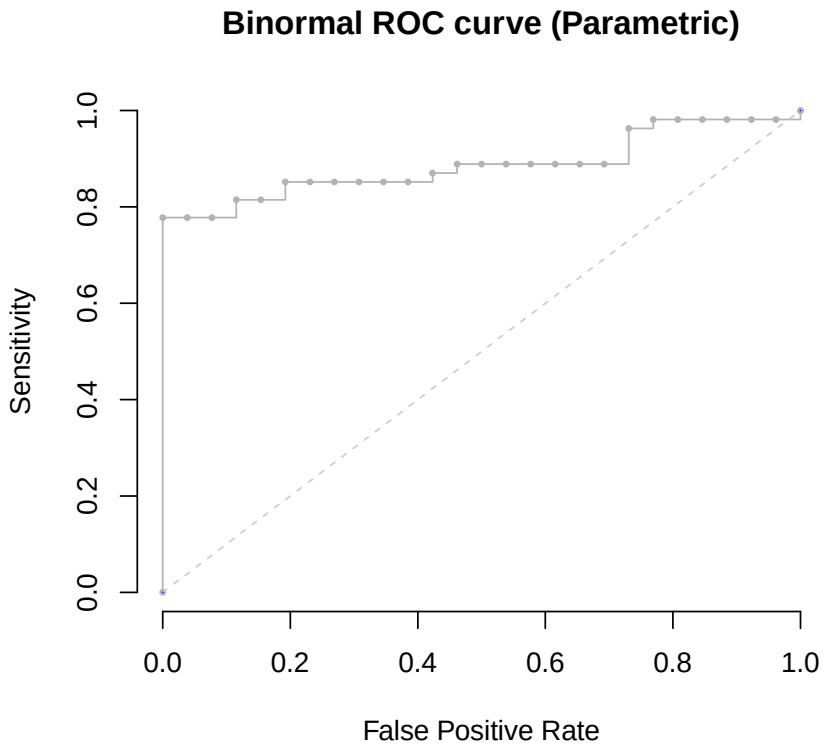

Figure 2.3: The parametric binormal model estimate of the ROC curve, along with confidence bands ($\bar{Y}_0 = 2.21$, $\bar{Y}_1 = 0.80$, $\hat{\sigma}_0 = 0.38$, $\hat{\sigma}_1 = 0.84$). The dashed line serves as a reference for a "useless" test, and the empirical ROC curve is also shown for reference in gray.

The parametric binormal model fit appears more closely (although certainly far from perfectly) aligned with the empirical data as compared to the exponential model. The improved performance of the binormal approach may be partly explained by the greater flexibility associated with more degrees of freedom (the normal model has two degrees of freedom where the exponential model has only one). Importantly, we must treat this heuristic interpretation with caution as the models are not nested.

2.5 The Binormal ROC Family

The family of binormal curves is quite flexible in that it captures a range of possible ROC curve shapes. To illustrate this, consider the following parametrization of the binormal ROC function, representing the mean difference, $\mu_0 - \mu_1$, by δ:

$$\mathrm{ROC}(p) = \Phi\left(\frac{\sigma_0 \Phi^{-1}(p) + \delta}{\sigma_1}\right).$$

By changing the form of the ROC function and reducing the number of parameters by one, we can more easily display the different shapes the binormal curve can take. Figure 2.4 shows how changes in the values of δ, σ_0, and σ_1 impact the shape of the binormal ROC model.

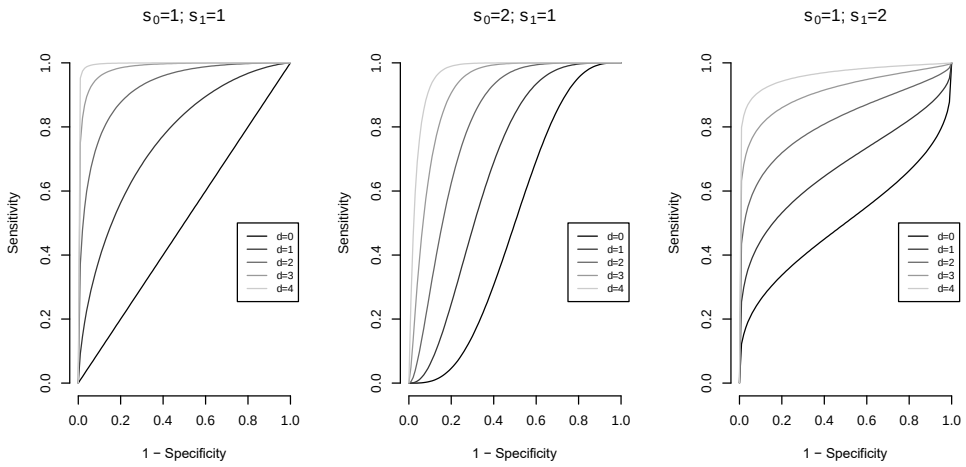

Figure 2.4: Shapes of binormal ROC curves. Left: standard deviations of healthy and SOT are equal. Center: the standard deviation of the healthy group is larger than the SOT group. Right: standard deviation in SOT group is higher than that of the healthy group. The major factor that separates the latter two examples is the direction of the concavity change across values of the FPR.

Given σ_0 and σ_1, fluctuations in the value of δ are associated with overall predictive ability. Given δ, the ratio $r = \sigma_0/\sigma_1$ controls the shape of the ROC curve, with the inverse of that ratio corresponding to a curve that is a reflection about the line $y = 1 - x$. Note that $\delta = 0$ corresponds to an AUC of 0.5, making this ROC curve as (in)effective as classifying the two groups purely randomly.

2.6 Induced Binormal ROC Curve Estimation

In this section, we illustrate how the binormal curve can be "induced" given a parametric form for either Y_0 or Y_1. Under the binormal model, $\text{ROC}(p) = \Phi\big(\alpha_0 + \alpha_1\Phi^{-1}(p)\big)$, so that, up to the link function $\Phi(\cdot)$ and the transformation $\Phi^{-1}(p)$, the ROC curve is linear in its target parameters. Suppose that RBD_2 follows an exponential distribution among SOT recipients (see Figure 1.2). We can then solve for the distribution function among HCs such that the ROC curve takes the binormal form. Under exponentially distributed RBD_2 for the SOT recipients (that is, under the assumption that $F_1(p) = 1 - \exp(-\lambda_1 t)$), it is straightforward to derive the quantile function for the HCs such that $F_1\big(F_0^{-1}(p)\big) = \text{ROC}(p) = \Phi\big(\alpha_0 + \alpha_1\Phi^{-1}(p)\big)$. Specifically, we have that $F_0^{-1}(p) = -\lambda_0 \log\Big(1 - \Phi\big(\alpha_0 + \alpha_1\Phi^{-1}(p)\big)\Big)$. To estimate $\boldsymbol{\theta} = (\lambda, \alpha_0, \alpha_1)$ by maximum likelihood requires an expression for the joint density of the data:

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{y}_0, \mathbf{y}_1) = \left(\prod_{i=1}^{n_0} f_0(y_{0i})\right)\left(\prod_{j=1}^{n_1} f_1(y_{1j})\right).$$

Since the quantile function for the HCs does not correspond to a known/named distribution, we determine the density function, $f_0(\cdot)$ as follows:

$$f_0(y; \boldsymbol{\theta}) = \phi\left(\frac{\Phi^{-1}(1 - \exp(-\lambda y) - \alpha_0)}{\alpha_0}\right)\frac{1}{\alpha_1}\frac{\lambda\exp(-\lambda y)}{\phi(\Phi^{-1}(1 - \exp(-\lambda y)))},$$

where $\phi(\cdot)$ denotes the standard normal density function. The log-likelihood is readily determined and can be maximized computationally in R using the *optim* function (see **Appendix I**). To obtain a 95% confidence interval, note that the *optim* function supplies the observed information for $\boldsymbol{\theta}$ (by way of the Hessian matrix). Since the ROC curve takes the form of a GLM, it is straightforward to determine a confidence interval on the scale of the linear predictor. Note that while λ is included as a parameter in the likelihood, the binormal form of the ROC

curve depends only upon α_0 and α_1. Let $\mathbf{a} = [0 \;\; 1 \;\; \Phi^{-1}(p)]^T$ so that $\mathbf{a}^T\boldsymbol{\theta} = \alpha_0 + \alpha_1\Phi^{-1}(p)$. We have that $\text{Var}(\mathbf{a}^T\boldsymbol{\theta}) = \mathbf{a}^T\text{Cov}(\boldsymbol{\theta})\mathbf{a}$. An asymptotically justified pointwise confidence interval on the scale of the ROC curve takes the following form:

$$\mathcal{G} = \left\{\left(p, \Phi\left(\hat{\alpha}_0 + \hat{\alpha}_1\Phi^{-1}(p) \pm z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}\left(\mathbf{a}^T\hat{\boldsymbol{\theta}}\right)}\right)\right) : p \in [0,1]\right\}$$

Figure 2.5 shows the 95% confidence bands plotted with the induced binormal ROC function:

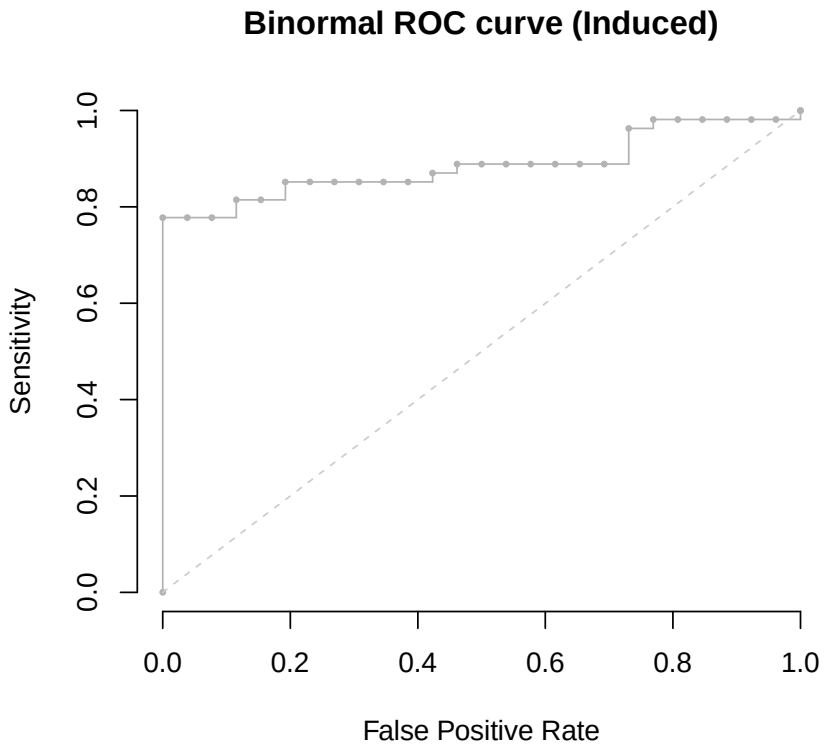

Figure 2.5: Estimated ROC curve under the induced exponential-induced binormal model ($\hat{\lambda}$ =1.256, $\hat{\alpha}_0$ =1.522, $\hat{\alpha}_1$ = 0.238). The dashed line serves as a reference for a "useless" test, and the empirical ROC curve is also shown for reference in gray.

This parametric choice provides a fit that matches the empirical fit reasonably well, particularly as compared to the model assuming the data emerged from an exponential distribution in each group.

CHAPTER 3

Semiparametric ROC Estimation and Regression

3.1 Semiparametric ROC Estimation

Our prior examples relied on distributional assumptions. We've previously noted the flexibility associated with the binormal family, and we've illustrated how it can emerge from non-normal data. We now present and implement a method for estimating the parameters of a binormal ROC semiparametrically (that is, without imposing assumptions directly on the individual groups beyond the form of the overall ROC curve). To that end, let $U_{ij} = 1(Y_{1j} \leq Y_{0i})$ denote the indicator of the j^{th} SOT recipient having an RBD_2 value no higher than the RBD_2 value of the i^{th} HC [6]. Note that our motivating dataset comprises $n_0 = 26$ HCs and $n_1 = 54$ SOT recipients for a total sample size of $n = n_0 + n_1 = 80$. On the other hand, the variable U_{ij} includes each SOT-HC pair, of which there are $N = n_0 \times n_1 = 1404$ values that cannot be presumed independent. Conveniently, we have the following result: $\mathrm{E}\left[U_{ij} | F_0(Y_{0i}) = p\right] = \mathrm{ROC}(p)$, which can be shown as follows:

$$\begin{aligned}
\mathrm{E}\left[U_{ij} | F_0(Y_{0i}) = p\right] &= \mathrm{P}\left(Y_{1j} \leq Y_{0i} | F_0(Y_{0i}) = p\right) \\
&= \mathrm{P}\left(Y_{1j} \leq Y_{0i} | Y_{0i} = F_0^{-1}(p)\right) \\
&= \mathrm{P}\left(Y_{1j} \leq F_0^{-1}(p)\right) \\
&= F_1\left(F_0^{-1}(p)\right) \\
&= \mathrm{ROC}(p).
\end{aligned}$$

Typically, the structure imposed on the ROC curve is that of the binormal form [7]:

$$\mathrm{E}\left[U_{ij} | F_0(Y_{0i}) = p\right] = \mathrm{ROC}(p) = \Phi\left(\alpha_0 + \alpha_1 \Phi^{-1}(p)\right).$$

The parameters can be estimated by via GLM for binary outcomes under a probit link. The standard errors associated with the *glm* function in R arise from an assumption of independence and are therefore inappropriate. A nonparametric bootstrap procedure can be implemented in order to estimate a variance-covariance matrix for $(\hat{\alpha}_0, \hat{\alpha}_1)$, from which a confidence interval can be formed (see Section 2.6; further see **Appendix II** for sample code). Figure 3.1 presents the estimated curve and 95% confidence interval under this semiparametric estimation procedure.

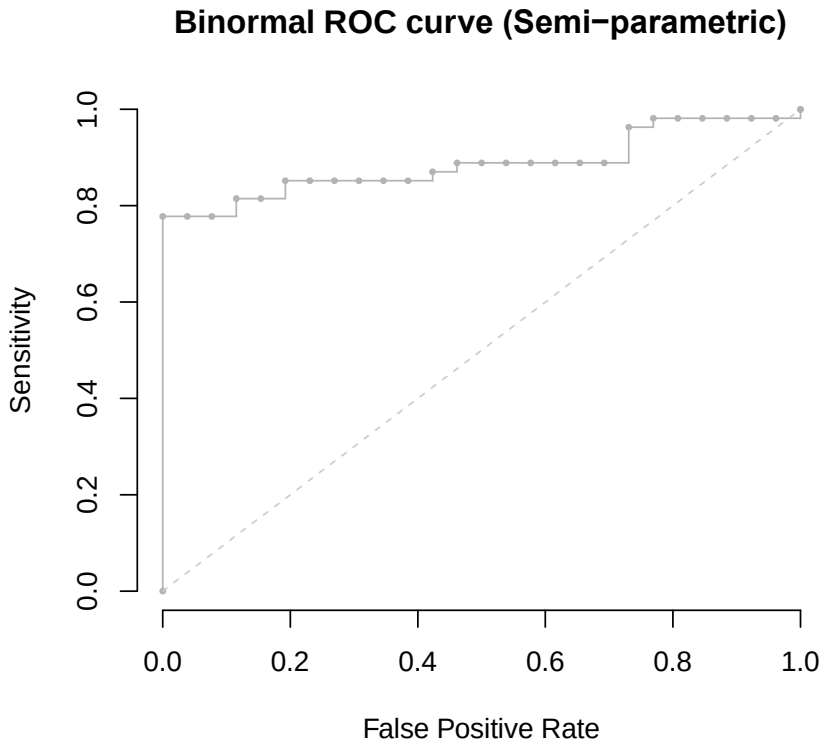

Figure 3.1: Estimated ROC curve under the semiparametric method ($\hat{\alpha}_0 = 1.30$, $\hat{\alpha}_1 = 0.433$). The dashed line serves as a reference for a "useless" test, and the empirical ROC curve is also shown for reference in gray.

Notably, the semiparametric fit to the data is closely calibrated to the empirical ROC curve. The confidence band is wider as compared to those shown in previous approaches (note that semiparametric models are sometimes less efficient as compared to parametric approaches).

3.2 Subgroup Comparison Using ROC Regression

We now aim to illustrate how the ROC curve varies across groups differing in some baseline covariate. There are two classes of covariates: those that apply to both the HCs and SOT recipients (e.g., age), and those that apply to SOT recipients only (e.g., time post-transplant). Although ROC regression methodology accommodates both forms of covariates, we restrict our attention to the latter, as the former requires a more in-depth discussion of quantile regression techniques that are outside the scope of this work. [4] A semiparametric regression model that includes some variable, X, in the model (i.e., one that only applies to SOT recipients) could be parameterized as follows:

$$\mathrm{E}\left[U_{ij} | F_0(Y_{0i}) = p, X_j = x_j\right] = \Phi\left(\alpha_0 + \alpha_1 \Phi^{-1}(p) + \alpha_2 x_j\right),$$

Including an additional coefficient into the semiparametric model allows us to compare subgroups and test whether the distribution of RBD_2 effectively classifies HCs and SOT recipients differentially by X. The estimation procedure is analogous to that described in Section 3.1 (i.e., using GLM machinery).

3.3 ROC Regression of RBD_2 on Time Post-transplant

As an example, we consider ROC regression of RBD_2 on time-post transplant, T.

$$\mathrm{E}\left[U_{ij} | F_0(Y_{0i}) = p, T_j = t_j\right] = \Phi\left(\alpha_0 + \alpha_1 \Phi^{-1}(p) + \alpha_2 t_j\right),$$

The median time post-transplant among SOT recipients was 7 months (IQR: [3, 13]); a histogram of time post-transplant is presented in Figure 3.2. We hypothesize that the ROC curve will better discriminate between HCs and SOT recipients having shorter time post-transplant. Namely, we hypothesize that $\alpha_2 < 0$ (recall the interpretation of parameters in the binormal model discussed in Section 2.5). Largely, the rationale for this hypothesis is that, although SOT

recipients are typically on immunosuppressants for the remainder of their life following transplantation to mitigate risks of antibody and cellular rejection, the early post-transplant period is typically associated with a higher degree of suppression. In turn, a patient in the earlier post-transplant period is likely less able to mount an immune response to antigen suppled in vaccines. This would make it easier, we suspect, to differentiate between the immunogenic vaccine response of HCs and recent SOT recipients than it would be to differentiate between HCs and SOT recipients having undergone transplantation long ago.

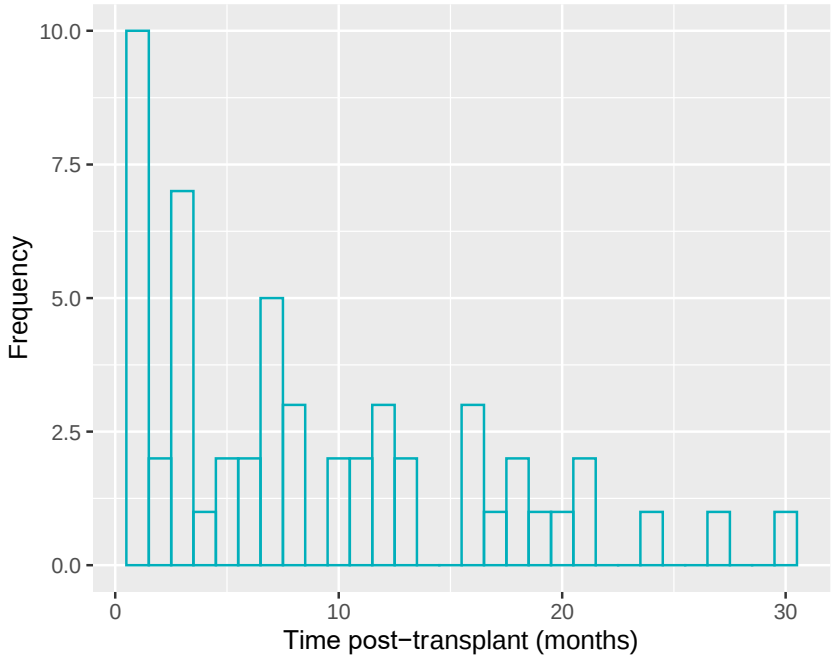

Figure 3.2: Histogram of time post-transplant among SOT recipients.

The results of the GLM fit are shown in Table 3.1. Our model fit confirms our pre-specified hypothesis that time post-transplant determines the predictive capacity of RBD_2 in comparing HCs and SOT recipients (with the distribution of responses among SOT recipients being more comparable to those of HCs when later post-transplant.

Table 3.1: Results of ROC regression with covariate months post-transplant

Coefficient	Estimate	95% CI	p-value
α_0	1.97	[1.13, 2.82]	<0.001
α_1	0.490	[0.21, 0.77]	<0.001
α_2	-0.058	[-0.11, -0.006]	0.037

Figure 3.3 further illustrates the estimated ROC curves in specific groups defined by their continuous time post-transplant. Confidence bands could also be obtained for these specific subgroups according to similar procedures as those previously described/implemented.

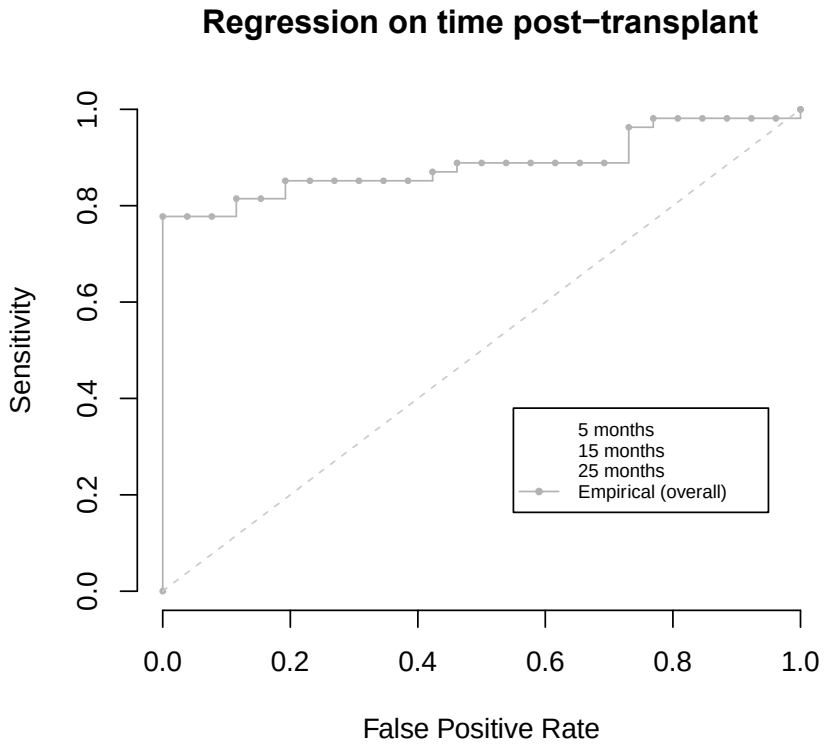

Figure 3.3: Estimated ROC curves displaying the results of the regression analysis on time post-transplant. Briefly, the longer the duration since transplantation, the more similar the distribution of RBD_2 is between HCs and SOT recipients.

CHAPTER 4

DISCUSSION

In this thesis, we presented and discussed a variety of ways to estimate the ROC curve beyond the traditionally implemented empirical/non-parametric approach. Our presentation of the empirical method serves as a reasonable benchmark to which to compare the other competing approaches, in the sense that the degree of alignment between the estimated model-based ROC curve and the empirical curve speaks in part to the reasonableness of assumptions put forth by that model.

The first parametric estimation methods we explored were the exponential model, in which both healthy and SOT groups are both assumed to follow an exponential distribution. This modeling approach provided the poorest fit of the empirical ROC curve of any of the methods we explored, at least visually. The reason for this poor fit is elucidated upon examining the empirical distribution of RBD_2 among HCs in particular. While the SOT recipients have an RBD_2 distribution that could be reasonably approximated by an exponential distribution, the HCs clearly do not. The parametric binormal method suffers, in some sense, from the reverse problem, in which the normal distribution is clearly a very poor representation of the distribution of RBD_2 among SOT recipients (though not as poor a representation of HCs). The fact that this model provides a fit more closely resembling the empirical ROC curve may be at least in part attributable to the higher degree of model complexity (there are two parameters per group rather than one—namely, the mean and the variance). However, this heuristic explanation comes with the important caveat that the exponential and binormal models are not nested, so a direct

comparison of model complexity as an explanation for this phenomenon may not be fully appropriate.

Our exploration of the induced-binormal method is, in some sense, a logical pathway toward the semiparametric approach. Just as a form for the ROC curve emerges from suppositions surrounding the distribution of data in each group, one can impose a structure on the ROC curve and exactly *one* of the two comparator groups; from this, the parametric form of the other group is implied. Based upon visual inspection, we elected to model the SOT group using an exponential distribution (the RBD_2 distribution among SOT recipients is clearly similarly right-skewed) and determine the induced form for the HCs that would produce an ROC curve of the binormal family. Had we done the reverse, the resulting binormal curve would have been a poor fit to the data (results not depicted).

The semiparametric approach to ROC analysis offers the most flexibility of the approaches we considered. The binormal family—which, as shown in one of our key examples, can emerge from non-normally distributed data—offers a diversity of shapes for the ROC curve. Heuristically, it is "easier" to index a class of monotone-increasing functions $f : [0,1] \to [0,1]$ than it is to parametrically model the underlying distributions of two groups. We found that the estimated ROC curve based on the semiparametric approach was most closely aligned with the empirical ROC curve. Of note, this may have come at the cost of variability, partly evidenced by the fact that the confidence intervals associated with the semiparametric approach were fairly wide.

The semiparametric estimation procedure served as the basis for our exploration of ROC regression. Notably, there are parametric ROC regression approaches, though we did not explore them in this work. Our primary example made use of time post-transplant as a key variable, although we could also examine the effect of variables that are directly measurable in both the SOT and HC groups (e.g., age). Modeling variables common to both SOT recipients and HCs would require quantile regression techniques that were not within the scope of this work.

Notably, Yanis et al. performed regression-based analyses based on contrasts in the mean. Methods to compare mean differences are already semiparametric in nature; part of the novelty of this work lies in the fact that we're exploring measures of stochastic ordering rather than measures of central tendency. Ultimately, we identified time post-transplant as a significant determinant of ROC curve performance, which supported our pre-specified hypothesis that vaccine-induced immune response in an SOT recipient is better among individuals who are further out from their time of transplantation. As evidenced by Figure 3.2, the model suggests that the vaccine-induced immunogenicity among SOT recipients more closely resembles the immunogenicity of HCs among those whose transplantation occurred a longer time ago from the time of study; this is intuitive from a clinical standpoint, as the intensity of immunosuppressants is higher in the early post-transplant period.

It is important to also acknowledge some of the limitations of the study upon which our example was based. There was a notable age gap between SOT recipients and healthy controls, with the SOT recipients being significantly older. This discrepancy was not accounted for in our regression analysis, though older age is highly unlikely to explain the substantial degree of

between SOT recipients and HCs. Another study limitation includes its inability to fully explore factors among the SOT group contributing to their collective lower immune response, though we used time post-transplantation as a proxy for such factors. A more rigorous analysis involving immunosuppressive regimen would require a larger sample than what was available.

Our work has broader applicability than to our motivating example of vaccine-induced immune responses. ROC-based methodology is most frequently used when seeking to determine the value of a diagnostic test. ROC estimation methods and regression can offer medical professionals a better idea of which factors may influence the effectiveness of a diagnostic test. Knowing what factors a patient may exhibit can allow for a more personalized form of care, knowing the most effective way to implement a diagnostic test, based on the patient's personal characteristics.

BIBLIOGRAPHY

[1] Yanis A., Haddadin Z., Spieker A.J., Waqfi D., Rankin D.A., Talj R., Thomas L., Birdwell K.A., Ezzell L., Blair M., Eason J., Varjabedian R., Warren C.M., … Halasa N.B. (2021). Humoral and cellular immune responses to the SARS-CoV-2 BNT162b2 vaccine among a cohort of solid organ transplant recipients and healthy controls. *Transplant Infectious Disease* 2022; 24(1), e13772.

[2] Pepe M.S. A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika* 1997; 84(3), 595-608.

[3] Pepe M.S. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* 1998; 54(1), 124-135.

[4] Koenker R.W., D'Orey D. Computing regression quantiles. *Applied Statistics* 1987; 36(3): 383-393.

[5] Janes H., Pepe M.S. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika* 2009; 96(2): 371-382.

[6] Alonzo T.A., Pepe M.S. Distribution-free ROC analysis using binary regression techniques. *Biostatistics* 2002; 3(3): 421-432.

[7] Pepe M.S. A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika* 1997; 84(3): 595-608.

[8] Greaney A.J., Loes A.N., Gentles L.E., Crawford K.H., Starr T.N., Malone K.D., Chu H.Y., Bloom, J.D. Antibodies elicited by mRNA-1273 vaccination bind more broadly to

the receptor binding domain than do those from SARS-CoV-2 infection. *Science Translational Medicine* 2021; 13(600).

APPENDIX I

R Code: Estimation of Induced Binormal ROC Curve

```
# Extract salient data
Y0 <- sot[which(sot$sot == 0),9]
Y1 <- sot[which(sot$sot == 1),9]

# Negative log-likelihood function
negative_ll <- function(theta, y0, y1){
  lambda <- theta[1]
  alpha0 <- theta[2]
  alpha1 <- theta[3]

  tmp1 <- dexp(y1,lambda)
  loglik1 <- log(tmp1)

  exp1 <- dnorm((qnorm(1-exp(-(lambda*y0)))-alpha0)/alpha1)
  exp2 <- 1/alpha1*lambda*exp(-lambda*y0)
  exp3 <- dnorm(qnorm(1-exp(-lambda*y0)))
  expr <- exp1*(exp2/exp3)
  negloglik <- -1*(sum(log(expr)) + sum(loglik1))
  return(negloglik)
}

# Optimization procedure
ROC_induced <- optim(c(0.8,0.5,0.8),
                     negative_ll,
                     y0=Y0, y1 = Y1,
                     hessian=TRUE)

# Extract estimates
a0 <- ROC_induced$par[2]
a1 <- ROC_induced$par[3]


# Variance estimation (point-wise)
p <- seq(0,1, length.out = 500)
CovHat <- solve(ROC_induced$hessian[1:3,1:3])
C <- cbind(0, 1, qnorm(p))
Vhat <- diag(C %*% CovHat %*% t(C))

# Plot ROC curve with confidence band
plot(p, pnorm(a0 + a1*qnorm(x)),
        col = "blue", type = "l", lty = 1,
        ylab = "Sensitivity", xlab = "1 - Specificity")

for (j in 1:length(p))
{
  segments(p[j], pnorm(a0 + a1*qnorm(p[j]) - qnorm(0.975) * sqrt(vhat)[j]),
           p[j], pnorm(a0 + a1*qnorm(p[j]) + qnorm(0.975) * sqrt(vhat)[j]),
           col = rgb(0,0,1,0.1))
}
```

APPENDIX II

R Code: ROC Regression on Time Post-transplant

```
# Extract salient data
Y0 <- sot[which(sot$sot == 0),9]
Y1 <- sot[which(sot$sot == 1),9]

# Sample sizes
n0 <- length(Y0)
n1 <- length(Y1)

nboot <- 10000
Bres0 <- data.frame(matrix(nrow = nboot, ncol = 2))

## Create U outcome and estimate S0(c|X)
 for (b in 1:nboot)
  {
    ## Re-sample with replacement
    b0 <- sample(1:n0, replace = TRUE)
    b1 <- sample(1:n1, replace = TRUE)
    y0 <- Y0[b0]
    y1 <- Y1[b1]

    ## Create U outcome and estimate F0(c|X)
    U <- as.numeric(t(outer(y1, y0, "<=")))
    #Use leq bc CDF
    Pij <- matrix(colMeans(outer(y0, y0, "<=")), nrow = n0, ncol = n1)
    P <- matrix(Pij, ncol = 1)
    Q <- qnorm(P)
    Q[P == 0] <- min(Q[P != 0])
    Q[P == 1] <- max(Q[P != 1])

    ## Run ROC-GLM
    zz <- glm(U ~ Q, family = binomial(link = "probit"))
    Bres0[b,] <- as.numeric(coef(zz))
 }
```