Non-likelihood based model evaluation and comparison with application to genetic and clinical HIV-1 outcomes

Ashley Elise Giambrone

University at Albany, State University of New York, agiambrone@albany.edu

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Biostatistics Commons

Recommended Citation


This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive. Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
NON-LIKELIHOOD BASED MODEL EVALUATION
AND COMPARISON WITH APPLICATION TO
GENETIC AND CLINICAL HIV-1 OUTCOMES

by

Ashley Elise Giambrone

A Dissertation
Submitted to the University at Albany, State University of New York
in Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy

School of Public Health
Department of Epidemiology and Biostatistics
2013


Abstract

Although treatment for human immunodeficiency virus type-1 (HIV-1) has undergone drastic change and morbidity and mortality has decreased over time, the development of drug-resistant HIV-1 is of concern for the long-term antiretroviral treatment of infected individuals. Drug-resistant virus is known to manifest with potentially complex mutational patterns in the HIV-1 genotype sequence and is associated with decreased response to therapy. Resistance occurs either as a result of development of mutations in the viral genome under selective drug pressure or as a result of naturally occurring polymorphisms. The most effective treatment methods are still debated at this time; however, current treatment methods are focused on drug cocktails and the importance of first-line regimens. This emphasizes the need for rapid discovery of virologic failure through clinical tests and statistical methods for interpreting genotypic resistance patterns.

The overall goal of the proposed research is to statistically quantify changes in HIV-1 genotype sequences over time that are associated with drug resistance utilizing AIDS Clinical Trial Group (ACTG) 398 data. One main aspect will be to investigate genotype changes from baseline to time of virologic failure through high-dimensional reduction techniques. The second and third goals are to predict both the outcome of RNA levels after 24 weeks of therapy and the right-censored time to virologic failure using newly developed, non-likelihood based methods to evaluate and compare statistical models.
Acknowledgments

Though only my name appears on the cover of this dissertation, many people have contributed to its production. I owe my gratitude to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

I would like to express the deepest appreciation to my committee chair, Dr. Greg DiRienzo, who has gone above and beyond the expectations of any doctoral advisor from the outset to the very end. I am grateful for his interest and assistance in the pursuit of these studies, and whose expertise, understanding, and dedication made it possible for me to work on a topic that was of great interest to me. Without his guidance and persistent help this dissertation would not have been possible.

I would like to thank my committee members, Dr. Recai Yucel and Dr. Perry Smith for their time, patience, interest and helpful comments, and whose vast expertise of statistical and clinical methods provided me with a real world context for the findings of this dissertation.

I am profoundly thankful for the love, support and patience from my parents, my grandmother, and most importantly my husband, Greg. His support, encouragement, quite patience and unconditional love were the foundation upon which the past six years of my life have been built. His tolerance of my occasional vulgar moods is a testament in itself of his unyielding affection and love. I know that one day soon I will be repaying the favor.

Last but not least I thank my friends and classmates for making sure that I have fun. Many friends have helped me stay sane through these difficult years. Their support and care helped me overcome setbacks and stay focused on my graduate study. I greatly value their friendship and I deeply appreciate their belief in me.
The work performed here would not have been possible without the use of data from ACTG 398 from the National Institute for Allergy and Infectious Diseases. This dissertation is part of the completion of DACS 288 "Statistical Methods for Identifying and Modeling Patterns of HIV-1 Genotype Evolution".
# Contents

Abstract ii

Acknowledgments iii

List of Tables viii

List of Figures ix

Preface xi

1 Statistical Methods for Identifying and Modeling Patterns of HIV-1 Genotype Evolution 1

1.1 Introduction ................................................. 1

1.2 Comparing Joint Distributions .......................... 5

1.3 Null Hypotheses ........................................... 9

1.4 Marginal Homogeneity ................................. 9

1.4.1 Conditional independence (CI) .................... 13

1.5 Multiple Testing Procedures .......................... 16

1.6 Analysis of ACTG 398 ............................... 19

1.7 Discussion .............................................. 23

2 New Statistical Methods to Evaluate and Compare Generalized Logit Models 25

2.1 Introduction ................................................. 25

2.2 Data ....................................................... 26

2.3 Logistic Regression: Polychotomous Response .......... 27

2.3.1 Proportional Odds Model ......................... 27

2.3.2 Polychotomous Logistic Regression .......... 32

2.4 Apparent Error ......................................... 36

2.4.1 Ordinary Logistic Regression ..................... 36
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4</td>
<td>2.4.2</td>
<td>Comparing Average Squared Prediction Error Between Polychotomous</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Binary Regression Models</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td></td>
<td>Model Checking</td>
<td>47</td>
</tr>
<tr>
<td>2.6</td>
<td></td>
<td>Model Comparisons</td>
<td>50</td>
</tr>
<tr>
<td>2.7</td>
<td></td>
<td>Application</td>
<td>53</td>
</tr>
<tr>
<td>2.8</td>
<td></td>
<td>Simulation</td>
<td>58</td>
</tr>
<tr>
<td>2.9</td>
<td></td>
<td>Discussion</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>New Statistical Methods to Evaluate and Compare Survival Models</td>
<td>63</td>
</tr>
<tr>
<td>3.1</td>
<td></td>
<td>Introduction</td>
<td>63</td>
</tr>
<tr>
<td>3.2</td>
<td></td>
<td>Survival Data</td>
<td>64</td>
</tr>
<tr>
<td>3.3</td>
<td></td>
<td>Notation</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>3.3.1</td>
<td>Common Families of Survival Distributions</td>
<td>68</td>
</tr>
<tr>
<td>3.4</td>
<td></td>
<td>Kaplan-Meier (KM) Estimator</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>3.4.1</td>
<td>Log-rank Test</td>
<td>71</td>
</tr>
<tr>
<td>3.5</td>
<td></td>
<td>Cox’s Proportional Hazards Model</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>3.5.1</td>
<td>Maximum Likelihood Estimation of ( \beta )</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>3.5.2</td>
<td>Inference</td>
<td>78</td>
</tr>
<tr>
<td>3.6</td>
<td></td>
<td>Prediction</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>3.6.1</td>
<td>Previous Estimates</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>3.6.2</td>
<td>Proposed Method</td>
<td>81</td>
</tr>
<tr>
<td>3.7</td>
<td></td>
<td>Model Checking</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>3.7.1</td>
<td>Martingale Residuals</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>3.7.2</td>
<td>Cumulative Martingale Residuals</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>3.7.3</td>
<td>Checking for Proportional Hazards</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>3.7.4</td>
<td>Calculating p-values</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>3.7.5</td>
<td>Schoenfeld Residuals</td>
<td>89</td>
</tr>
<tr>
<td>3.8</td>
<td></td>
<td>Model Comparisons</td>
<td>91</td>
</tr>
</tbody>
</table>
3.9 Application ......................................................... 94
3.10 Simulation ......................................................... 98
3.11 Discussion ......................................................... 100

4 Future Work ......................................................... 104

5 Tables ................................................................. 121

6 Figures ............................................................... 129
# List of Tables

1. 1-way Changes ................................................. 121
2. 2-way Joint Changes for RT108 and RT190 .................. 122
3. 2-way Joint Changes for RT181 and RT190 .................. 122
4. 3-way Joint Changes ............................................. 122
5. Predictors Included in Polychotomous Models M0-M4 ........ 123
6. P-values for Checking Functional Form of Continuous Covariates and the Link Function for Polychotomous Models M0-M3 .......... 123
7. P-values for Checking Functional Form of Continuous Covariates and the Link Function for Polychotomous Model M4 .......... 124
8. Prediction Error (PE) Results for Polychotomous Models ........ 124
9. Selected Model Results for $M_3$ ................................ 125
10. Model Comparisons for Polychotomous Models ................ 125
11. Polychotomous Model Simulation Results .................... 126
12. Predictors Included in Survival Models M0-M4 ............... 126
13. Supremum Test for Functional Form .......................... 126
14. Supremum Test for Proportional Hazards Assumption .......... 127
16. Model Comparisons for Survival Models ...................... 127
17. Survival Model Results for $M_3$ ............................... 128
18. Survival Model Simulation Results ............................ 128
List of Figures

1  Model M0, Level 1 .................................................. 129
2  Model M0, Level 2 .................................................. 129
3  Model M0, Level 3 .................................................. 129
4  Model M1, Level 1 .................................................. 129
5  Model M1, Level 2 .................................................. 130
6  Model M1, Level 3 .................................................. 130
7  Model M2, Level 1 .................................................. 130
8  Model M2, Level 2 .................................................. 130
9  Model M2, Level 3 .................................................. 130
10 Model M3, Level 1 .................................................. 130
11 Model M3, Level 2 .................................................. 131
12 Model M3, Level 3 .................................................. 131
13 Predicted Probabilities Against Baseline RNA ...................... 131
14 Predicted Probabilities Against Week 4 RNA ....................... 131
15 Predicted Probabilities Against Baseline RT103 ...................... 131
16 Predicted Probabilities Against Week 4 RT103 ...................... 131
17 Generalized Extreme Plot of Errors ................................ 132
18 Functional Form: Baseline CD4 .................................. 132
19 Functional Form: Baseline RNA .................................. 132
20 Proportional Hazards: Baseline CD4 .............................. 133
21 Proportional Hazards: Baseline RNA .............................. 133
22 Proportional Hazards: Week 4 RNA .............................. 133
23 Proportional Hazards: Baseline RT100 ........................... 133
24 Proportional Hazards: Baseline RT103 ........................... 133
25 Proportional Hazards: Week 4 RT103 ............................ 133
26 Predicted Survival Function - Week 4 RNA ....................... 134
Preface

Treatment of human immunodeficiency virus type 1 (HIV-1) infection has undergone considerable changes due to widespread usage of highly active antiretroviral therapy (HAART) (Palella, 1998; Miiro, 2009). National surveillance data have shown a drastic decrease in morbidity and mortality associated with acquired immunodeficiency syndrome (AIDS) and these reductions have been attributed to the use of more intensive antiretroviral therapies (Palella, 1998). A study conducted in Uganda from 1999-2006 found overall morbidity during the first year was reduced by 80% among those receiving HAART (Miiro, 2009). However, there are many factors that have been found to limit the success of HAART including poor adherence, regimen complexity, pharmacodynamic interactions, drug tolerability and toxicity, therapy costs, and the presence of comorbid conditions such as substance abuse and addiction (Palella, 2002). Also, of major concern is virologic failure (Palella, 2002; Foulkes, 2003; Harrigan, 2005). Patients receiving HAART to suppress HIV-1 may stop responding to treatment due to virologic failure (Foulkes, 2003). This resistance occurs either as a result of development of mutations in the viral genome under selective drug pressure or as a result of naturally occurring polymorphisms (Foulkes, 2003). It is estimated that virologic failure occurs in as many as 25% - 50% of HIV-1 infected individuals within two years of initiation of HAART (Harrigan, 2005). Among viremic patients, a recent study found 76% had resistance to one or more antiretroviral drugs (Richman, 2004). The development of resistance is influenced by adherence to therapy, the pharmacokinetics of antiretrovirals, and high baseline plasma virus loads (Harrigan, 2005). HIV-1 resistance is of major concern not only for the patient’s overall health and well being, but also for public health due to the transmission of resistant virus (Richman, 2004; Booth, 2007; Conlon, 1994; Grant, 2002; Yerly, 2001; Weinstock, 2004; Wensing, 2005). It has been estimated that the prevalence from 1995 to 1998 of transmitted resistant virus was 3.4% and increased to 12.4% from 1999 to 2000 (Little, 2002; Barbour, 2004).
Viral resistance is known to manifest with potentially complex mutational patterns in the HIV-1 genotype sequence. Due to increased prevalence of resistant HIV-1, it has become common practice to conduct HIV resistance testing on serum samples of infected individuals using genotypic assays that detect mutations known to confer drug resistance (Shafer, 2000; Tural, 2002). Genotypic information has important implications, not only for patient treatment plans, but also for relating HIV viral genotype to viral susceptibility as new drugs are developed and identifying new important mutations leading to virologic failure (Sevin, 2000). A major clinical trial, Aids Clinical Trial Group Study 398 (ACTG 398), has utilized genotype sequencing and gene expression arrays to identify patients’ time to virologic failure. ACTG 398 randomized treatment experienced subjects with loss of virological suppression to one of four salvage regimens (Hammer, 2002). The primary endpoint was time from randomization to virologic failure. An HIV-1 genotype sequence from the protease and reverse transcriptase (RT) regions was obtained for each subject at baseline and for most subjects at one or more post-baseline times, including the time of virologic failure for those subjects experiencing this event. In addition to various baseline variables recorded, CD4 count and viral load were also recorded longitudinally.

The overall goal of the proposed research is to statistically quantify changes over time in HIV-1 genotype sequences that are associated with drug resistance utilizing ACTG 398 data. One main aspect will be to investigate genotype changes from baseline to the time of virologic failure. The secondary goal is to predict both the outcome of RNA levels after 24 weeks of therapy and the right-censored time to virologic failure using new, non-likelihood based methods to evaluate and compare models.

Presented in paper 1 is the development of a computationally efficient non-parametric statistical method that can be used to identify simultaneous changes in the joint distribution of HIV-1 genotype between baseline and time of virologic failure. Virologic failure will
be defined as either confirmed rebound of viral load after confirmed suppression, or failure to suppress by 24 weeks. Those subjects who do not experience virologic failure will be assumed to not have drug-resistant virus. Thus, examining the changes in the joint distribution of HIV-1 genotype sequence between baseline and time of virologic failure among subjects who experienced virologic failure, we can identify those simultaneous genotype changes that are apparently important contributors to drug-resistance.

Presented in paper 2 is an investigation of the association between HIV-1 RNA genotype at week 24 and measured characteristics such as baseline and post-baseline markers, as well as clinical measurements. Due to the high cost of obtaining RNA genotype measures, the objective of this paper is to examine if an additional RNA genotype measure at week 4 assists in the prediction of treatment failure at week 24, thus allowing for timely change to a more effective treatment regimen leading to a better prognosis. To achieve this goal, prediction error estimates are derived for generalized logit models.

It is of further scientific and clinical interest to assess the prediction performance of survival models that analyze censored survival times. Presented in paper 3 is the development of prediction error estimation for survival models to accommodate the missing data issue due to censoring. To adjust for censoring, a new prediction error estimate is derived to ensure censored individuals’ information is maintained in the prediction error.

The findings of these analyses apply to a particular cohort of subjects with a particular drug exposure, not to the entire population of HIV-1 infected subjects. However, the statistical methods used to obtain such results can in general be used on any cohort of subjects.
1 Statistical Methods for Identifying and Modeling Patterns of HIV-1 Genotype Evolution

1.1 Introduction

Previous statistical methods for genotype sequencing and gene expression arrays were inefficient and complex; however, in recent years with the development of new technologies, scientific and statistical advances have become more attainable (Shafer, 2000; Sevin, 2000). On the other hand, such technologies produce datasets with many more variables than data points resulting in datasets with high-dimensionality and complicated dependency structures producing several statistical complications (Sevin, 2000). The high dimensionality of viral genotypes generate an infinite number of variable/level combinations, and even if the analysis were limited to mutations in positions (codons) of the genotype sequence previously identified or those that cause resistance, the number of possible genotypes remains insurmountable. For example, suppose there are R variables of interest. A full comparison between groups gives consideration to all marginal (first-order) variable/level pairs and all two-way thru R-way combinations of variable/level pairs, which in general will not be computationally feasible. Also of concern, comparing two high-dimensional multivariate distributions results in the possibility of redundant differences between groups. For example, suppose event A has a different prevalence between groups and event B is independent of A within each group, but has equal prevalence between groups. Then, event \( (A \cap B) \) has different prevalence between groups, but is redundant because it is solely driven by event A. The issues related to the use of high dimensional data are explored further during this research study analyzing AIDS Clinical Trials Group study 398 (ACTG 398), an example of a high dimensional dataset.

ACTG 398 randomized treatment experienced subjects with loss of virological suppression
to one of four salvage regimens (Hammer, 2002). The primary endpoint was time from randomization to virologic failure. Virologic failure is defined as either confirmed rebound of viral load after confirmed suppression, or failure to suppress by 24 weeks. An HIV-1 genotype sequence from the protease (PR) and reverse-transcriptase (RT) regions was obtained for each subject at baseline and at time of virologic failure for each subject experiencing this event. The categorical variable at each codon in the genotype sequence is an amino acid, taking one of 20 possible unordered values. Drug-resistant virus is known to manifest with potentially complex mutational patterns in the HIV-1 genotype sequence (Lerma, 1999; Foulkes, 2003). Because sequencing HIV-1 has dropped to a cost that allows its use in routine patient management, identifying those patterns of mutations evident of resistance is potentially useful for patient-specific drug selection. Therefore, an important question during the analysis of ACTG 398 data is what specific genotype changes are associated with the risk of virologic failure. That is, for those subjects observed to experience virologic failure, what components of the joint distribution of HIV-1 genotype sequence differ between baseline and post-baseline populations. Post-baseline genotypes are not available for subjects not observed to experience virologic failure, since there is not enough virus for amplification. For such cases, virologists typically assume no changes in genotype sequence.

The phenotype used to indicate drug resistance will be the endpoint of virologic failure. Subjects who do not experience virologic failure will be assumed to not have drug-resistant virus. Thus, by examining the changes in the joint distribution of HIV-1 genotype sequence between baseline and the time of virologic failure among those subjects who experienced virologic failure, we can identify those simultaneous genotype changes that are apparently important contributors to drug-resistance.

Proposed in this paper is a computationally efficient nonparametric method for approximat-
ing a full parsimonious comparison of two large-scale joint distributions. More specifically, the approximation limits the number of variable/level pairs in a combination at some value $K < R$, where $K$ is chosen by the user. If the largest order of dependency among variables is of order $K$ or less, then limiting the number of variable/level pairs in a combination at $K$ actually results in a full (not approximate) parsimonious comparison of the distributions. This is because differences between groups with respect to prevalence of combinations of more than $K$ variable/level pairs would be redundant in the sense that the difference is completely explained by differences in prevalence of combinations of order $K$ or less. Since in practice, one never knows the largest order of dependency among variables, limiting the number of variable/level pairs in a combination at $K$ in general results in an approximate comparison. Further approximations are also presented that consider only those combinations for which at least a pre-specified number of variable/level pairs have a significant marginal (one-way) group effect. The computer algorithm proposed to execute the method converts a potentially computer-memory intensive method to a computer-processor intensive method, the latter resource being much cheaper than the former.

For both the paired and independent sample cases, the approximate parsimonious comparison method necessarily tests two types of multiple null hypotheses: equal prevalence and conditional independence. For each combination of variable/level pairs under consideration, one first tests if the prevalence of the combination is the same in the two groups. If the null hypothesis of equal prevalence is not rejected, the combination is excluded from further analysis. On the other hand, if the combination is declared to occur more frequently in one group, then a parsimonious comparison method requires testing if the prevalence of the combination can be written as the same product of prevalence’s of lower-order combinations in each group. That is, it is required to test for conditional independence of events defining the combination given group membership. If the hypothesis of conditional independence is not rejected in both groups, then even though the combination was declared to
have different prevalence between groups, the difference is considered redundant and the combination is discarded. This is because the difference in prevalence is completely driven by lower-order combinations. For a given combination, the number of conditional independence null hypotheses to test in each group depends on the number of variable/level pairs in the combination. For example, a combination of two variable/level pairs has one corresponding conditional independence null hypothesis per group, whereas a combination of three variable/level pairs has three corresponding conditional independence null hypotheses per group.

We will use multiple statistical testing procedures to simultaneously make a decision on a sequence of null hypotheses. There will be two types of null hypotheses to test. The first concerns marginal homogeneity and the other conditional independence. These are defined in more detail below. Simultaneous testing of multiple hypotheses requires defining a type I error rate. The simplest is the family-wise error rate (FWER), defined as the probability of making one or more false rejections among all the hypotheses when performing multiple hypothesis tests. In order to control the FWER, \( \alpha \), in an analysis involving more than one model comparison, the error rate for each comparison must be more stringent than \( \alpha \). Boole’s theorem says that the probability of the union of several events occurring is less than or equal to the sum of the probabilities of each event. Namely, the inequality implies that if each test performed is to have type I error rate \( \alpha/Q \), the total error rate will not exceed \( \alpha \). This is known as the Bonferroni correction, and is one of the most naive and commonly used approaches for multiple comparisons. The correction is based on the idea that if an experimenter is testing \( Q \) dependent or independent hypotheses on a set of data, then one way of maintaining the FWER is to test each individual hypothesis at a statistical significance level of \( 1/Q \) times what it would be if only one hypothesis were tested.

Because simple techniques such as the Bonferroni method can be too conservative, there
has been a great deal of attention paid to developing more powerful techniques, such that the overall rate of false positives can be maintained without inflating the rate of false negatives unnecessarily. These methods include Romano and Shaikh step-up methods (2006) and Holm (1979) step-down methods for controlling the FWER. For these methods, total $\alpha$ can be proved to never exceed 0.05 (or some other chosen value) under any conditions. These methods provide "strong" control against Type I error, in all conditions including a partially correct null hypothesis, under arbitrary data-generating distributions.

The goal of the proposed method is to estimate the set of variable/level combinations that have a non-redundant different prevalence between the two populations, in this paper, baseline and post-baseline. For each combination in the estimated set, it is shown that the probability that the combination either has equal prevalence between groups, or corresponds to a redundant group difference, is bounded above asymptotically by any pre-specified level $\alpha$, for any data-generating mechanism.

This paper is set up as follows. Section 1.2 is an overview of comparing joint distributions. Section 1.3 describes the null hypotheses. Marginal homogeneity and conditional independence are discussed in Section 1.4. Multiple testing procedures are explored in Section 1.5. An analysis of ACTG398 and discussions are presented in Sections 1.6 and 1.7, respectively.

1.2 Comparing Joint Distributions

Some specific notations used to describe the data in this section are:
a large R-vector of categorical variables: \( Y_1, \ldots, Y_R \)

\( W \) group membership \( \{0, 1\} \)

\( N^0 \) number of independent realizations of \( Y \) for group 0

\( Y^0_n \) set of independent realizations: \( Y^0_{n,1}, \ldots, Y^0_{n,R} \), for \( n = 1, \ldots, N^0 \)

\( P^0 \) unknown probability distribution for group 0

\( N^1 \) number of independent realizations of \( Y \) for group 1

\( Y^1_n \) set of independent realizations: \( Y^1_{n,1}, \ldots, Y^1_{n,R} \), for \( n = 1, \ldots, N^1 \)

\( P^1 \) unknown probability distribution for group 1

Each of the \( N^1 = N^0 \) independent realizations \( Y^1_n \), for the paired sample case, are from the unknown conditional probability distribution of \( Y \) in group 1 given \( Y^0 = Y^0_n \), denoted \( P^1|Y^0_n, n = 1, \ldots, N^1 \).

The collection of distributions \( P = (P^0, P^1) \) (independent sample case) or \( P = \{P^0, (P^1|y)_y\} \) (paired sample case), where \( y \) is a potential value for \( Y \) is the complete data-generating mechanism. For both the paired and independent sample case, the main goal of this paper is to estimate the parsimonious set of differences between the joint distributions \( P^0 \) and \( P^1 \). The collection \( P \) is considered to belong to a family of probability distributions, say \( \Omega \); the model \( \Omega \) may be nonparametric, parametric, or semiparametric. Only the nonparametric model is considered in this paper.

Let \( y_{r,1}, \ldots, y_{r,L_r} \) denote the possible values of variable \( Y_r, r = 1, \ldots, R \), with \( L_r \geq 2 \).

Define \( J_{r;\ell} = I(Y_r = y_{r,\ell}), \ell = 1, \ldots, L_r, r = 1, \ldots, R, I(.) \) the indicator function. More generally, the variable indicating the \( k \)th-order variable/level combination, \( R \geq k \geq 1 \), is denoted by \( J_{r_1;\ell_1, \ldots, r_k;\ell_k} = J_{r_1;\ell_1} \times \cdots \times J_{r_k;\ell_k} \).

A crude method for comparing \( P^0 \) and \( P^1 \) is one that attempts to identify all those \( R \)th-order variable/level combinations for which \( P(J_{r_1;\ell_1, \ldots, r_R;\ell_R} = 1|W = 0) \) does not equal
\[ P(J_{r_1;\ell_1,\ldots,r_R;\ell_R} = 1 | W = 1). \]

However, if \( P(J_{r_1;\ell_1,\ldots,r_R;\ell_R} = 1 | W) \) can be written as the same product of prevalences of combinations of order \( R - 1 \) or less for \( W = 0, 1 \), then the statement \( P(J_{r_1;\ell_1,\ldots,r_R;\ell_R} = 1 | W = 0) \neq P(J_{r_1;\ell_1,\ldots,r_R;\ell_R} = 1 | W = 1) \) is redundant in the sense that the difference is completely explained in terms of prevalences of lower-order combinations. For example, suppose \( P(J_{r_1;\ell_1,r_2;\ell_2} = 1 | W = 0) \neq P(J_{r_1;\ell_1,r_2;\ell_2} = 1 | W = 1) \), but that

\[ P(J_{r_1;\ell_1,r_2;\ell_2} = 1 | W) = P(J_{r_1;\ell_1} = 1 | W) P(J_{r_2;\ell_2} = 1 | W) \]

for \( W = 0, 1 \). Then the difference between groups in \( P(J_{r_1;\ell_1,r_2;\ell_2} = 1 | W) \) is completely driven by a difference between groups with respect to \( P(J_{r_1;\ell_1} = 1 | W) \) and/or \( P(J_{r_2;\ell_2} = 1 | W) \).

For third-order variable/level combinations, for each group \( W \), there are three ways for \( P(J_{r_1;\ell_1,r_2;\ell_2,r_3;\ell_3} = 1 | W) \) to be written as a product of prevalences of lower-order combinations, namely,

i) \( P(J_{r_1;\ell_1} = 1 | W) P(J_{r_2;\ell_2,r_3;\ell_3} = 1 | W) \), or

ii) \( P(J_{r_2;\ell_2} = 1 | W) P(J_{r_1;\ell_1,r_3;\ell_3} = 1 | W) \), or

iii) \( P(J_{r_3;\ell_3} = 1 | W) P(J_{r_1;\ell_1,r_2;\ell_2} = 1 | W) \).

Case i) states that the event \( J_{r_1;\ell_1} = 1 \) is conditionally independent of the event \( J_{r_2;\ell_2,r_3;\ell_3} = 1 \) given \( W \), similarly for case ii) and iii). For fourth-order variable/level combinations, the seven ways for \( P(J_{r_1;\ell_1,r_2;\ell_2,r_3;\ell_3,r_4;\ell_4} = 1 | W) \) to be written as a product of prevalences of lower-order combinations are given in Appendix I.

The best method for comparing the two distributions \( P^0 \) and \( P^1 \), is a method that first
identifies first-order marginal probabilities, $P(J_{r,\ell} = 1|W)$, that differ between these two
groups. The method then would attempt to identify those second- thru $R$th-order com-
bbinations whose prevalence is different between groups but also cannot be written as the
same product of prevalences of lower order combinations for both groups $W = 0, 1$. It
becomes challenging to consider all possible combinations due to computational ineffi-
ciencies. However, redundant differences between groups are not identified, allowing for
the most parsimonious method.

To ensure this method is practical for comparing $P^0$ and $P^1$, a possible starting point would
be to assign a value for the largest considered combination of variable/level pairs, at some
value $K$, with $R \geq K \geq 1$, e.g. $K = 3$ or $4$. A full comparison of the joint distributions
cannot be obtained due to computational inefficiencies; however, by limiting the number of
combinations to a pre-specified order $K$, then an approximate full parsimonious compari-
on of order $K$ can be obtained. Although the value of $K$ is independent of sample size,
because of dimensionality complications, the practical range for $K$ certainly increases with
sample size. The value of $K$ may be obvious in a particular subject matter of interest, or
determined by the computational and/or data resources at hand.

Although the total number of combinations of variable/level pairs has been limited, the
computational resources necessary to perform the comparisons may still be too demand-
ing. Note that, for each group $W$, the total number of first-order variable/level pairs
is $\sum_{r=1}^{R} L_r$; the total number of first- and second-order variable/level combinations is
$\sum_{r=1}^{R} L_r \{ (\sum_{s=r+1}^{R} L_s) + 1 \}$; and so-on. Thus the total number of combinations rapidly
increases with the combination order. Consider the following motivation to approximating
the set of all second- thru $K$th-order combinations. Suppose some $k$th-order combination
has different prevalence between groups, $K \geq k \geq 2$; it is not necessarily the case that
each of the $k$ marginal (first-order) variable/level pairs has different prevalence between
groups. In fact, it could be the case that none of the $k$ variable/level pairs has a marginal group effect. Thus, those $k$th-order combinations for which at least $\max\{(k - K_0), 0\}$ variable/level pairs are marginally associated with group membership, $K \geq K_0 \geq 0$, represent a subset of all possible $k$th-order combinations and, when $K_0 < k$, constitutes an approximation to this set. For each $k$th-order combination in this subset, at most $K_0$ of the $k$ variable/level pairs have no marginal group association. In a sense $K_0$ can be thought of as degrees of freedom; $K_0 = K$ corresponds to no approximation, while $K_0 = 0$ considers only those $k$th-order combinations for which all $k$ variable/level pairs have a marginal association with group membership.

1.3 Null Hypotheses

Comparing $P^0$ and $P^1$ requires testing two types of multiple null hypotheses: those corresponding to equal prevalence (marginal homogeneity) between groups, and those corresponding to conditional independence of events between groups. After we obtain the p-values for the corresponding multiple null hypotheses, simultaneous testing will be performed with control of a pre-specified error rate.

1.4 Marginal Homogeneity

For $K \geq k \geq 1$, define

$$H_{r_1;\ell_1, \ldots, r_k;\ell_k}^{(k)} : P(J_{r_1;\ell_1, \ldots, r_k;\ell_k} = 1|W = 0) = P(J_{r_1;\ell_1, \ldots, r_k;\ell_k} = 1|W = 1),$$

with $r_i \in (1, \ldots, R)$, $r_i \neq r_j$ ($i \neq j$) and $\ell_i \in (1, \ldots, L_{r_i})$. Each alternative hypothesis can be either two-sided or one-sided, and is denoted by $\overline{P}_{r_1;\ell_1, \ldots, r_k;\ell_k}^{(k)}$.

Marginal homogeneity occurs when the marginal probabilities for one response equals
the corresponding marginal probabilities for the other response. This implies that the off-diagonal joint probabilities are also equal (symmetry) (Agresti, 2007). For each of n matched pairs, let $\pi_{i,j}$ denote the probability of outcome i for the first observation and outcome j for the second. Let $n_{i,j}$ count the number of such pairs with $p_{i,j} = n_{i,j}/n$ the sample proportion. We treat $n_{i,j}$ as a sample from a multinomial $(n; \pi_{i,j})$ distribution. Then, for a 2x2 table, $p_{i,+}$ is the proportion in category i for observation 1, and $p_{+,i}$ is the corresponding proportion for observation 2. We compare samples by comparing marginal proportions $p_{i,+}$ with $p_{+,i}$. With matched samples, these proportions are correlated and methods for independent samples are inappropriate. When $\pi_{1,+} = \pi_{+,1}$, then $\pi_{2,+} = \pi_{+,2}$ also, and there is marginal homogeneity. Since

$$\pi_{1,+} - \pi_{+,1} = (\pi_{11} + \pi_{12}) * (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21},$$

marginal homogeneity in 2 x 2 tables is equivalent to $\pi_{1,2} = \pi_{2,1}$.

For 2 x 2 tables a p-value on the null hypothesis of marginal homogeneity is easily obtained with McNemar’s (1947) test to compare the corresponding dependent proportions. The score test statistic is defined as

$$z_0 = \frac{\pi_{2,1} - \pi_{1,2}}{(\pi_{2,1} + \pi_{1,2})^{1/2}}$$

The square of the score statistic $z_0$ is the test statistic of McNemar’s Test, which has a $\chi^2$ distribution with 1 degree of freedom (df). Therefore, the McNemar’s statistic uses the sample proportions and is defined as

$$\chi^2 = \frac{(p_{1,2} - p_{2,1})^2}{(p_{1,2} + p_{2,1})}$$

Statistical significance is determined by evaluating the probability of $\chi^2$ with reference to a
table of cumulative probabilities of the chi-squared distribution or a comparable computer function. A significant result implies that marginal frequencies (or proportions) are not homogeneous. The test is inherently two-tailed. For a one-tailed test, one could divide the obtained p value by two. The McNemar statistic depends only on cases classified in different categories for the two observations. The $\pi_{1,1} + \pi_{2,2}$ on the main diagonal are irrelevant to inference about whether $\pi_{1,+}$ and $\pi_{+,1}$ differ.

For larger than 2 x 2 tables, the approaches of Bhapkar(1966) or Stuart (1955) can be used to obtain a p-value. Bhapkar(1966) tested marginal homogeneity by exploiting the asymptotic normality of marginal proportions. Let $d_i = p_{i,+} - p_{i,+}$ and let $d' = (d_1, \ldots, d_{J-1})$. It is redundant to include $d_j$ since $\sum d_i = 0$. The sample covariance matrix $\hat{S}$ of $\sqrt{n}d$ has elements

$$\hat{s}_{i,j} = -(p_{i,j} + p_{j,i}) - (p_{+,i} - p_{i,+})(p_{+,j} - p_{j,+})$$

$$\hat{s}_{i,i} = p_{+,i} + p_{i,+} - 2p_{i,i} - (p_{+,i} - p_{i,+})^2$$

Now $\sqrt{n}[d - E(d)]$ has an asymptotic multivariate normal distribution with estimated covariance matrix $\hat{S}$. Under marginal homogeneity, $E(d) = 0$, and

$$Z = nd'\hat{S}^{-1}d$$

is asymptotically chi-squared with degrees of freedom (df) = $J - 1$. Stuart(1955) proposed $W_0 = nd'\hat{S}_0^{-1}d$ which uses the sample null covariance matrix $\hat{S}_0$ and is the score test. This has
\[ s_{i,j} = -(p_{i,j} + p_{j,i}), \quad \text{for } a \neq b \]

\[ s_{i,i} = p_{i,+} + p_{+,i} - 2p_{i,i} \]

Ireland et al. (1969) noted that \( Z = Z_0/(1 - Z_0/n) \). For \( J = 2 \), \( Z_0 \) is McNemar’s test statistic.

These tests use all \( J-1 \) degrees of freedom available for comparisons of \( J \) pairs of marginal proportions. With ordered categories, when \( J \) is large and the dependence between classifications is strong, ordinal tests (with df=1) can be much more powerful (Agresti, 1984).

If there is perfect agreement for any category \( j \), that category must be omitted in order to invert matrix \( S \). (Note that if there is perfect agreement on a category, the corresponding row and column marginal frequencies are equal.) Such categories should be ignored in calculations and the Stuart-Maxwell test performed with respect to the remaining categories. The df in this case can still be considered \( J - 1 \), where \( J \) is the number of original categories; this treats omitted categories as if they were included but contributed 0 to the value of \( \chi^2 \) even though such categories have equal row and column marginals.

To define the null hypothesis for all 1-way, 2-way, etc. changes, let \( X_{r_1} \) and \( X_{r_2} \) equal the value of the combinations of codons at baseline and post-baseline, respectively. Then the null hypotheses for combinations \( 1, \ldots, K \) are as follows:
where JxJ is the dimension of the joint distribution of \((X_{r_1}, \ldots, X_{r_k})\), for \(k = 1, \ldots, K\).

### 1.4.1 Conditional independence (CI)

If the null hypothesis of marginal homogeneity is not rejected, the combination is excluded. If the combination is declared to occur more frequently in one group, we test if the prevalence of the combination can be written as the same product of prevalences of lower-order combinations in each group, also known as CI. When we reject marginal homogeneity, in order to say the rejection is non-redundant, we need to reject CI also. The number of hypotheses to test for CI depends on the order of the combination.

For a second-order combination, there is one hypothesis to test for baseline and one for post-baseline. In order to keep the pair, we would need to reject CI at baseline and post-baseline. Corresponding to each second-order variable/level combination, the null hypothesis is defined as:

\[
Q_{r_1;\ell_1,r_2;\ell_2}^{W,1} : P(J_{r_1;\ell_1,r_2;\ell_2} = 1|W) = P(J_{r_1;\ell_1} = 1|W)P(J_{r_2;\ell_2} = 1|W),
\]

for \(r_i \in (1, \ldots, R)\), \(r_1 \neq r_2\), \(\ell_i \in (1, \ldots, L_{r_i})\) and \(W = 0, 1\).

Note that when \(Q_{r_1;\ell_1,r_2;\ell_2}^{W,1}\) is true, the following three statements are also true:
i) \( P(J_{r_1;\ell_1} = 0, J_{r_2;\ell_2} = 1|W) = P(J_{r_1;\ell_1} = 0|W)P(J_{r_2;\ell_2} = 1|W), \)

ii) \( P(J_{r_1;\ell_1} = 1, J_{r_2;\ell_2} = 0|W) = P(J_{r_1;\ell_1} = 1|W)P(J_{r_2;\ell_2} = 0|W), \) and

iii) \( P(J_{r_1;\ell_1} = 0, J_{r_2;\ell_2} = 0|W) = P(J_{r_1;\ell_1} = 0|W)P(J_{r_2;\ell_2} = 0|W). \)

For a third-order combination, there are three hypotheses to test for baseline and post-baseline. To keep the triplet, we would need to reject baseline and post-baseline for at least one test. Corresponding to each third-order variable/level combination are three null hypotheses of interest in each group \( W = 0, 1 \). These are

\[
Q_{r_1;\ell_1, r_2;\ell_2, r_3;\ell_3}^{W_1} : P(J_{r_1;\ell_1, r_2;\ell_2, r_3;\ell_3} = 1|W) = P(J_{r_1;\ell_1} = 1|W)P(J_{r_2;\ell_2, r_3;\ell_3} = 1|W),
\]

\[
Q_{r_1;\ell_1, r_2;\ell_2, r_3;\ell_3}^{W_2} : P(J_{r_1;\ell_1, r_2;\ell_2, r_3;\ell_3} = 1|W) = P(J_{r_2;\ell_2} = 1|W)P(J_{r_1;\ell_1, r_3;\ell_3} = 1|W),
\]

\[
Q_{r_1;\ell_1, r_2;\ell_2, r_3;\ell_3}^{W_3} : P(J_{r_1;\ell_1, r_2;\ell_2, r_3;\ell_3} = 1|W) = P(J_{r_3;\ell_3} = 1|W)P(J_{r_1;\ell_1, r_2;\ell_2} = 1|W),
\]

for \( r_i \in (1, \ldots, R), r_i \neq r_j, (i \neq j) \) and \( \ell_i \in (1, \ldots, L_{r_i}) \).

When \( Q_{r_1;\ell_1, r_2;\ell_2, r_3;\ell_3}^{W_1} \) is true, the following three statements are also true:

i) \( P(J_{r_1;\ell_1} = 0, J_{r_2;\ell_2, r_3;\ell_3} = 1|W) = P(J_{r_1;\ell_1} = 0|W)P(J_{r_2;\ell_2, r_3;\ell_3} = 1|W), \)

ii) \( P(J_{r_1;\ell_1} = 1, J_{r_2;\ell_2, r_3;\ell_3} = 0|W) = P(J_{r_1;\ell_1} = 1|W)P(J_{r_2;\ell_2, r_3;\ell_3} = 0|W), \) and

iii) \( P(J_{r_1;\ell_1} = 0, J_{r_2;\ell_2, r_3;\ell_3} = 0|W) = P(J_{r_1;\ell_1} = 0|W)P(J_{r_2;\ell_2, r_3;\ell_3} = 0|W). \)

Fisher’s exact test is used to determine whether the proportions of those falling into each category differ by group. For a 2x2 table assuming fixed marginals, the exact probability of being in a given cell is

\[
\frac{(a + b)!((c + d)!(a + c)!(b + d)!)}{n!a!b!c!d!}
\]
where \( a, b, c, d \) are the cell frequencies. To test \( H_0 \): independence, the p-value is the sum of these probabilities for outcomes at least as extreme to the observed frequencies (Agresti, 2007)

The elements of the test statistic for a second-order combination corresponding to \( Q_{r_1;\ell_1,r_2;\ell_2}^{W,1} \), are:

\[
\begin{align*}
a &= NW - \sum_{n=1}^{NW} J_{r_2;\ell_2}^{n,W} - J_{r_1;\ell_1,r_2;\ell_2}^{n,W} + J_{r_1;\ell_1,r_2;\ell_2}^{n,W} \\
b &= \sum_{n=1}^{NW} J_{r_2;\ell_2}^{n,W} - J_{r_1;\ell_1,r_2;\ell_2}^{n,W} \\
c &= \sum_{n=1}^{NW} J_{r_1;\ell_1}^{n,W} - J_{r_1;\ell_1,r_2;\ell_2}^{n,W} \\
d &= \sum_{n=1}^{NW} J_{r_1;\ell_1,r_2;\ell_2}^{n,W}
\end{align*}
\]

The elements of the test statistic for a third-order combination corresponding to \( Q_{r_1;\ell_1,r_2;\ell_2,r_3;\ell_3}^{W,1} \), are:

\[
\begin{align*}
a &= NW - \sum_{n=1}^{NW} J_{r_2;\ell_2,r_3;\ell_3}^{n,W} - J_{r_1;\ell_1,\ell_2,r_3;\ell_3}^{n,W} + J_{r_1;\ell_1,\ell_2,r_3;\ell_3}^{n,W} \\
b &= \sum_{n=1}^{NW} J_{r_2;\ell_2,r_3;\ell_3}^{n,W} - J_{r_1;\ell_1,\ell_2,r_3;\ell_3}^{n,W} \\
c &= \sum_{n=1}^{NW} J_{r_1;\ell_1}^{n,W} - J_{r_1;\ell_1,\ell_2,r_3;\ell_3}^{n,W} \\
d &= \sum_{n=1}^{NW} J_{r_1;\ell_1,\ell_2,r_3;\ell_3}^{n,W}
\end{align*}
\]
Similarly, a two-sided p-value can be calculated for $Q_{r_1,l_1,r_2,l_2,r_3,l_3}^{W,i}$, $i = 2, 3$.

Fisher’s exact test was used here since considerably less data is used compared to the test for equal prevalence. Additionally, one of the row sums in this $2 \times 2$ table may be small, corresponding to sparse data.

### 1.5 Multiple Testing Procedures

We will need to test all of the one-way, two-way, etc., null hypotheses simultaneously using a multiple testing procedure that controls a given error rate. The error rate can be the FWER, the generalized FWER (gFWER), or the proportion of false positives among the rejected null hypotheses (PFP). These error rates and corresponding multiple testing procedures are described in Romano and Shaikh (2006) for step-up testing approaches and Romano and Shaikh (2004) and Lehmann and Romano (2005) for step-down testing approaches.

Null hypotheses to test include:
1-way hypotheses: \( \{H_1, \ldots, H_R\} \)

2-way hypotheses: \( \{H_{i,j} \}, i \neq j \)

3-way hypotheses: \( \{H_{i,j,k} \}, i \neq j \neq k \)

\[ \vdots \]

K-way hypotheses: \( \{H_{i,\ldots,K} \} \)

where K is the largest combination order considered, e.g. with 399 codons there would be \(2^{399}\) hypothesis tests.

Accounting for the multiplicity of individual hypothesis tests can be attained by controlling an appropriate error rate. A classical approach to examining the multiplicity problem is to restrict to procedures that control the FWER, defined to be the probability of one or more false rejections.

Romano and Shaikh (2006) propose step-up procedures that provide strong control, that is, for arbitrary \( P(Y, X) \), of either the number or proportion of true null hypotheses among the rejected hypotheses at level \( \alpha \). If \( PFP \) represents the proportion of false positivies of among the rejected hypotheses (which equals 0 if there are no rejections), then control of the \( PFP \) satisfies \( P(PFP > \phi) \leq \alpha \), where \( \phi \in [0, 1] \) is user defined. Note that choosing \( \phi = 0 \) corresponds to the usual FWER. Although use of the FWER leads to a simple interpretation of the covariate screening result, it can be considerably less powerful
compared to choosing $\phi > 0$. For simplicity, only procedures for controlling the FWER are presented below. Note that it is not necessary to estimate the covariance between p-values to execute any of these testing procedures (whether or not $\phi$ is taken to equal 0). To control the FWER, Romano and Shaikh (2006) propose the following step-up algorithm. Given p-value $\hat{\pi}(q)$ for testing $H(q)$, order the observed p-values in descending order to obtain $\hat{\pi}(Q) \geq \cdots \geq \hat{\pi}(1)$ with corresponding $H(Q), \ldots, H(1)$. If $\hat{\pi}(Q) \leq \frac{\alpha}{a(Q)}$, then reject all $Q$ hypotheses. Otherwise, if $\hat{\pi}(Q) > \frac{\alpha}{a(Q)}$, do not reject $H_Q$ and next consider if $\hat{\pi}(Q-1) > \frac{\alpha}{a(Q-1)}$.

In general, reject hypotheses $H(q^*), \ldots, H(1)$, where $q^*$ is the smallest index satisfying

$$\hat{\pi}(Q) > \frac{\alpha}{a(Q)}, \ldots, \hat{\pi}(q^*+1) > \frac{\alpha}{a(q^*+1)}.$$ 

If, for all $q$, $\hat{\pi}(q) > \frac{\alpha}{a(q)}$, then reject no hypotheses. That is, a step-up procedure begins with the least significant p-value and continues accepting hypotheses as long as their corresponding p-values are large. The formula for the constant $a(Q)$ is provided in Romano and Shaikh (2006) and equals 2.13 for $Q \geq 25$.

Holm (1979) introduced a step-down procedure that provides strong control of the FWER. The algorithm is as follows. A step-down procedure begins with the most significant test statistic and determines which null hypotheses to reject. If $\hat{\pi}(1) > \alpha/Q$, then reject no hypotheses. Otherwise, reject hypotheses $H(1), \ldots, H(q^*)$, where $q^*$ is the largest index such that

$$\hat{\pi}(1) \leq \frac{\alpha}{Q}, \ldots, \hat{\pi}(q^*) \leq \frac{\alpha}{Q - q^* + 1}.$$ 

That is, a step-down procedure begins with the most significant p-value and continues rejecting hypotheses as long as their corresponding p-values are small. Holm’s method is based on the Bonferroni method and is valid regardless of the joint distribution of the test statistics (strong control).
Lehmann and Romano (2005) proposed step-down procedures that provide strong control over the number or proportion of false positives. Romano and Shaikh (2004) proposed step-down methods that provide strong control of the proportion of false positives that can be considerably more powerful than those of Lehmann and Romano (2005).

Note that a valid p-value on each null hypothesis is all that is required to use these multiple testing approaches, an arbitrary dependence structure among the p-values is allowed.

A codon/ amino acid combination that is claimed to violate marginal homogeneity using a multiple testing procedure may be a redundant genotype change in the sense that the violation is completely explained by a lower-order codon/ amino acid combination. For example, if a two-way combination is claimed to violate marginal homogeneity, the violation is completely explained by the corresponding one-way changes if the two pairs in the two-way combination occur independently at baseline and at post-baseline.

Thus, we will need to test independence of the elements in a combination that is claimed to violate marginal homogeneity and discard those combinations that can be written as a product of lower order combinations at both baseline and post-baseline.

1.6 Analysis of ACTG 398

The data used are from AIDS Clinical Trials Group (ACTG) 398, a cohort of 481. ACTG 398 was a randomized, double-blind, placebo-controlled study of saquinavir, indinavir, or nelfinavir added as a second protease inhibitor to the 4-drug class regimen of amprenavir, abacavir, efavirenz, and adefovir dipivoxil in patients with virologic failure. Patients were randomized between October 31, 1998, and April 14, 1999, with a study completion date of April 2000 (Hammer, 2002). Virologic failure was defined as a viral load above 1000
copies/mL while receiving saquinavir, nelfinavir, indinavir, or ritonavir (Hammer, 2002). The main objective of ACTG 398 was to assess whether the addition of a second protease inhibitor to a new 4-drug class regimen including amprenavir would improve virologic response in patients failing protease inhibitor containing regimens. The study incorporated patients recruited from 31 AIDS Clinical Trial Units that were 13 years of age or older, had laboratory documentation of an HIV-1 infection, had prior exposure to a maximum of 3 protease inhibitors from among saquinavir, ritonavir, indinavir, and/or nelfinavir for a cumulative period of protease inhibitor therapy of at least 16 weeks, received the failing protease inhibitor containing regimen at time of screening, and had certain laboratory parameter levels (Hammer, 2002). Patients were also required to be treatment-naive to amprenavir, abacavir, and adefovir dipivoxil. Participants received follow-up visits (including clinical assessment, routine laboratory monitoring, and viral load measurement) at weeks 2, 4, 8, 16, and every 8 weeks until the last participant completed 48 weeks on study (Hammer, 2002). The CD4 cell counts of participants were measured twice at baseline and at weeks 4, 8, 16 and every 8 weeks until the study end (Hammer, 2002). The study found that antiretroviral-experienced patients with advanced immunodeficiency could achieve viral suppression with regimens containing four or five new drugs, with 148 (31%) of the 481 participants achieving a viral load below 200 copies/mL at week 24 (Hammer, 2002).

Those subjects who did not experience virologic failure while on study are assumed to not have developed drug-resistant virus. We assume each subject who experienced virologic failure had a genotype sequence recorded at baseline and at the time of failure. Among those who experienced virologic failure, we compare the joint distribution of HIV-1 genotype sequence between baseline and the time of failure. This is generally a difficult problem since there are 99 codons in the protease region and 300 codons in the RT region. Within an individual there is a great deal of viral heterogeneity since multiple strains of the virus make up the infected population. The most commonly used sequencing methods are able to
identify only the dominant amino acid at each codon or at best detect the presence of amino acid mixtures. If one simplifies the data at each codon to be either mutant or wildtype, then there are $2^{399}$ possible genotype patterns. Thus, it is not feasible to investigate the full joint distribution of genotype changes. Instead, we take a step-up approach that examines all one-way genotype changes, that is, one codon at a time, all two-way changes, that is, all pairs of codons at a time, up to all three-way simultaneous changes.

For ease of exposition we classify each amino acid as either mutant or wildtype. The one-way change at a given codon can be represented as a $2 \times 2$ table of 4 cell counts, with rows the baseline amino acid and columns the amino acid at virologic failure. Next, for two-way joint changes, for each pair of codons, the two-way (joint) change can be represented as a $4 \times 4$ table (16 cells), with rows the baseline codon combinations and columns the codon combinations at virologic failure. Moving on in a similar fashion, when considering the (joint) change of a given triplet of codons, the data can be represented as an $8 \times 8$ table (64 cells).

For any table of frequency counts as described above, we can define the null hypothesis of marginal homogeneity, that is, equal distributions of the codon/amino acid combinations at baseline and virologic failure.

Of the 481 subjects in ACTG 398, 58 participants were missing baseline genotype or never started treatment. The remaining 423 subjects had an HIV-1 genotype sequence recorded that consisted of all 99 codons from the PR region and 300 codons from the RT region. When sequencing the HIV-1 genome, within an individual there is a good deal of viral heterogeneity as multiple strains of the virus make up the infecting population. It is possible that the amino acid value at a given codon is not the same for all viruses in the sample given there is a good deal of viral heterogeneity; when there is no majority amino acid value at
a codon, the value reported by the assay is a mixture of the values observed in the sample. For the purpose of this analysis, all mixtures were labeled other, resulting in 21 possible amino acid levels at each codon. A total of $N^0 = N^1 = 302$ subjects experienced virologic failure while on study and had an HIV-1 genotype recorded at the time of failure.

There were 399 first-order variable/level pairs observed in the data; of these, six were claimed, at the global FWER $\alpha = 0.10$ level, to have different prevalence between baseline and post-baseline. These include the following RT codons: 74, 100, 101, 103, 108, and 190. The largest magnitude of change was from RT103;K to RT103;N, a well-known resistance mutation for the drug Efavirenz, a component of the ACTG 398 treatment regimen. These results can be found in Table 1.

There were $\binom{399}{2} = 79,401$ total marginal homogeneity null hypotheses. Of these, 814 were rejected. Using these, we tested for conditional independence which, at the global 10% FWER, there were two non-redundant second-order combinations rejected, $RT108/RT190$ and $RT181/RT190$. Codon RT181 was not singly more prevalent post-baseline; however, the combination arising from 190;A was more prevalent post-baseline. The empirical conditional probability of RT190;A given RT181;C is 0.53; given RT181;Y, the empirical probability of RT190;A is 0.14, corresponding to a relative risk of about 4.0. RT108 was found to have a higher prevalence post-baseline alone and in combination with RT190;A. RT103 was found to have a higher prevalence post-baseline alone and in combination with RT190;A. However, the empirical risk of RT190;A given RT103;I relative to RT103;V was 0.45, meaning there was a decrease in the risk of RT190;A mutating when RT103 mutated. These results can be found in Tables 2 and 3.

There were $\binom{399}{3} = 10,507,399$ total 3-way combinations. We considered all 3-ways for which at least one had a significant 1- or 2-way significant association, which resulted
in over 544,000 triplets. Of those, about 80,000 rejected marginal homogeneity and 264 rejected conditional independence. Recall that there are three different ways to reject conditional independence for a 3-way combination. At least one of the three tests had to be rejected for baseline and post-baseline for the triplet to reject conditional independence. This resulted in one non-redundant third-order combination, $PR5/RT44/RT190$, that was claimed to have different prevalence between baseline and post-baseline. Table 4 shows these results.

1.7 Discussion

These results identify changes to known resistance mutations to drugs in ACTG 398 regimens among subjects observed to experience virologic failure (Hammer, 2002). This info can be used by clinicians in active patient management by tailoring patient-specific regimens in an attempt to optimize therapeutic response.

Analyses focused on joint changes in HIV RNA including 1-way, 2-way, and 3-way changes. In order to have a full multivariate comparison, one would need to assess the joint changes up to the number of variables available, in this case 399. However, it is likely that combinations greater than four could have sparse data, which may result in lower power to statistically identify changes. As shown previously, even the results presented from 3-way changes had many zero cells, so data from 4-way changes would be rather sparse. So, to ensure statistical power was maintained, an approximate full comparison was estimated up to a pre-defined order $K(3)$.

The findings of these analyses apply to a particular cohort of subjects with a particular drug exposure history, and are not interpretable for the entire population of HIV-1 infected subjects. In particular, subjects were 13 years of age or older, had laboratory documenta-
tion of an HIV-1 infection, had prior exposure to a maximum of 3 protease inhibitors from among saquinavir, ritonavir, indinavir, and/or nelfinavir for a cumulative period of protease inhibitor therapy of at least 16 weeks, receiving the failing protease inhibitor containing regimen at time of screening, and having certain laboratory parameter levels (Hammer, 2002). Patients were also required to be treatment-naive to amprenavir, abacavir, and adefovir dipivoxil. However, the methods used during these analyses can be applied to any cohort of subjects.

Valid inference was ensured over all tests performed by using strong control multiple testing procedures for the FWER. Two types of null hypotheses were needed to be tested for this methodology, marginal homogeneity and conditional independence. We controlled each one separately with its own MTP. Redundant associations were also eliminated by testing for conditional independence.

It is important to identify early bio-markers associated with future treatment failure because even with advances in treatment options, a significant proportion of patients will not respond to treatment. Therefore, timely identification of these bio-markers may have a profound effect on the selection of appropriate care and management of patients. These findings can prove useful to clinicians in helping with drug regimens and therapy management. This ultimately benefits the patient by potentially preventing future morbidity and mortality.
New Statistical Methods to Evaluate and Compare Generalized Logit Models

2.1 Introduction

It is important to identify early biomarkers associated with virologic failure of HIV-1 since, even with advances in antiretroviral therapies, a significant proportion of patients will not achieve viral suppression. Timely identification of treatment failure biomarkers may have a profound effect on the selection of appropriate alternative therapy for patients. Presented in paper 1 was the development of reliable statistical methods to identify simultaneous changes in the joint distribution of HIV-1 genotype between baseline and time of virologic failure, among subjects experiencing virologic failure. Identified were simultaneous genotype changes found to be important contributors to virologic failure. In this paper, utilizing the important baseline and post-baseline codons from paper 1, in addition to clinical variables, the prediction of week 24 RNA levels will be assessed using polychotomous modeling techniques.

Polychotomous models are commonly used to fit nominal data with greater than two categories. A limitation of this methodology is the difficulty to assess model fit. For ordinary logistic regression, it is trivial to calculate numerous model checking statistics such as prediction error when checking prediction between multiple working models. This is difficult to assess for polychotomous modeling, due to the multi-logit structure, and sparse literature is available describing the calculation of prediction error for polychotomous models. It has become common practice to use individualized regressions in place of a polychotomous model due to the ease of computation and interpretation (Begg, 1984). However, this technique estimates regression coefficients separately, reducing the efficiency, and does not ensure that the probability of choosing all possible outcome categories is equal to one.
Due to the high cost of obtaining RNA genotype measures, the primary objective of this paper is to examine if an additional RNA measure at week 4 (referred to as week\(_4\)) assists in the prediction of treatment failure at week 24 (referred to as week\(_{24}\)), thus allowing for the timely implementation of an effective treatment regimen leading to a better prognosis. Furthermore, this paper aims to use the dimension reduction selection from paper 1 to form candidate working models for future virologic response as predicted by clinical and genotypic information, quantify week\(_{24}\) outcome of virologic failure in one of four categories, and extend prediction error estimation techniques for ordinary logistic regression to incorporate a polychotomous structure to derive robust variance estimates.

This paper is organized as follows. The observed data is described in Section 2.2. Sections 2.3.1 and 2.3.2 include proportional odds and polychotomous logistic regression models. Described in Section 2.4 is apparent error for comparing multiple ordinary and polychotomous logistic models. Section 2.5 discusses model checking techniques to assess the adequacy of the fitted models. Model comparisons and controlling an appropriate error rate are discussed in Section 2.6 and an analysis of ACTG 398 is provided in Section 2.7. A simulation study is presented in Section 2.8 followed by a final discussion in Section 2.9.

### 2.2 Data

It is assumed that a random sample of \((y, x)\) is observed for \(n\) individuals from a well defined population, denoted \((y_i, x_i)\), for \(i = 1 \ldots n\). That is, the observations \((y_i, x_i)\) are independent and identically distributed (i.i.d) for \(i = 1 \ldots n\).
2.3 Logistic Regression: Polychotomous Response

While the typical logistic regression analysis models a dichotomous response, logistic regression is also applicable to multilevel responses. The response may be ordinal (no pain, moderate pain, severe pain) or nominal (math, science, English). For ordinal response outcomes, one can model functions called cumulative logits by performing ordered logistic regression using the proportional odds model. For nominal response outcomes, one forms generalized logits and perform a traditional logistic analysis except that you model multiple logits per subpopulation.

2.3.1 Proportional Odds Model

The proportional odds model is a class of generalized linear models used for modeling the dependence of an ordinal response on discrete or continuous covariates. Like in binary and multinomial logistic regression, predictors may be categorical and/or continuous, and the computation of crude or adjusted odds ratios is the typical goal. The unique feature of the proportional odds model is that the odds ratio for each predictor is taken to be constant across all possible collapsings of the response variable.

Let $Y$ denote the response category with $J$ levels denoted $y_1, \ldots, y_J$. Also, denote the probability (mass) density function (PDF) by $\pi_j = P(Y = y_j)$, for $j = 1, \ldots, J$, with $\sum_{j=1}^{J} \pi_j = 1$. When the levels of $Y$ are ordinal, one can take advantage of this property and construct models that are simpler to interpret and possibly more efficient (when the model is properly specified) than models that assume no ordering (e.g., polychotomous logistic regression). Note however that one can always fit a polychotomous logistic regression, whether $Y$ is ordered or not, but proportional odds models require ordered $Y$.

Proportional odds models incorporate the ordinal structure of $Y$ to invoke modeling assumptions that reduce the number of parameters that need to be estimated. For a dichoto-
mous response, a logit function is computed for each subpopulation. For a multi-level response, one creates more than one logit function for each subpopulation. With ordinal data, one can compute cumulative logits, which are based on the cumulative probabilities. For three response levels, one would compute two logits:

$$\text{logit}(\theta_1) = \log \left[ \frac{\pi_1}{\pi_2 + \pi_3} \right], \text{logit}(\theta_2) = \log \left[ \frac{\pi_1 + \pi_2}{\pi_3} \right].$$

These cumulative logits are the log odds of category 1 to categories 2 and 3, and the log odds of category 1 and 2 to category 3. Both log odds focus on more favorable to less favorable response. The proportional odds model takes both of these odds into account.

In general, define the cumulative (probability) distribution function (CDF) of $Y$ as $P(Y \leq y_j) = \pi_1 + \ldots + \pi_j = F_j, j = 1, \ldots, J$, where $F_j = 1$.

Define the cumulative logits:

$$\text{logit}\{P(Y \leq y_j)\} = \log \left\{ \frac{P(Y \leq y_j)}{1 - P(Y \leq y_j)} \right\}$$

$$= \log \left\{ \frac{P(Y \leq y_j)}{P(Y > y_j)} \right\}$$

$$= \log \left\{ \frac{F_j}{1 - F_j} \right\}$$

$$= \log \left\{ \frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J} \right\}, j = 1, \ldots, J - 1$$

Each of the $J - 1$ cumulative logits uses all $J$ response levels, unlike polychotomous logistic regression where each of the $J - 1$ models only considers data from those two response categories.
The $j$th cumulative logit essentially splits the data into two groups, those with outcome levels less than or equal to $y_j$ and those with outcome levels greater than $y_j$.

Let $F_j(x) = P(Y \leq y_j | X = x)$ denote the conditional CDF of $Y$, $j = 1, \ldots, J - 1$.

Consider the $J - 1$ cumulative logit models

$$
\log \left\{ \frac{F_1(x)}{1 - F_1(x)} \right\} = \alpha_1 + \beta_1 x \\
\vdots \\
\left\{ \frac{F_{J-1}(x)}{1 - F_{J-1}(x)} \right\} = \alpha_{J-1} + \beta_{J-1} x
$$

Inverting these relationships yields the logistic functions

$$F_j(x) = \frac{\exp(\alpha_j + \beta_j x)}{1 + \exp(\alpha_j + \beta_j x)}, j = 1, \ldots, J - 1$$

Consider the case when $X = 0$, corresponding to the baseline logit, where logit$\{F_j(0)\} = \alpha_j$ so that $F_j(0) / \{1 - F_j(0)\} = \exp(\alpha_j)$.

Consider a fixed value of $X = x$, then
\[ \frac{F_{J-1}(x)}{1 - F_{J-1}(x)} = \exp(\alpha_j) \exp(\beta_j x) \]

\[ = \left[ \frac{F_{J-1}(0)}{1 - F_{J-1}(0)} \right] \exp(\beta_j x), \ j = 1, \ldots, J - 1 \]

The proportional odds model constrains

\[ \beta_1 = \ldots = \beta_{J-1} = \beta \]

so that it is assumed the effects of covariates are the same on all response categories:

\[ \frac{F_{J-1}(x)}{1 - F_{J-1}(x)} = \left[ \frac{F_{J-1}(0)}{1 - F_{J-1}(0)} \right] \exp(\beta_j x), \ j = 1, \ldots, J - 1. \]

The model implies that for different response values, say \( y_1 \) and \( y_2 \), the model-based logistic function \( F_2(x) \) is simply a shift (to the left or right) of \( F_1(x) \). \( F_1(x), \ldots, F_{J-1}(x) \) all have the same shape, they differ only by location-shifts.

Other link functions on \( \{F_j(x)\} \) besides the logit(.) can be used, such as the complementary log-log link function that corresponds to the proportional hazards model

\[ \log[-\log\{F_j(x)\} - \gamma_j + \theta_x], \quad j = 1, \ldots, J - 1. \]

**Estimation:**
The \(J - 1\) models are estimated simultaneously:

\[
\log\left\{ \frac{F_1(x)}{1 - F_1(x)} \right\} = \alpha_1 + \beta x \\
\log\left\{ \frac{F_2(x)}{1 - F_2(x)} \right\} = \alpha_2 + \beta x \\
\vdots \\
\log\left\{ \frac{F_{J-1}(x)}{1 - F_{J-1}(x)} \right\} = \alpha_{J-1} + \beta x
\]

Assuming that \(\beta_j = \beta\) for all \(j\), this model simplifies to \(\text{logit}(\theta) = \alpha_j + \beta x\).

For a unit increase in \(X\), the odds of a response below the \(j\)th level is multiplied for \(\exp(\beta)\), for any category \(y_j, j = 1, \ldots, J - 1\).

**Inference:**

When \(\beta = 0\), \(Y\) and \(X\) are not associated so that inference may proceed by testing the null hypothesis \(H_0 : \beta = 0\) against \(H_1 : \beta \neq 0\). This may be accomplished with a 1 d.f. likelihood ratio test, a 1 d.f. Wald test, or a 95% CI for \(\beta\).

The differences between proportional odds models and polychotomous logit models are that the CDF is used as response instead of the PDF and the effect of covariates remains constant for each model instead of varying. Polychotomous logit models are described in detail below.
2.3.2 Polychotomous Logistic Regression

When you have nominal response variables, you can also use logistic regression to model your data. Instead of fitting a model to cumulative logits, you fit a model to generalized logits. Recall the probability density function (PDF) of $Y$ is $\pi_j = P(Y = y_j), j = 1, \ldots, J$ with $\sum_{j=1}^{J} \pi_i = 1$.

When $n$ independent realizations of $Y$ are observed with PDF $\{\pi_1 \ldots \pi_J\}$, the number of observations falling into the $j$th level of $Y$, denoted by $n_j$, has the multinomial distribution:

$$P(n_1 = m_1, \ldots, n_J = m_J) = \frac{n!}{m_1! \ldots m_J!} \pi_1^{m_1} \ldots \pi_J^{m_J}$$

The ordinary logistic regression model takes the form

$$\log \left\{ \frac{P(Y = y_1 | X = x)}{P(Y = y_2 | X = x)} \right\} = \log \left\{ \frac{\pi_1(x)}{\pi_2(x)} \right\} = \alpha + \beta x$$

and thus the generalized logit is defined as $\text{logit}(\theta_j) = \log \left[ \frac{\pi_j}{\pi_J} \right]$, for $j = 1, \ldots, J$. A logit is formed for the probability of each succeeding category over the last response category.

Thus, the generalized logits for say a three-level response is

$$\text{logit}(\theta_1) = \log \left[ \frac{\pi_1}{\pi_3} \right], \text{logit}(\theta_2) = \log \left[ \frac{\pi_2}{\pi_3} \right].$$
The model you fit for generalized logits is \( \logit(\theta_j) = \alpha_j + \beta_j x \), where \( j \) indexes the two (or more) logits. This says that there are separate intercept parameters (\( \alpha_j \)) and different sets of regression parameters (\( \beta_j \)) for each logit. The matrix \( x \) is the set of explanatory values for the \( j \)th group. Instead of estimating one set of parameters for one logit function, as in logistic regression for a dichotomous response variable, you are estimating sets of parameters for multiple logit functions. Whereas for the proportional odds model you estimated multiple intercept parameters for the cumulative logit functions but only one set of parameters corresponding to the explanatory variables, for the generalized logits model you are estimating multiple sets of parameters for both the intercept and the explanatory variables.

**Estimation:**

For the polychotomous logistic regression model, we equate the linear component to the log of the odds of a \( j^{th} \) observation compared to the \( J^{th} \) observation. That is, we will consider the Jth category to be the omitted or baseline category, where logits of the first \( J - 1 \) categories are constructed with the baseline category in the denominator.

The polychotomous logistic regression model fits simultaneously the \( J - 1 \) ordinary logistic models:
Simultaneous estimation is more efficient (more powerful, lower variance) than fitting each of the $J - 1$ equations separately. Notice each model has separate coefficient values $(\alpha_j, \beta_j), \ j = 1, \ldots, J - 1$.

It does not matter what category is selected for the reference category since logistic equations can be constructed for any pair of outcome categories given the chosen reference category:

$$
\begin{align*}
\log \left\{ \frac{\pi_1(x)}{\pi_J(x)} \right\} & = \alpha_1 + \beta_1 x \\
\log \left\{ \frac{\pi_2(x)}{\pi_J(x)} \right\} & = \alpha_2 + \beta_2 x \\
& \vdots \\
\log \left\{ \frac{\pi_{J-1}(x)}{\pi_J(x)} \right\} & = \alpha_{J-1} + \beta_{J-1} x
\end{align*}
$$

That is, given that the response falls into either category $Y = a$ or $Y = b$, the log odds that the response is $Y = a$ when $X = x$ is $(\alpha_a - \alpha_b) + (\beta_a - \beta_b)x$.  

---

34
Inference:

We would test the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against the alternative hypothesis $H_1 : H_0$ fails to hold (at least one $\beta_j \neq 0$).

Suppose we are interested in estimating the conditional probabilities $\pi_j(x) = P(Y = y_j | X = x)$.

The logistic equations can be inverted to obtain

$$
\pi_j(x) = \frac{\exp(\alpha_j + \beta_j x)}{\sum_{i=j}^{J} \exp(\alpha_i + \beta_i x)} \quad j = 1, \ldots, J
$$

where $(\alpha_j, \beta_j)$ corresponding to the reference category are taken to be $(0,0)$, e.g. $(\alpha_J, \beta_J) = (0,0)$ so that

$$
\pi_j(x) = \frac{\exp(\alpha_j + \beta_j x)}{\exp(\alpha_1 + \beta_1 x) + \ldots + \exp(\alpha_{J-1} + \beta_{J-1} x)} \quad j = 1, \ldots, J.
$$

Estimated conditional probabilities are obtained by substituting MLEs from the unknown population regression coefficients. In order to find the MLE’s of the logistic equation we must find the set of parameters for which the probability of the observed data is greatest. The maximum likelihood equation is derived from the probability distribution of the dependent variable. Since each $y_i$ represents a multinomial count in the $i^{th}$ population, the joint probability density function (PDF) of $Y$ is:
\[ f(y|\beta) = \prod_{i=1}^{N} \left[ \frac{n_i!}{\prod_{j=1}^{J} y_{ij}!} \prod_{j=1}^{J} \pi_{ij}^{y_{ij}} \right] \]

The likelihood function is algebraically equivalent to the PDF, the only difference being that the likelihood function expresses the unknown values of \( \beta \) in terms of known fixed constant values for \( y \). Since we want to maximize the PDF with respect to \( \beta \), the factorial terms that do not contain any of the \( \pi_{ij} \) terms can be treated as constants. For a detailed explanation of the iterative procedure of Newton-Raphson for parameter estimation as it applies to polychotomous regression models, see Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation, Czepiel, section 2.2.3.

### 2.4 Apparent Error

#### 2.4.1 Ordinary Logistic Regression

Commonly in statistics and model building, one wishes to know: How do we measure the adequacy of a model? The natural approach is to look at the average squared error in predicting a future response, commonly called the prediction error or apparent error.

For a response category \( Y \) with levels \( y_1, \ldots, y_n \), for \( i = 1, \ldots, n \), and covariates \( x_1, \ldots, x_n \), the apparent error is defined as

\[ D(\hat{\beta}) = AE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2, \quad (1) \]

where \( \hat{Y}_i = \hat{E}[Y|X_i] \)

\[ \equiv y_1 \ast \hat{P}(Y = y|X = x_i) \quad (2) \]
To compare the prediction error between two models to assess which models fit the data best, hypothesis testing may be utilized.

The average prediction error can be defined as

\[
\hat{\Phi}_0 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i^0)^2 = D(\hat{\beta}_0)
\]

and

\[
\hat{\Phi}_1 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i^1)^2 = D(\hat{\beta}_1),
\]

for models 0 and 1, respectively, where \(\hat{Y}_i^0\) is the prediction of \(Y_i\) from model 0 and \(\hat{Y}_i^1\) is the prediction of \(Y_i\) from model 1. See below for a more formal definition. Model 1 (competing model) is assumed to have more predictors than model 0 (reference model).

The null hypothesis \(H_0: \Phi_0 = \Phi_1\) can be tested against the alternative hypothesis \(H_1: \Phi_0 > \Phi_1\) where \(\Phi_0\) represents the population average squared prediction error for model 0 and \(\Phi_1\) represents the population average squared prediction error for model 1. In order to obtain a test statistic concerning the significance for this estimate in the null hypothesis, calculate the ratio of the difference in sample estimation of these parameters divided by an estimate of the corresponding standard error:

\[
w = \frac{\hat{\Phi}_1 - \hat{\Phi}_0}{\sqrt{\text{var}(\hat{\Phi}_1 - \hat{\Phi}_0)}},
\]

This statistic has approximately a standard normal distribution in large samples (Tian et al., 2007). Alternatively, we can treat the square of this statistic as approximately a chi-squared
with one degree of freedom.

Combining the equations in (3) and (4) we get

$$\hat{\Phi}_1 - \hat{\Phi}_0 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\gamma}_i^1)^2 - (Y_i - \hat{\gamma}_i^0)^2$$ (6)

$$= \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{y}_i^1(\hat{\beta}_1))^2 - (Y_i - \hat{y}_i^0(\hat{\beta}_0))^2$$ (7)

$$= \frac{1}{n} \sum_{i=1}^{n} A_i - A_0$$ (8)

$$= \frac{1}{n} \sum_{i=1}^{n} A_i$$ (9)

It is easy to obtain an estimate for $\Phi_1 - \Phi_0$. However, the variance is a little more complicated. Define working model 0 as

$$\text{logit}\left[P(Y = 1|x_i)\right] = \beta_0^0 + (\beta_0^1)^\top x_i^0$$

and

$$\beta^0 = (\beta_0^0, \beta_1^0)$$

Let $\hat{\beta}^0$ be the m.l.e. of $\beta^0$. Similarly for model 1,

$$\text{logit}\left[P(Y = 1|x_i)\right] = \beta_0^1 + (\beta_1^1)^\top x_i^1$$

and

$$\beta^1 = (\beta_0^1, \beta_1^1)$$

with m.l.e. $\hat{\beta}^1$. So, what is the variance of $\hat{\Phi}_1 - \hat{\Phi}_0$, i.e.,

$$\text{vår} \left[ \frac{1}{n} \sum (Y_i - \hat{y}_i^1)^2 - \frac{1}{n} \sum (Y_i - \hat{y}_i^0)^2 \right]$$ (10)
Assuming the data above are i.i.d., fitting the data to the logistic equation yields

\[ \hat{Y}_i^0 = g((\beta_1^\top x_i^0)) \]

and

\[ \hat{Y}_i^1 = g((\beta_1^\top x_i^1)), \]

where \( g(\cdot) \) represents the link function, \( g(u) = \frac{e^u}{1 + e^u} \) the anti-logit link. Note that other link functions can be used, e.g. the proportional hazard link, \( g(u) = 1 - \exp(\exp(u)) \) and the probit link, \( g(u) = \Phi^{-1}(u) \). The form of the m.l.e is given below.

Consider the following estimating equation for the regression parameters. For simplicity of exposition, assume a scalar covariate:

\[
\begin{pmatrix}
0 \\
0
\end{pmatrix} = U\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \left[ Y_i - \hat{E}(Y|X_i) \right] = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \left[ Y_i - \frac{e^{x_i \hat{\beta}}}{1 + e^{x_i \hat{\beta}}} \right], \tag{11}
\]

where \( \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) \), \( x_i = (1 \ x_i) \), and \( U(\cdot) \) is the first derivative of the log-likelihood when \( g(\cdot) \) is the canonical link function, represented here as the anti-logit link. \( U(\cdot) \) is not the first derivative of the log-likelihood for other links. However, it is acceptable to use \( U(\cdot) \) as an estimating equation for all link functions \( g(\cdot) \) since the resulting \( \hat{\beta} \) have good statistical properties. This results in estimators that converge in probability to a well defined vector even if the working model is incorrect (Tian et al., 2007).

The delta method uses first-order Taylor series expansions to approximate the variance of a function of one or more random variables. The Taylor series expansion of a function \( f(\cdot) \) about a value \( a \) as \( x \to a \) is given as:

39
\[ f(x) = f(a) + f'(a)(x-a) + f''(a) \frac{(x-a)^2}{2!} + \ldots, \]

where we can often drop the higher order terms to give the approximation:

\[ f(x) \approx f(a) + f'(a)(x-a). \]

Letting \( f(x) = U(\hat{\beta}_0, \hat{\beta}_1) \), then

\[ \begin{pmatrix} 0 \\ 0 \end{pmatrix} = U(\beta_0, \beta_1) + (\hat{\beta}_0 - \beta_0) U'(\beta_0, \beta_1), \]

where \( U'(\beta_0, \beta_1) \) is the first derivative of \( U \) with respect to \( \beta_0 \) and \( \beta_1 \), which results in

\[ -U(\beta_0, \beta_1) \left[ U'(\beta_0, \beta_1) \right]^{-1} = (\hat{\beta}_0 - \beta_0, \hat{\beta}_1 - \beta_1). \]  

(12)

Referring to equation (1), the variance of

\[ D(\hat{\beta}) = \text{var} \left( D(\hat{\beta}_0, \hat{\beta}_1) \right) = \text{var} \left( \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - e^{X_i \hat{\beta}} \right)^2 \right), \]

where \( \hat{\beta} \) was calculated using \( U(\cdot) \).

Assuming there is a random sample of data from a given population, then \( \{Y_i, X_i\}_{i=1}^{n} \) are i.i.d.. A first-order Taylor series expansion can be applied to \( D(\hat{\beta}) \), resulting in:

\[ D(\hat{\beta}_0, \hat{\beta}_1) = D(\beta_0, \beta_1) + (\hat{\beta}_0 - \beta_0) D'(\beta_0, \beta_1) + \text{Op}(1), \]

where \( D'(\beta_0, \beta_1) \) is the first derivative of \( D \) with respect to \( \beta_0 \) and \( \beta_1 \), and \( \text{Op}(1) \) denotes the terms that converge to 0 in probability.

Substituting for \( (\hat{\beta}_0 - \beta_0, \hat{\beta}_1 - \beta_1) \) from (12) results in:

\[ D(\hat{\beta}_0, \hat{\beta}_1) \approx D(\beta_0, \beta_1) + \left[ -U(\beta_0, \beta_1) \left[ U'(\beta_0, \beta_1) \right]^{-1} \right] D'(\beta_0, \beta_1), \]
with the objective of obtaining the variance of \( D(\hat{\beta}) \), approximately equal to:

\[
\text{var} \left( D \left( \frac{\beta_0}{\beta_1} \right) + \left[ -U' \left( \frac{\beta_0}{\beta_1} \right) \left[ U' \left( \frac{\beta_0}{\beta_1} \right) \right]^{-1} \right] D' \left( \frac{\beta_0}{\beta_1} \right) \right),
\]

where \( U'(\beta) \) is the square matrix of first-order partial derivatives of \( U(\beta) \) also known as the Hessian, so that

\[
U'(\beta) = \frac{\partial U(\beta)}{\partial \beta} = -\frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{x_i} \right) \left( 1 \times x_i \right) \left[ \frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}} \cdot \frac{1}{1 + e^{x_i' \beta}} \right],
\]

and

\[
D'(\beta) = \frac{\partial D}{\partial \beta} = -\frac{1}{n} \sum_{i=1}^{n} 2 \left( Y_i - \frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}} \right) \left[ \frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}} \cdot \frac{1}{1 + e^{x_i' \beta}} \right] \left( 1 \times x_i \right)
\]

If \( D(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} A_i(\hat{\beta}) \), where \( A_i = A^1_i - A^0_i \) and are i.i.d. since we are expanding around \( \hat{\beta} \), then

\[
\text{var}(D(\hat{\beta})) \approx \text{var} \left( \frac{1}{n} \sum_{i=1}^{n} A_i \right)
\]

\[
= \frac{1}{n^2} \text{var} \sum_{i=1}^{n} A_i(\hat{\beta})
\]

\[
= \frac{1}{n^2} \sum_{i=1}^{n} \text{var} A_i(\hat{\beta})
\]

\[
= \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2
\]

\[
= \frac{1}{n^2} n \sigma^2
\]

\[
= \sigma^2 / n,
\]

where \( \sigma^2 = \text{var}(A_i) \). To estimate \( \sigma^2 \), take the sample variance of \( \{A_1, \ldots, A_n\} \), denoted
Each \( A_i \) takes the form

\[
A_i = (Y_i - \hat{Y}_i)^2 - \left( \frac{1}{x_i} \right)^\top (Y_i - \hat{Y}_i(\beta)) \left[ U'(\beta) \right]^{-1} D'(\hat{\beta})
\]

so that \( \tilde{\text{var}}(A_i) = \frac{1}{n} \sum_{i=1}^n A_i^2 - \left( \frac{1}{n} \sum_{i=1}^n A_i \right)^2 \).

### 2.4.2 Comparing Average Squared Prediction Error Between Polychotomous Binary Regression Models

In the context of polychotomous logistic regression, we use the results from section 5.1. Namely it will be shown how to obtain estimates of prediction error and corresponding variance estimators.

Suppose the outcome of interest \( Y \) has 4 possible outcomes:

\[
(i) = \log \frac{P(Y = 2|x)}{P(Y = 1|x)} = \beta_1^{(1)} + \beta_1 \mathbf{X}^\top = \tilde{\mathbf{X}}^\top \beta^{(1)}
\]

\[
(ii) = \log \frac{P(Y = 3|x)}{P(Y = 1|x)} = \beta_2^{(2)} + \beta_2 \mathbf{X}^\top = \tilde{\mathbf{X}}^\top \beta^{(2)}
\]

\[
(iii) = \log \frac{P(Y = 4|x)}{P(Y = 1|x)} = \beta_3^{(3)} + \beta_3 \mathbf{X}^\top = \tilde{\mathbf{X}}^\top \beta^{(3)}
\]

where \( \tilde{\mathbf{X}}^\top = (1 \ x_i) \), \( \beta^{(j)} = (\beta_0^{(j)}, \beta_j) \), and \( (i), (ii), \) and \( (iii) \) represent the 3 ordinary logistic models that are simultaneously being fit. Allowing \( Y = 1 \) to be the reference group, the conditional probabilities \( \pi_j(x) = P(Y = y_j|X = x) \) are:

\[
P(Y = 1|X = x) = \frac{1}{1 + \exp(\tilde{\mathbf{x}}^\top \beta^{(1)}) + \exp(\tilde{\mathbf{x}}^\top \beta^{(2)}) + \exp(\tilde{\mathbf{x}}^\top \beta^{(3)})},
\]
\[ P(Y = 2 | X = x) = \frac{\exp(\tilde{x}^\top \beta^{(1)})}{1 + \exp(\tilde{x}^\top \beta^{(1)}) + \exp(\tilde{x}^\top \beta^{(2)}) + \exp(\tilde{x}^\top \beta^{(3)})}, \]

\[ P(Y = 3 | X = x) = \frac{\exp(\tilde{x}^\top \beta^{(2)})}{1 + \exp(\tilde{x}^\top \beta^{(1)}) + \exp(\tilde{x}^\top \beta^{(2)}) + \exp(\tilde{x}^\top \beta^{(3)})}, \]

\[ P(Y = 4 | X = x) = \frac{\exp(\tilde{x}^\top \beta^{(3)})}{1 + \exp(\tilde{x}^\top \beta^{(1)}) + \exp(\tilde{x}^\top \beta^{(2)}) + \exp(\tilde{x}^\top \beta^{(3)})}. \]

These can be estimated by substituting \( \hat{\beta}^{(j)} \) for \( \beta^{(j)} \) everywhere, where \( \beta^{(j)} \) are defined below, and denoted \( \hat{P}(Y = y | X = x) \).

For polychotomous logistic regression, the apparent error is defined as
\[ D(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2, \]
similar to ordinary logistic regression but where \( \hat{Y}_i = \hat{E}[Y | x] \)
\[ = \hat{P}(Y = 1 | X = x)+2*\hat{P}(Y = 2 | X = x)+3*\hat{P}(Y = 3 | X = x)+4*\hat{P}(Y = 4 | X = x). \]

Similarly, \( U(\cdot) \) is a vector consisting of 3 vectors \( U_1(\cdot), U_2(\cdot), U_3(\cdot) \), where
\[ 0 = U_1(\hat{\beta}_0^{(1)}, \hat{\beta}_1^{(1)}) = \frac{1}{n} \sum_{i=1}^{n} (1 x_i) \{ Y_i^{(1)} - g(\hat{\beta}^{(1)} \tilde{x}_i) \} I(Y_i \leq 2), \]
\[ 0 = U_2(\hat{\beta}_0^{(2)}, \hat{\beta}_1^{(2)}) = \frac{1}{n} \sum_{i=1}^{n} (1 \mathbf{x}_i) \{ Y_i^{(2)} - g(\hat{\beta}_0^{(2)} \tilde{\mathbf{x}}_i^\top) \} I(Y_i = 1) I(Y_i = 3), \]

\[ 0 = U_3(\hat{\beta}_0^{(3)}, \hat{\beta}_1^{(3)}) = \frac{1}{n} \sum_{i=1}^{n} (1 \mathbf{x}_i) \{ Y_i^{(3)} - g(\hat{\beta}_0^{(3)} \tilde{\mathbf{x}}_i^\top) \} I(Y_i = 1) I(Y_i = 4) \]

where \( Y_i^{(j)} = 1 \) if \( Y_i = j + 1 \), \( Y_i^{(j)} = 0 \) if \( Y_i = 1 \), for \( j = 1, 2, 3 \), and \( I(\cdot) \) represents the indicator function, i.e. \( I(A) = 1 \) if the event A occurs, 0 otherwise.

Therefore in this example, \( U'(\cdot) \) is the analog of (13) adapted to this polychotomous setting, where \( U'(\beta) = \)

\[
\begin{bmatrix}
\frac{\partial U_1(\cdot)}{\partial \beta_1} & \frac{\partial U_1(\cdot)}{\partial \beta_2} & \frac{\partial U_1(\cdot)}{\partial \beta_3} \\
\frac{\partial U_2(\cdot)}{\partial \beta_1} & \frac{\partial U_2(\cdot)}{\partial \beta_2} & \frac{\partial U_2(\cdot)}{\partial \beta_3} \\
\frac{\partial U_3(\cdot)}{\partial \beta_1} & \frac{\partial U_3(\cdot)}{\partial \beta_2} & \frac{\partial U_3(\cdot)}{\partial \beta_3}
\end{bmatrix}
\]

This matrix is a diagonal matrix since \( \frac{\partial U_j(\cdot)}{\partial \beta_k(\cdot)} = 0 \) for \( j \neq k \).

\[
\frac{\partial U_1}{\partial \beta_1} = -\frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{\mathbf{x}_i^\top} \right) (1 \mathbf{x}_i) g(\tilde{\mathbf{x}}_i \beta^{(1)}) [1 - g(\tilde{\mathbf{x}}_i \beta^{(1)})] I(Y_i \leq 2)
\]

\[
\frac{\partial U_2}{\partial \beta_2} = -\frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{\mathbf{x}_i^\top} \right) (1 \mathbf{X}_i) g(\tilde{\mathbf{x}}_i \beta^{(2)}) [1 - g(\tilde{\mathbf{x}}_i \beta^{(2)})] I(Y_i = 1) I(Y_i = 3)
\]

\[
\frac{\partial U_3}{\partial \beta_3} = -\frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{\mathbf{x}_i^\top} \right) (1 \mathbf{X}_i) g(\tilde{\mathbf{x}}_i \beta^{(3)}) [1 - g(\tilde{\mathbf{x}}_i \beta^{(3)})] I(Y_i = 1) I(Y_i = 4)
\]
Furthermore, 

\[ D(\beta) = \left( \frac{\partial D}{\partial \beta_1}, \frac{\partial D}{\partial \beta_2}, \frac{\partial D}{\partial \beta_3} \right), \]

where, when evaluated at the estimators \( \hat{\beta}^{(j)} \), equals

\[
\frac{\partial D}{\partial \beta_1} = -\frac{2}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i) \frac{\partial \hat{Y}_i}{\partial \beta_1},
\]
\[
\frac{\partial D}{\partial \beta_2} = -\frac{2}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i) \frac{\partial \hat{Y}_i}{\partial \beta_2},
\]
\[
\frac{\partial D}{\partial \beta_3} = -\frac{2}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i) \frac{\partial \hat{Y}_i}{\partial \beta_3},
\]

and

\[
\frac{\partial \hat{Y}_i}{\partial \beta_1} = \frac{1}{S^2} + \frac{2 - S - \exp(\hat{\beta}^{(1)} \tilde{x}_i)}{S^2} - \frac{3 \exp(\hat{\beta}^{(2)} \tilde{x}_i)}{S^2} - \frac{4 \exp(\hat{\beta}^{(3)} \tilde{x}_i)}{S^2}
\]

\[
\times \tilde{x}_i \exp(\hat{\beta}^{(1)} \tilde{x}_i)^	op
\]

\[
\frac{\partial \hat{Y}_i}{\partial \beta_2} = \frac{1}{S^2} - \frac{2 \exp(\hat{\beta}^{(1)} \tilde{x}_i)}{S^2} + \frac{3 \exp(\hat{\beta}^{(2)} \tilde{x}_i)}{S^2} - \frac{4 \exp(\hat{\beta}^{(3)} \tilde{x}_i)}{S^2}
\]

\[
\times \tilde{x}_i \exp(\hat{\beta}^{(2)} \tilde{x}_i)^	op
\]

\[
\frac{\partial \hat{Y}_i}{\partial \beta_3} = \frac{1}{S^2} - \frac{2 \exp(\hat{\beta}^{(1)} \tilde{x}_i)}{S^2} - \frac{3 \exp(\hat{\beta}^{(2)} \tilde{x}_i)}{S^2} + \frac{4 \exp(\hat{\beta}^{(3)} \tilde{x}_i)}{S^2}
\]

\[
\times \tilde{x}_i \exp(\hat{\beta}^{(3)} \tilde{x}_i)^	op
\]

where \( S = (1 + \exp(\hat{\beta}^{(1)} \tilde{x}_i) + \exp(\hat{\beta}^{(2)} \tilde{x}_i) + \exp(\hat{\beta}^{(3)} \tilde{x}_i)). \)
The expanded formulas above can be presented as

\[
D(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - \hat{Y}_i]^2
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} [Y_i - \hat{Y}_i(\hat{\beta}^{(j)})]^2
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} A_i(\hat{\beta}^{(j)}),
\]

where \(\beta^{(j)} = \begin{pmatrix} \hat{\beta}^{(1)} \\ \hat{\beta}^{(2)} \\ \hat{\beta}^{(3)} \end{pmatrix} \).

then,

\[
D(\hat{\beta}) = D(\beta) + (\hat{\beta} - \beta)D'(\beta) + \text{op}(1)
\]

\[
\approx D(\beta) + (\hat{\beta} - \beta)D'(\beta).
\]

It follows that

\[
\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} U_1(\hat{\beta}^{(1)}) \\ U_2(\hat{\beta}^{(2)}) \\ U_3(\hat{\beta}^{(3)}) \end{pmatrix} = U(\hat{\beta}) \approx U(\beta) + (\hat{\beta} - \beta)U'(\beta),
\]

resulting in

\[
-U(\beta) = (\hat{\beta} - \beta)U'(\beta)
\]

\[
-U(\beta)[U'(\beta)]^{-1} = (\hat{\beta} - \beta).
\]

Therefore, \(D(\hat{\beta}) = D(\beta) + (-U(\beta)[U'(\beta)]^{-1}) [D'(\beta)^\top] \), if \(\beta\) is replaced with \(\hat{\beta}\) everywhere, since \(\hat{\beta} \xrightarrow{p} \beta\) as \(n \to \infty\).

Simplifying, if \(D(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} A_i(\hat{\beta})\), then \(\text{var}(D(\hat{\beta})) = \text{var}(\frac{1}{n} \sum_{i=1}^{n} A_i(\hat{\beta}))\), which
results in $\frac{\sigma^2}{n}$, where $\sigma^2$ is taken to be the sample variance, so that

$$\text{var} \left( \sum_{i=1}^{n} A_i(\hat{\beta}) \right) = \frac{1}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} A_i^2(\hat{\beta}) - \left( \frac{1}{n} \sum_{i=1}^{n} A_i(\hat{\beta}) \right)^2 \right].$$

### 2.5 Model Checking

The Hosmer-Lemeshow (1980) (HL) test is a statistical test for goodness of fit for logistic regression models. This test divides subjects into deciles based on predicted probabilities, then computes a chi-square from observed and expected frequencies. It tests the null hypothesis that there is no difference between the observed and predicted values of the response variable. Therefore, when the test is not significant we cannot reject the null hypothesis and say that the model fits the data well. This test statistic is defined as:

$$\sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)}$$

where

- $n_j = \text{number of observations in the } j^{\text{th}} \text{ group}$
- $O_j = \sum_i y_{ij} = \text{observed number of cases in the } j^{\text{th}} \text{ group}$
- $E_j = \sum_i \hat{p}_{ij} = \text{expected number of cases in the } j^{\text{th}} \text{ group}$

Given that the HL test depends on the estimated probabilities from the model, it depends heavily on the number of new groups and the calculating algorithm and thus different implementations might lead to different conclusions regarding the fit of the model (Kuss, 2002). Groups can be equal in size (i.e. percentiles of predicted probabilities) or be of equal prediction-intervals (i.e. 0-10%, 10-20%, ..., 90-100%). Most statistical packages use percentiles of groups to calculate the HL statistic, because grouping by intervals may result in groups with very few or no observations at all. Unfortunately, the HL statistic has
little power to discover mis-calibration in small samples (Vergouwe, 2010).

Another method for testing model fit is to look at cumulative sums of residuals. Residuals have commonly been used to assess the fit of regression models. However, conventional residual analysis based on the plots of raw residuals or their smoothed curves is highly subjective and most other numerical goodness-of-fit tests provide little information about the nature of model misspecification. In this paper, more objective and informative model-checking techniques by Lin et al. are utilized to assess the adequacy of the fitted model. These techniques utilize cumulative sums of residuals over certain covariates. These methods can be used for a variety of statistical models and data structures, including generalized linear models with independent or dependent observations. Through computer simulation of observations, the distributions of these stochastic processes under the assumed model can be approximated by the distributions of certain zero-mean Gaussian processes. Each observed process can then be compared, both graphically and numerically, with a number of realizations from the Gaussian process allowing for an objective assessment into whether a trend seen in a residual plot reflects model misspecification or natural variation. The techniques developed by Lin et al. are particularly useful in checking the functional form of a covariate and the link function (Lin et al., 2002).

Let the model for the mean be $g^{-1}(\mu_i) = x_i' \beta$, here $g^{-1}(u) = \logit(u) = \log(\frac{u}{1-u})$ and $\mu_i$ is the mean of the response $y_i$ and $x_i$ is the vector of covariates for the $i^{th}$ observation. Denote the raw residual resulting from the fitted model as

$$e_i = y_i - \hat{\mu}_i = y_i - g(\tilde{x}_i \hat{\beta})$$

where $\hat{\mu}_i = \frac{e^{\tilde{x}_i \hat{\beta}}}{1 + e^{\tilde{x}_i \hat{\beta}}}$, and let $x_{ij}$ be the value of the $j^{th}$ covariate in the model for observation $i$. 

48
To check the functional form of the $j^{th}$ covariate, cumulative sums of residuals may be utilized with respect to $x_{ij}$,

$$W_j(x) = n^{-1/2} \sum_{i=1}^n I(x_{ij} \leq x)e_i,$$

where $I(\cdot)$ is the indicator function. Here, the $e_i's$ are cummulatively summed by covariate $x$, and $w_j(x)$ is the sum of the residuals with values of $x_j$ less than or equal to $x$.

Recall the score function from (11) also denoted as $U(\beta) = \sum_{i=1}^n h(x'\beta)x_i(y_i - g(x'\beta))$ from above and $h(u) = 1$.

Recall $U'(\beta)$ from (13) as $\frac{\partial U(\beta)}{\partial \beta}$. Define

$$\hat{W}_j(x) = n^{-1/2} \sum_{i=1}^n \left[ I(x_{ij} \leq x) + \nu'(x^T; \hat{\beta})[U'(\hat{\beta})]^{-1}(\hat{\beta})x_i \right] e_i Z_i$$

where $Z_i$ are independent $N(0, 1)$ random variables and $\nu'(x; \beta) = \sum_{i=1}^n I(x_{ij} \leq x)\frac{\partial \nu(x'\beta)}{\partial \beta}$, which is the cumulative sum through $x$ of $g(x\beta)[1 - g(x\beta)]x_i^T$. Then the conditional distribution of $\hat{W}_j(x)$, given $(y_i, x_i), i = 1, \ldots, n$, under the null hypothesis $H_0$ that the model for the mean is correct, is the same asymptotically, as $n \to \infty$, as the unconditional distribution of $W_j(x)$ (Lin et al., 2002).

Realizations can be approximated from the null hypothesis distribution of $W_j(x)$ by repeatedly generating normal samples $Z_i, i = 1, \ldots, n$, while holding $(y_i, x_i), i = 1, \ldots, n$ at their observed values and computing $\hat{W}_j(x)$ for each sample.

Functional form of covariate $j$ can be assessed by plotting a few realizations of $\hat{W}_j(x)$ on the same plot as the observed $W_j(x)$ and visually comparing to see how typical the ob-
served $W_j(x)$ is of the null distribution samples.

A formal test of the null hypothesis that $x_j$ has the correct functional form can be the Kolmogorov-type supremum test. Let $s_j$ be the observed value of $S_j = sup_x|W_j(x)|$. The p-value $Pr[S_j \geq s_j]$ is approximated by $Pr[\hat{S}_j \geq s_j]$, where $\hat{S}_j = sup_x|\hat{W}_j(x)|$. $r[\hat{S}_j \geq s_j]$ is estimated by generating realizations of $\hat{W}_j(\cdot)$.

The link function can also be assessed instead of the $j^{th}$ covariate by using values of the linear predictor $x'_i\hat{\beta}$ in place of the values of the $j^{th}$ covariate $x_{ij}$. The graphical and numerical methods described previously are then sensitive to inadequacies in the link function.

### 2.6 Model Comparisons

Traditional approaches for model comparisons include Akaike's Information Criteria (AIC), Schwarz Bayesian Criteria (SC), and likelihood ratio tests (-2 Log L) which are all likelihood based measures that estimate a measure of the difference between a given model and the true underlying model (Agresti, 2002). The model with the smallest AIC, SC, or LRT among all competing models is deemed the best model. The likelihood ratio test statistic is defined as:

$$-2\log L = -2 \sum w_j f_j \log(\hat{p}_j),$$

where $w_j$ and $f_j$ are the weight and frequency values of $j^{th}$ observation and $\hat{p}_j$ is the estimated event probability.

**AIC:**

$$= -2 \log L + 2(k + s),$$

where $k$ is the total number of response levels minus one, and $s$ is the number of explanatory effects.
SC:

\[-2 \log L + (k + s) \log \sum_j (f_j).\]

The -2 Log Likelihood statistic has a chi-square distribution under the null hypothesis (that all the explanatory effects in the model are zero). The AIC and SC statistics give two different ways of adjusting the 2 Log Likelihood statistic for the number of terms in the model and the number of observations used. These statistics are used when comparing different models for the same data. Lower values of the statistic indicate a more desirable model.

Our goal is to conduct model comparisons that do not require proper specification of likelihoods for valid inference. The methods use a 2-step Delta method to derive consistent variance estimates of prediction error estimates. Multiple hypothesis testing procedures are employed that maintain a chosen error rate at some pre-specified level for arbitrary data generating estimates, or so-called strong control.

Multiple testing refers to any instance that involves the simultaneous testing of several hypotheses. Multiple testing procedures adjust p-values derived from multiple statistical tests to correct for the occurrence of false positives. Suppose there are several model comparisons of interest denoted \( H_1 : \Phi_0^{(1)} = \Phi_1^{(1)}, \ldots, H_Q : \Phi_0^{(Q)} = \Phi_1^{(Q)} \). To obtain a test statistic for each hypothesis, define \( W_q = \frac{\hat{\Phi}_0^{(q)} - \hat{\Phi}_1^{(q)}}{\sqrt{\text{var}(\hat{\Phi}_0^{(q)} - \hat{\Phi}_1^{(q)})}}, q = 1, \ldots, Q \) where \( \text{var}(\hat{\Phi}_0^{(q)} - \hat{\Phi}_1^{(q)}) \) is the sample variance of \( \{A_{i,0}^{(q)} - A_{i,1}^{(q)}\}, i = 1, \ldots, n \) and \( A_{i,j}^{(q)} \) is defined for \( j = (0, 1), q = 1, \ldots, Q \) as taking the form shown in (14). The marginal p-values \( \hat{\pi}_q = P(W_q < Z) \) can be obtained where \( Z \sim N(0, 1) \). With \( \hat{\pi}_1, \ldots, \hat{\pi}_Q \) we can simultaneously test \( H_1, \ldots, H_Q \) using the following multiple testing procedures (MTP).

Accounting for the multiplicity of individual hypothesis tests can be attained by controlling an appropriate error rate. A classical approach to examining the multiplicity problem
is to restrict to procedures that control the familywise error rate (FWER), defined to be the probability of one or more false rejections.

Romano and Shaikh (2006) propose step-up procedures that provide strong control, that is, for arbitrary $P(Y, X)$, of either the number or proportion of true null hypotheses among the rejected hypotheses at level $\alpha$. If $PFP$ represents the proportion of false positives among the rejected hypotheses (which equals 0 if there are no rejections), then control of the $PFP$ satisfies $P(PFP > \phi) \leq \alpha$, where $\phi \in [0, 1]$ is user defined. Note that choosing $\phi = 0$ corresponds to the usual FWER. Although use of the FWER leads to a simple interpretation of the covariate screening result, it can be considerably less powerful compared to choosing $\phi > 0$. For simplicity, only procedures for controlling the FWER are presented below. Note that it is not necessary to estimate the covariance between p-values to execute any of these testing procedures (whether or not $\phi$ is taken to equal 0). To control the FWER, Romano and Shaikh (2006) propose the following step-up algorithm. Given p-value $\hat{\pi}(q)$ for testing $H(q)$, order the observed p-values in descending order to obtain $\hat{\pi}(Q) \geq \cdots \geq \hat{\pi}(1)$ with corresponding $H(Q), \ldots, H(1)$. If $\hat{\pi}(Q) \leq \frac{\alpha}{a(Q)}$, then reject all $Q$ hypotheses. Otherwise, if $\hat{\pi}(Q) > \frac{\alpha}{a(Q)}$, do not reject $H_Q$ and next consider if $\hat{\pi}(Q-1) > \frac{\alpha}{a(Q-1)}$.

In general, reject hypotheses $H_{(q^*)}, \ldots, H_{(1)}$, where $q^*$ is the smallest index satisfying

$$\hat{\pi}(Q) > \frac{\alpha}{a(Q)}, \ldots, \hat{\pi}(q^*+1) > \frac{\alpha}{a(q^* + 1)}.$$  

If, for all $q$, $\hat{\pi}(q) > \frac{\alpha}{a(q)}$, then reject no hypotheses. That is, a step-up procedure begins with the least significant p-value and continues accepting hypotheses as long as their corresponding p-values are large. The formula for the constant $a(Q)$ is provided in Romano and Shaikh (2006) and equals 2.13 for $Q \geq 25$.  

52
Holm (1979) introduced a step-down procedure that provides strong control of the FWER. The algorithm is as follows. A step-down procedure begins with the most significant test statistic and determines which null hypotheses to reject. If \( \hat{\pi}(1) > \alpha/Q \), then reject no hypotheses. Otherwise, reject hypotheses \( H(1), \ldots, H(q^*) \), where \( q^* \) is the largest index such that

\[
\hat{\pi}(1) \leq \frac{\alpha}{Q}, \ldots, \hat{\pi}(q^*) \leq \frac{\alpha}{Q - q^* + 1}.
\]

That is, a step-down procedure begins with the most significant p-value and continues rejecting hypotheses as long as their corresponding p-values are small. Holm’s method is based on the Bonferroni method and is valid regardless of the joint distribution of the test statistics (strong control).

Lehmann and Romano (2005) proposed step-down procedures that provide strong control over the number or proportion of false positives. Romano and Shaikh (2004) proposed step-down methods that provide strong control of the proportion of false positives that can be considerably more powerful than those of Lehmann and Romano (2005).

### 2.7 Application

The data used are from AIDS Clinical Trials Group (ACTG) 398, a cohort of 481 HIV-infected patients. ACTG 398 was a randomized, double-blind, placebo-controlled study of saquinavir, indinavir, or nelfinavir added as a second protease inhibitor to the 4-drug class regimen of amprenavir, abacavir, efavirenz, and adefovir dipivoxil in patients with virologic failure. Patients were randomized between October 31, 1998, and April 14, 1999, with a study completion date of April 2000 (Hammer, 2002). Virologic failure was defined as a viral load above 1000 HIV-1 RNA copies/mL while receiving saquinavir, nelfinavir, indinavir, or ritonavir (Hammer, 2002). The main objective of ACTG 398 was to assess
whether the addition of a second protease inhibitor to a new 4-drug class regimen including amprenavir would improve virologic response in patients failing a protease inhibitor containing regimen. The study incorporated patients recruited from 31 AIDS Clinical Trial Units, who were 13 years of age or older, had laboratory documentation of an HIV-1 infection, had prior exposure to a maximum of 3 protease inhibitors from among saquinavir, ritonavir, indinavir, and/or nelfinavir for a cumulative period of protease inhibitor therapy of at least 16 weeks, receiving the failing protease inhibitor containing regimen at time of screening, and having certain laboratory parameter levels (Hammer, 2002). Patients were also required to be treatment-naive to amprenavir, abacavir, and adefovir dipivoxil. Participants received follow-up visits (including clinical assessment, routine laboratory monitoring, and viral load measurement) at weeks 2, 4, 8, 16, and every subsequent 8 weeks until the last participant completed 48 weeks on study (Hammer, 2002). The CD4 cell counts of participants were measured twice at baseline and at weeks 4, 8, 16 and every subsequent 8 weeks until the study end (Hammer, 2002). The study found that antiretroviral-experienced patients with advanced immunodeficiency could achieve viral suppression with regimens containing four or five new drugs, with 148 (31%) of the 481 participants achieving a viral load below 200 copies/mL at week 24 (Hammer, 2002).

In this paper, the response variable Y is the HIV-1 RNA measure at week 24, an important measure of the patient’s response to treatment from baseline administration.

As previously described in *Statistical Methods for Identifying and Modeling Patterns of HIV-1 Genotype Evolution*, conditional independence tests were utilized to predict week 24 RNA based on significant codons found using MTP for 1-way, 2-way, and 3-way changes in HIV-1 RNA. Codons found to have a statistically significant change between baseline and post-baseline values included *PR*5, *RT*44, *RT*74, *RT*100, *RT*101, *RT*103, *RT*108, *RT*181, and *RT*190. In an attempt to improve the accuracy of predicting week 24 RNA
(\(RNA_{24}\)), additional week 4 genotype variables were included in a forward step-wise logistic regression model. For this analysis, patients with complete records for the above mentioned important predictors were included (339/481). The observed Y, HIV-1 RNA, for these patients ranged from 2.301 to 6.620 on the log scale. Variables included in these efforts were the previously described significant baseline codons (\(RT44_{(0)}\), \(RT74_{(0)}\), \(RT100_{(0)}\), \(RT101_{(0)}\), \(RT103_{(0)}\), \(RT108_{(0)}\), \(RT181_{(0)}\), \(RT190_{(0)}\)), and several clinical variables (baseline and week 4 RNA and baseline and week 4 CD4, denoted RNA\(_{(0)}\), RNA\(_{(4)}\), CD4\(_{(0)}\) and CD4\(_{(4)}\), respectively). The outcome variable, \(RNA_{24}\), was dummy coded into four groups (levels) based on log RNA values; 1 = RNA ≤ 2.302, 2 = 2.302 < RNA ≤ 3.1, 3 = 3.1 < RNA ≤ 4.3, 4 = RNA > 4.3. Using the forward stepwise regression in a proportional odds model, several baseline and week four variables were identified as predictive of \(RNA_{24}\): \(RT100_{(0)}\), \(RT103_{(0)}\), \(CD4_{(0)}\), RNA\(_{(0)}\), \(RT103_{(4)}\) and RNA\(_{(4)}\). However, the score test for the proportional odds assumption was significant (p-value = < 0.0001) indicating the model does not hold and suggesting that separate parameters are needed across the logits for a least one predictor. Therefore, polychotomous logistic regression was implemented utilizing the variables identified as significant predictors in the proportional odds model. With these results various working models (M0 - M4) were posited to investigate if adding week 4 RNA improves model fit. Each level of \(RNA_{24}\) was modeled in a separate logistic model, where the number of models is equal to one less than the number of levels as discussed in Section 2.3.2. Predictors for these models are described as continuous baseline CD4 and RNA, categorical week 4 RNA (1 = RNA ≤ 2.302, 2 = 2.302 < RNA ≤ 4.3, 3 = RNA > 4.3), and categorical baseline codons RT100, RT103, and week 4 RT103.

Model M4 differs from models M0-M3 by not only taking into account results from the proportional odds model, but instead looking at the effect of all significant codons as represented from *Statistical Methods for Identifying and Modeling Patterns of HIV-1 Genotype Evolution*, including higher order 2-way and 3-way interactions. The significant 1-way,
2-way, and 3-way codons included in model $M_4$ are: $RT_{74}$, $RT_{100}$, $RT_{101}$, $RT_{103}$, $RT_{108}$, $RT_{190}$, $RT_{103}/RT_{190}$, $RT_{181}/RT_{190}$, and $PR_{5}/RT_{44}/RT_{190}$, respectively. It is of interest to see if modeling all the codons in a full model as they are from Paper 1, are predictive of $RNA_{24}$. To assess whether the addition of week 4 information improves model fit, separate logistic models were created, each adding in more week 4 predictors (Table 5).

For simplicity, the logit models for $M_3$ are shown here:

\[
\ln \left( \frac{P(Y = 2 | X)}{P(Y = 1 | X)} \right) = \alpha_{01} + \beta_{01} (\text{baseline CD4}) + \beta_{02} (\text{baseline RNA}) \\
+ \beta_{03} (\text{baseline RT100}) + \beta_{04} (\text{baseline RT103}) \\
+ \beta_{05} (\text{week 4 RT103}) + \beta_{06} (\text{week 4 RNA})
\]

\[
\ln \left( \frac{P(Y = 3 | X)}{P(Y = 1 | X)} \right) = \alpha_{11} + \beta_{11} (\text{baseline CD4}) + \beta_{12} (\text{baseline RNA}) \\
+ \beta_{13} (\text{baseline RT100}) + \beta_{14} (\text{baseline RT103}) \\
+ \beta_{15} (\text{week 4 RT103}) + \beta_{16} (\text{week 4 RNA})
\]

\[
\ln \left( \frac{P(Y = 4 | X)}{P(Y = 1 | X)} \right) = \alpha_{21} + \beta_{21} (\text{baseline CD4}) + \beta_{22} (\text{baseline RNA}) \\
+ \beta_{23} (\text{baseline RT100}) + \beta_{24} (\text{baseline RT103}) \\
+ \beta_{25} (\text{week 4 RT103}) + \beta_{26} (\text{week 4 RNA})
\]

As described in Section 2.5, to better assess model fit, functional form of continuous variables and the link function were assessed by the use of cumulative residuals. This showed that the functional form of all continuous variables and the link function fit well.

Statistical and graphical representations of the functional form of continuous covariates and the logit link are presented in Table 6 and Figures 1-12. With the exception of model
M0, level 3 (1 vs 4), the logit link fits well. In addition, given the large covariate structure of model M4, separating the model into three binary models was not appropriate. Instead, levels 1 and 2 were compared to levels 3 and 4, to create one binary model. These results are also in Table 7.

From here, model comparisons using adapted techniques proposed in section 2.4.2 and MTP are performed to identify which models with different predictors are statistically different from each other. Prediction error results can be found in Table 8.

Interestingly, as week 4 information is added to the baseline model, the prediction error decreases, indicating that the more week 4 information that is used to predict $R.N.A_{24}$, the lower the error rate. Also, there is a dramatic decrease in the prediction error for the full model. For ease of interpretation, both the squared and absolute difference are shown.

All models were found to fit well based on HL. Furthermore, calculated odds ratios revealed interesting results. Compared to those with week 4 RNA levels below detection, those with high RNA levels (>10,000 copies/mL), have almost 14 times the odds (CI: 2.2,86.8) of having a high (level 4) week 24 RNA level compared to an undetectable week 24 RNA measure (level 1) (Table 9). Additionally, when looking at week 4 RT103, compared to those with wild-type, those with mutant have 50 times the odds (CI: 2.1,>100.0) of having a high (level 4) week 24 RNA level compared to an undetectable week 24 RNA measure (level 1) (Table 9).

These numeric results are supported graphically (Figures 13 and 14). Figure 13 shows the plotted predicted probabilities for baseline 4 RNA, and Figure 14 shows the predicted probabilities for week 4 RNA. Where there is an effect of baseline RNA measures on week 24 RNA, it is much more evident that the probability of being in level 1 for week 24 RNA
decreases as week 4 RNA values increase, showing the effect of week 4 RNA on week 24 RNA. Also notable, as baseline and week 4 RNA measures increase, the probability of being in level 4 for week 24 RNA increases. Figures 15 and 16 show the plotted predicted probabilities for baseline and week 4 RT103, respectively. As week 4 RT103 goes from wildtype to mutant, the probability of being in level 4 for week 24 RNA increases, and the probability of being in level 1 for week 24 RNA decreases. Arguably, baseline RT103 shows the opposite effect: there is a higher probability of having high week 24 RNA levels for baseline RT103 being wildtype. However, this may be due to Simpson’s Paradox, in the sense that a trend is seen for the marginal effect of baseline RT103, but if stratified by drug experience, this effect may disappear.

In all, there were \( \binom{5}{2} = 10 \) model comparisons for testing the null hypothesis that the prediction error between any two given models is equal, \( H_0 : \Phi_0^{(q)} = \Phi_1^{(q)} \), against the alternative \( H_1 : \Phi_0^{(q)} > \Phi_1^{(q)} \). Test results can be found in Table 10.

After applying the Holm step-down multiple testing correction, comparisons 3, 4, 7, 9 and 10 are rejected (indicated with a single *). Similar results were found using the Romano and Shaikh (2006) step-up method. In addition, the Lehmann and Romano (2005) step-down method for controlling the number or proportion of false positives was conducted and rejected an additional two comparisons, 1 and 8 (indicated with a double **).

### 2.8 Simulation

In order to demonstrate performance of results from the previous section and to numerically validate the analytic results, we perform a limited simulation study to evaluate properties of the derived test statistic for calculating prediction error in polychotomous models, where the measure for prediction error is denoted \( D \). We calculate the empirical distribution and
variance of $D$ for a specific set of fixed covariates by repeated sampling of $n$ observations from a polychotomous model with conditional probabilities

$$\pi_j(x) = P(Y = y_j|X = x) = \frac{\exp(\alpha_j + \beta_j x)}{\sum_{i=j}^J \exp(\alpha_i + \beta_i x)} \quad j = 1, \ldots, J$$

where $J = 4$ and $X$ consists of two covariates, $X_1, X_2$. Given four levels for the outcome, three beta’s are needed for each covariate. The beta’s used are as follows:

<table>
<thead>
<tr>
<th></th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>log(1)</td>
<td>log(1.5)</td>
<td>log(2)</td>
</tr>
<tr>
<td>X2</td>
<td>log(1)</td>
<td>log(1/2)</td>
<td>log(1/3)</td>
</tr>
</tbody>
</table>

The intercept, $\alpha$ is chosen to be 0 for all levels of $Y$ indicating a uniform baseline distribution.

At each iteration, we calculate the average prediction error ($D$) and variance. In practice, consistent estimates of the variance are desired in order to make inferences about prediction error estimates. To examine the effects of sample size, we conducted simulations with $n = 200, 500, 800, \text{ and } 2000$. In order to make the simulation study feasible, we considered three polychotomous models with covariates as follows:

$$S1 = X_1$$
$$S2 = X_2$$
$$S3 = \{X_1, X_2\}$$

Note that models $S1$ and $S2$ omit a covariate and thus are mis-specified, model $S3$ is the properly specified model. The goal of the simulation study was to show first that the measure we derived for estimating prediction error is capable of choosing the most parsimonious model with the minimum prediction error with probability approaching 1 as $n \to \infty$. 

59
Second, in order to do inference, we need to assess the accuracy of the variance estimators for $D$.

A total of 500 independent simulation iterations were performed to ensure the derived analytical results are supported numerically. Results of the simulation are presented in Table 11. Across all sample sizes, model $S3$ presents with the lowest prediction error and is the correct model. Additionally, model $S3$ consistently had the smallest average prediction error as well as the smallest prediction error for each iteration and, as $N$ gets larger, this proportion approaches, and even reaches, 1. Meaning, that model $S3$ will have the lowest prediction error for every iteration 100% of the time as $N$ gets larger. For assessing the variance estimates for $D$, as $N$ gets large, the variance approaches 1 indicating appropriate variance estimates were used.

2.9 Discussion

The early identification of biomarkers associated with virologic failure is vital to decreasing the high proportion of individuals who do not respond to treatment effectively. Timely identification of these biomarkers may have a profound effect on the selection of appropriate antiviral treatment of patients. Presented here was compelling evidence that the addition of week 4 information is important to the prediction of week 24 RNA measures. Resistance mutations to drugs utilized in ACTG 398 among subjects observed to experience virologic failure were identified at 4 weeks of treatment. This information is important and can be utilized during clinical management of HIV-1 patients.

This research project utilized information found to have significant changes between baseline and post-baseline genotypes. Week 4 information was added to multiple modeling techniques to assess if the addition of this information improved the ability to predict week
24 RNA measures. Proportional odds models identified that the proportional odds assumption failed and was not the appropriate statistical modeling technique for these data. The failure of the proportional odds assumption was due to the fact that there is at least one covariate for which the effect on all levels of the outcome variable RNA\textsubscript{24} is not proportional, i.e., the slopes for all the covariates are equal across logit equations. The variables found to be significant in the proportional odds model were felt to be important predictors and were utilized in polychotomous models.

The addition of week 4 information to polychotomous modeling was found to be significant in the prediction of week 24 RNA measures, thus potentially allowing for the early identification of virologic failure. Polychotomous model checking was performed using multiple techniques indicating results presented here are accurate and from appropriate models. Although the p-value for the link function for model $M_0$ (level 3) was significant, this can be expected given a 5% type 1 error rate. One alternative to using the logistic link function would be to use the probit link function and use the normal CDF

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

However, this would be computationally more intensive and result in minimal to no gain in model accuracy.

Comparing prediction errors across models interesting results, showing that as week 4 information was added, the prediction error decreased with the full model having the lowest prediction error. Hypothesis tests were generated to identify which models were statistically different from each other. Valuable information about adding in week 4 information and using the full model was gained from these analyses. It was also apparent that models with more week 4 predictors, both genotype and clinical, fit better than models with less
information. More importantly, it is evident that model $M_4$ fits statistically better than any other model. The included interactions in this model were extremely important in lowering the prediction error drastically. This shows that the codons found to be statistically significant in previous research are important in the prediction of week 24 RNA measures, but the interaction of these codons is the most important information. Biologically, codons interact with each other and the addition of these interactions greatly improves the prediction of week 24 RNA measures.

In addition, a large contribution of this paper is the adaptation of a distribution free model comparison methodology by extending the ordinary logistic prediction error to the poly- 
chotomous setting. Through simulation, we demonstrated the adaptation of this technique performed well, producing good prediction error estimates with appropriate variance estimates. Valid inference was ensured with consistent variance estimation even when models were mis-specified. The implementation of this technique has far reaching uses beyond the scope of this paper and can be used in other polytomous outcome settings.

Furthermore, the information found during these research efforts are essential to clinicians monitoring treatment regimens for patients with HIV-1. The addition of week 4 information is vital to predicting week 24 RNA measures, thus allowing for the identification of imminent virologic failure. Using this information, clinicians may alter the treatment course leading to a better outcome for patients and potentially preventing virologic failure and associated morbidity and mortality. The findings of this paper support the current NIH treatment guidelines, *Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents*, which propose viral load testing at 2-8 weeks after initiating or changing antiretroviral therapy, with HIV-1 RNA genotyping performed if viral load remains non-suppressed. The early identification of virologic failure may assist in placing patients on the appropriate treatment regimens quickly.
3 New Statistical Methods to Evaluate and Compare Survival Models

3.1 Introduction

Many medical survival studies have the objective of identifying prognostic factors that can be used to guide clinical management of patients (Schemper, 2000). Regression models are an important technique used by statisticians to approximate the relationship between a set of covariates and the variable of interest. Statistical modeling has two main objectives: to explain and to predict the outcome, based on the information provided by individuals in the data. Most statisticians focus on the estimation of parameters and interpretability of the model. However, most clinicians and non-statistical researchers want a model that is efficient, easily built and interpreted, and, most importantly, predicts well. A model may fit the data at hand well but cannot be generalizable to the general population of individuals who are not in the data. Utilized in this paper is survival analysis methodology to estimate prediction error. Examining prediction errors allows us to evaluate the performance of a model with respect to predictions. A model that predicts well is expected to produce low prediction errors. The prediction errors are generally computed as the squared difference between a future response and its prediction from the fitted model. The focus of this paper is survival models, also known as failure time models, which provide statistical methods for the analysis and prediction of data when the variable of interest is the time to an event.

A possible estimate of the prediction error is the proportion of errors from the fitted model compared to the original data points. In ordinary logistic regression models, the prediction error may be obtainable through various statistical methods involving cross validation and bootstrapping. However, survival models have the added complication of censoring, which could make prediction error estimation arduous.
Estimation of prediction error for survival models has been examined previously (Lawless and Yuan, 2010; Schumacher, Binder and Gerds, 2007; Henderson and Keiding, 2005; Henderson, Jones, and Stare, 2001). The main advantage of the new measure developed is that it properly accounts for censoring in prediction error estimates which is an important component to survival analysis. Furthermore, simulation methods are used to test the new method on generated data.

This paper is setup as follows. Section 3.2 describes the setting of survival data and censoring. Section 3.3 provides some definitions of common notations of survival analysis including the survivor, hazard, and cumulative hazard functions. Sections 3.4 describes the Kaplan-Meier estimator for estimating the survival function and the log-rank test for comparing the survival distributions for two or more groups. The Cox proportional hazards model is described in Section 3.5. Section 3.6 provides a section on prediction error estimates including the adaptation of the Brier score (1950) by Gerds and Schumacher (2006) to survival data and our proposed estimator of prediction error for survival data. Section 3.7 depicts goodness of fit and various numerical and graphical model checking techniques. Performing model comparisons and hypothesis testing is described in Section 3.8. Application to an AIDS clinical trials data set is performed in Section 3.9 followed by a simulation/validation study and discussion in sections 3.10 and 3.11, respectively.

3.2 Survival Data

In many medical studies an outcome of interest is the time to an event. Time to event, also called survival time, is the time since the beginning of follow-up, after enrollment for an individual, until the time to the first observation of the event of interest. A sample of individuals from the study population is followed for a period of time until the event occurs, which could be any specific experience of an individual in the study. In the simplest case
the event is death, but it could also include diagnosis of disease, tumor development, or cancer remission. Survival data results from a study with a limited period of follow-up, such as a clinical trial or cohort study, and requires the observation of an event to ensure the appropriate association is identified.

Naturally, due to the limited period of follow-up, a key feature to survival analysis is censored observations, defined as an observation with incomplete information. There are four different types of censoring possible: right truncation, left truncation, right censoring and left censoring. For this analysis, we will focus exclusively on right censoring. When an observation is right censored it means that the information is incomplete because the subject did not have an event during the time the subject was part of the study. Utilizing survival analysis, subjects may be followed over time and observed at which point in time an event occurred. Often, studies do not span enough time in order to observe the event in question for all the subjects under study. This may be due to study drop-out related to death, non-compliance, or lost to follow-up. Incomplete observations resulting from the previous examples may be due to the time-limited nature of scientific studies, and if subjects were able to be followed for an indefinite amount of time, then the event of interest could eventually be observed.

Even if the event was not observed during the study period, the censored subjects bring some information about the actual survival time. We know that it is longer than the time to the end of study, longer than the time to loss to follow-up, or longer than the time to withdrawal. In these situations, the censored time is recorded instead of the actual survival time.
3.3 Notation

Suppose $T$ is a non-negative random variable ($T \geq 0$) representing the time until some event of interest. For example, $T$ might denote the elapsed time from diagnosis of a disease until death, the elapsed time between administration of a vaccine and development of an infection, or the elapsed time from the start of treatment of a symptomatic disease and the suppression of symptoms. We shall assume $T$ is continuous (i.e. not discrete/categorical) unless we specify otherwise. The probability density function (PDF) and cumulative distribution function (CDF) are most commonly used to characterize the distribution of any random variable, and we shall denote these by $f(\cdot)$ and $F(\cdot)$, respectively,

$$PDF : f(t)$$
$$CDF : F(t) = P(T \leq t)$$

where $F(0) = P(T = 0)$.

Because $T$ is non-negative and usually denotes the elapsed time until an event, it is commonly characterized in other ways as well:

**Survivor function:**

$$S(t) = 1 - F(t) = P(T > t), \text{ for } t > 0$$

The survivor function simply indicates the probability that the event of interest has not yet occurred by time $t$; thus, if $T$ denotes time until death, $S(t)$ denotes probability of surviving beyond time $t$. 

66
Note that, for an arbitrary $T$, $F(\cdot)$ and $S(\cdot)$ as defined above are right continuous in $t$. For continuous survival time $T$, both functions are continuous in $t$. However, even when $F(\cdot)$ and $S(\cdot)$ are continuous, the non-parametric estimators, say $\hat{F}(\cdot)$ and $\hat{S}(\cdot)$, that we will consider are discrete distributions.

**Hazard function:**

$$h(t) = \lim_{\Delta \downarrow 0} \frac{P[t \leq T < t + \Delta | T \geq t]}{\Delta} = \lim_{\Delta \downarrow 0} \frac{P[t \leq T < t + \Delta]}{\Delta} = \frac{f(t)}{S(t-)}$$

with $S(t-) = \lim_{s \uparrow t} S(s)$. That is, the hazard function is a conditional density, given that the event has not occurred prior to time $t$. Note that for continuous $T$, $h(t) = -\frac{d}{dt} \ln[1 - F(t)] = -\frac{d}{dt} \ln S(t)$.

**Cumulative hazard function:**

$$H(t) = \int_0^t h(u)du \quad t > 0$$

$$= -\ln[1 - F(t)] = -\ln S(t)$$

Note that

$$S(t) = e^{-H(t)}$$

$$f(t) = h(t)e^{-H(t)}$$
and \( h(t)dt = f(t)dt/S(t) \approx P(\text{fail in } [t, t + dt] \mid \text{survive until } t) \)

Thus, the hazard function might be of more intrinsic interest than the PDF to a patient who had survived a certain time period and wanted to know something about his/her prognosis.

3.3.1 Common Families of Survival Distributions

**Exponential Distribution:** denoted \( T \sim NE(\lambda) \). For \( t > 0 \),

\[
\begin{align*}
  f(t) & = \lambda e^{-\lambda t} \text{ for } \lambda > 0 \text{ (scale parameter)} \\
  F(t) & = 1 - e^{-\lambda t} \quad S(t) = e^{-\lambda t} \\
  h(t) & = \lambda \text{ constant hazard function} \\
  H(t) & = \lambda t \\
  E(t) & = \frac{1}{\lambda} \\
  V(t) & = \frac{1}{\lambda^2}
\end{align*}
\]

Note: the exponential distribution has the ”lack of memory” property meaning the probability of surviving another \( t \) time units does not depend on how long you’ve lived so far:

\[ P[T > t] = P[T > t + t_0|T > t_0] \]

Also, the exponential family is closed to scale changes, that is: \( T \sim NE(\lambda), c > 0 \Rightarrow c \cdot T \sim NE(\lambda/c) \)

**Exponential Distribution:**
The Weibull distribution can be viewed as a generalization of the exponential distribution, and is denoted $W(\lambda, p)$. It is defined as follows:

\[
F(t) = 1 - e^{(-\lambda t)^p}
\]
\[
f(t) = p\lambda t^{p-1} e^{(-\lambda t)^p}
\]
\[
h(t) = p\lambda t^{p-1} \text{ (power of } t)\]
\[
H(t) = (-\lambda t)^p
\]

for $t > 0$, $\lambda > 0$ (scale), $p > 0$ (shape).

When the shape parameter $p$ equals 1, the Weibull distribution reduces to the exponential distribution. When $p > 1$, the hazard function is increasing; when $p < 1$ it is decreasing.

### 3.4 Kaplan-Meier (KM) Estimator

We now consider non-parametric estimation of the survivor function $S(\cdot)$ based on $n$ i.i.d. survival times that can be non-informatively right censored. The resulting estimator, known as the Kaplan-Meier Estimator, is probably one of the most commonly-used estimators in medical/public health studies involving failure time data.

Suppose that $T_1, T_2, \ldots, T_n$ are i.i.d. survival times with survivor function $S(\cdot)$, and suppose that our observations are denoted $(U_i, \delta_i)$ for $i = 1, 2, \ldots, n$. $U_i = \min(T_i, C_i)$ is the observed survival portion and ($\delta_i = 1$) indicates that $U_i = T_i$, ($U_i = 0$) indicates that $U_i = C_i$, i.e. ($T_i > U_i$).

For simplicity, let us suppose that the survival time $T$ is discrete (categorical) with possible values $0 \leq v_1 \leq v_2 < \ldots$
Define the discrete hazard functions:

\[
h_1 = P[T = v_1] = P[T = v_1 | T > v_0 = 0]
\]

and

\[
h_j = [T = v_j | T > v_{j-1}]
\]

for \( j > 1 \).

Note that for \( t \in [v_j, v_{j+1}), \)

\[
S(t) = P(T > t) = P(T > v_j) = P(T > v_j, T > v_{j-1}) = P(T > v_j | T > v_{j-1}) P(T > v_{j-1}) = P(T > v_j, T > v_{j-1}) P(T > v_{j-2}) P(T > v_{j-2}) \cdots = (1 - h_j)(1 - h_{j-1}) \cdots (1 - h_1) = \prod_{i=1}^{j} (1 - h_i)
\]

The KM estimator for the censored distribution may be obtained by coding the indicator of censoring, \( \delta \) in reverse order. Meaning, the creation of a new complement indicator \( \delta_c \), such that \( \delta_c = 0 \) if the event is observed, and \( \delta_c = 1 \) if censoring occurs. With the above described indicator, \( \delta_c \), the KM survival function can be computed. This estimator is used in applications that adjust estimated measures by the loss of individuals during the follow-up time (Schemper and Henderson, 2000), and can be viewed as the survival probability for censoring. This estimator will be utilized further in Section 6.2.
3.4.1 Log-rank Test

The log-rank test is the most commonly-used statistical test for comparing the survival distributions of two or more groups (such as different treatment groups in a clinical trial). Assume we have 2 groups of individuals, say group \( i = 0 \) and group \( i = 1 \). In group \( i \), there are \( N_i \) i.i.d. underlying survival times with common CDF denoted \( F_i(\cdot) \), for \( i = 0,1 \). The corresponding hazard and survival functions for group \( i \) are denoted \( h_i(\cdot) \) and \( S_i(\cdot) \), respectively.

We assume observations are subject to non-informative right censoring.

We want a non-parametric test of

\[
H_0 : F_0(\cdot) = F_1(\cdot)
\]

or equivalently, of

\[
S_0(\cdot) = S_1(\cdot)(or)h_0(\cdot) = h_1(\cdot)
\]

If we knew \( F_0 \) and \( F_1 \) were in the same parametric family (e.g., for \( T \sim \text{Exp} \), \( S_i(t) = e^{-\lambda_i t} \)), then \( H_0 \) is expressible as the equality of two finite dimensional parameters, say \( H_0 : \lambda_0 = \lambda_1 \). However, we instead want a non-parametric test; a test whose validity does not depend on any parametric assumptions about \( F_j(\cdot) \).

We introduce the log-rank test from the perspective of Mantel (1966) (Cochran-Mantel-Haenszel test (CMH)). This approach adapts methods for analyzing 2x2 contingency tables to accommodate censoring.

**Log-rank Test Construction:**
Denote the distinct times of observed failures as $\tau_1 < \tau_2 < \ldots < \tau_k$, and define

\[ \bar{Y}_i(\tau_j) = \# \text{ persons in group } 1 \text{ who are at risk at } \tau_j \]
\[ \bar{Y}(\tau_j) = Y_0(\tau_j) + Y_i(\tau_j) = \# \text{ at risk at } \tau_j \text{ (both groups)} \]
\[ d_{ij} = \# \text{ in group } I \text{ who fail (uncensored) at } \tau_j \]
\[ d_j = d_{0j} + d_{1j} = \text{ total } \# \text{ failures at } \tau_j \]

The information at time $\tau_j$ can be summarized in the following 2x2 table:

<table>
<thead>
<tr>
<th></th>
<th>Observed to fail $\tau_j$</th>
<th>Survived at risk at $\tau_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 0</td>
<td>$d_{0j}$</td>
<td>$Y_0(\tau_j) - d_{0j}$</td>
</tr>
<tr>
<td>Group 1</td>
<td>$d_{1j}$</td>
<td>$Y_0(\tau_j) - d_{1j}$</td>
</tr>
<tr>
<td></td>
<td>$d_j$</td>
<td>$Y(\tau_j) - d_j$</td>
</tr>
</tbody>
</table>

NOTE: $d_{0j}/Y_0(\tau_j)$ can be viewed as an estimator of $h_0(\tau_j)$.

Suppose $H_0 : F_0(\cdot) = F_1(\cdot)$ holds. Conditional on the 4 marginal totals, a single element (say $d_{1j}$) defines the table. Furthermore, with this conditioning and assuming $H_0$ is true, $d_{1j}$, the number of events in group 1 at time $\tau_j$, has the hyper-geometric distribution, that is:

\[ P[d_{1j} = d] = \frac{\binom{d_j}{d} \binom{Y(\tau_j)d_j}{Y(\tau_j)d_j} \binom{Y_1(\tau_j)}{Y_1(\tau_j)}}{Y(\tau_j)Y_1(\tau_j)} \]

for $d = \max(0, d_j Y_0(\tau_j)), \ldots, \min(d_j, Y_1(\tau_j))$. 


The mean and variance of $d_{1j}$ under $H_0$ are thus

$$
E_j = \frac{Y_1(\tau_j)}{Y(\tau_j)} d_j
$$

$$
V_j = \frac{Y(\tau_j) - Y_1(\tau_j)}{Y(\tau_j)1} \cdot Y_1(\tau_j) \left( \frac{d_j}{Y(\tau_j)} \right) \left( 1 - \frac{d_j}{Y(\tau_j)} \right)
$$

$$
= \frac{Y_0(\tau_j) Y_1(\tau_j) d_j (Y(\tau_j) d_j)}{Y(\tau_j)^2 (Y(\tau_j) 1)}
$$

Define $O_j = d_{1j}$. This is analogous to Fishers exact test. Fishers test would tell us to consider extreme values of $d_{1j}$ as evidence against $H_0$.

Thus, we sum over the distinct failure times and define

$$
O = \sum_{j=1}^{k} O_j = \text{total # failures in group 1}
$$

$$
E = \sum_{j=1}^{k} E_j
$$

$$
V = \sum_{j=1}^{k} V_j
$$

and let

$$
Z = \frac{OE}{\sqrt{V}} = \frac{\sum_{j} (O_j E_j)}{\sqrt{\sum_{j} V_j}}.
$$

Then under $H_0$, it is argued that $Z \sim N(0, 1)$ (or that $Z^2 \sim \chi^2_1$). This approximation can be used to obtain an approximate test for $H_0$ by comparing the observed value of $Z$ (or $Z^2$) to the tail area of the standard normal (chi-square) distribution.
3.5 Cox’s Proportional Hazards Model

The Cox (1972) proportional hazards model is a semi-parametric method that models the association between risk factors and survival times. This model is semi-parametric because while the baseline hazard can take any functional form, the covariates enter the model linearly, and beta-coefficients must be estimated. The main assumption of the model is that constant hazard ratios hold for all the covariates in the model, which is considered the proportional hazards (PH) assumption, where a specific probability distribution for the hazard rate does not need to be specified. There are two main components to the Cox regression model. First, it models in terms of the hazards; and second, it provides hazard ratios, which allow comparing the occurrence of events between groups of individuals, and furthermore, allow examining the impact of risk factors on occurrence of events.

For each of \( n \) subjects we have the value of some covariate vector \( Z \) and the survival outcome \((U, \delta)\) representing non-informatively right-censored values of a survival time \( T \). That is, for subject \( i \), \( Z_i \) denotes the value of the covariate vector \( Z \), and \( T_i \) and \( C_i \) denote the underlying survival time and potential censoring time, respectively, and we observe \((Z_i, U_i, \delta_i)\), where \( U_i = \min\{T_i, C_i\} \) and \( \delta_i = 1[T_i \leq C_i] \), and where \( T_i \perp C_i | Z_i \), that is non-informative censoring is defined by the conditional independence of \( T_i \) and \( C_i \) given \( Z_i \). Note that this is a weaker condition (and more realistic) than \( C_i \perp T_i \).

Suppose \( T \sim NE(\lambda_Z) \), then the working Cox model is

\[
h(t|z) = \lambda_0(t)e^{\beta Z} \tag{15}
\]

where \( \lambda_0(t) \) is the baseline hazard as a function of \( t \), \( \beta = (\beta_1, \ldots, \beta_K)' \) is the vector of regression coefficients, and \( Z = (Z_1, \ldots, Z_K)' \) is the vector of \( K \) covariates. There is wide flexibility in the choice of the regression variables \( Z(t) \) in the model and the specification...
of a suitable model is an important step in data analysis.

Mentioned previously, the main assumption of a Cox model is the proportional hazards assumption. The proportional hazards model can be used when the primary goal of the analysis is to estimate the effect of study variables on survival time. Suppose that we have a Cox model containing a single covariate. Because the hazard function for the proportional hazards regression model is

\[ h(t|z) = \lambda_0(t)e^{\beta Z}, \]

it follows that the link function is the natural log transformation, so that \( \ln[h(t|Z)] = \ln[\lambda_0(t)] + \beta Z \). Thus, we might assume that the \( T_i \) are independent with \( T_i \sim NE(\lambda_0(t)e^{\beta Z}) \).

In general, suppose that

\[ h(t|Z) = h_0(t) \cdot g(Z) \]

where \( h_0(t) \) is a function of time, but not \( Z \), and \( g(Z) \) is a function of \( Z \) but not time. This factorization implies that

\[ \frac{h(t|Z = Z_1)}{h(t|Z = Z_2)} = \frac{g(Z_1)}{g(Z_2)} \]

Specifically, the estimated hazard ratios corresponding to any 2 values of \( Z \) do not depend on time, i.e. proportional hazards. In other words, the proportional hazards assumption characterizes the model as a function of time, not of the covariates per se. For the special case where \( g(Z) = e^{\beta Z} \), this gives \( h(t|Z) = h_0(t)e^{\beta Z} \), which is the Cox proportional hazards model. Here

\[ \frac{h(t|Z = Z_1)}{h(t|Z = Z_2)} = e^{\beta(Z_1-Z_2)} \]

For scalar \( Z \), \( e^\beta \) is the hazard ratio corresponding to a unit change in \( Z \). That is \( h(t|Z = z + 1) = e^\beta h(t|Z = z) \).
3.5.1 Maximum Likelihood Estimation of $\beta$

Given $n$ independent observations from the PH model $h(t|Z) = h_0(t)e^{\beta Z}$, with data $(U_i, \delta_i, Z_i)$, the likelihood function is

$$L(\beta, h_0(\cdot)) = \prod_i \left\{ f(u_i|Z_i)^{\delta_i} S(u_i|Z_i^{1-\delta_i}) \right\}$$

$$= \prod_i \left\{ h(u_i|Z_i)^{\delta_i} S(u_i|Z_i) \right\}$$

$$= \prod_i \left\{ [h_0(u_i)e^{\beta Z_i}]^{\delta_i} \left[ e^{-\int_{0}^{u_i} h_0(t)dt} \right] e^{\beta Z_i} \right\}$$

$$= \text{function}\{\text{data}, \beta, h_0(\cdot)\}$$

The estimation of the parameter $\beta$ is derived via maximum likelihood estimation and is denoted by $L$. The likelihood describes the joint probability of all observations in the data as a function of the unknown parameter $\beta$. The likelihood function for the Cox model is a partial likelihood function because the computation involves individual likelihoods only for subjects where censoring is not observed. However, the information of subjects with censored survival times ($\delta_i = 0$) is not entirely left out. The hazards of those subjects are included in the computation of the likelihood $L_i$ for $\delta_i = 1$, as long as they are at risk at time $t$.

Define $\tau_1 < \tau_2 < \ldots < \tau_k = \text{distinct failure times}$, $d_j = \text{number of failures at time } \tau_j$, $R_j = \text{risk set at } \tau_j = \{\ell|U_\ell \geq \tau_j\}$, that is, the set of observation labels of those $Y(\tau_j)$ subjects at risk at $\tau_j$, and $Z_{(j)} = \text{the value of } Z \text{ for the subject who fails at } \tau_j$. Note that knowledge of the $\tau_j$, $R_j$, and $Z_{(j)}$ allows us to reconstruct the original data for this setting.

The partial likelihood equals

$$L_1(\beta) = \prod \left\{ \frac{h_0(t)e^{\beta Z_{(j)}}}{\sum_{t \in R_j} h_0(t)e^{\beta Z_i}} \right\} = \prod \left\{ \frac{e^{\beta Z_{(j)}}}{\sum_{t \in R_j} e^{\beta Z_i}} \right\}$$
which is a product over the observed failure times of the conditional probability of failure, given the risk set and that a failure occurs at $\tau_j$. In order to get the estimates of $\beta$, the partial likelihood $L_1$ is maximized by using Newton-Raphson techniques. We do not need to know or estimate $h_0(t)$ to estimate and conduct inference for $\beta$. For ease of computation, the log-likelihood is computed and is equal to

$$\ell(L_1(\beta)) = \sum (Z\beta - \ln(\sum e^{Z\beta}))$$

We then maximize $\ell(L_1(\beta))$ by solving the score equation:

$$U(\beta) = \frac{\partial \ell}{\partial \beta} = 0 = \sum \left( Z_i - \frac{\sum Ze^{Z_i\beta}}{\sum e^{Z_i\beta}} \right) = \sum (Z_i - \bar{Z}_i)$$

where $\bar{Z}_i$ is the weighted average of the $Z$ covariates in the risk set $R_j$. The solution for $U(\beta) = 0$ is the maximum likelihood estimator $\hat{\beta}$.

The variance of $\hat{\beta}$ can be obtained by using the second derivative of $\ell$ from the information matrix $\hat{I}(\beta)$ so that

$$\text{var}(\hat{\beta}) = \left[ -\frac{\partial^2 \ell}{\partial \beta^2} \right]^{-1}$$
where
\[
\frac{\partial^2 \ell}{\partial \beta^2} = \sum \left( -\frac{(\sum Z_j - \bar{Z}_j)^2 e^{Z_i \beta}}{\sum e^{Z_i \beta}} \right)
\]

3.5.2 Inference

We can use formal statistical tests to make inference about $\beta$. Such tests include the Wald test and Score test, where both test the null hypothesis of no association between risk factors and survival.

Wald test of $H_0 : \beta = 0$:

\[
\frac{\hat{\beta} - 0}{\sqrt{I^{-1}(\hat{\beta})}} \approx N(0, 1) \text{ under } H_0
\]

based on $\hat{\beta} \approx N(\beta, I^{-1}(\hat{\beta}))$

Score test of $H_0 : \beta = 0$:

based on assuming $U(0) \approx N(0, \hat{I}(0))$ under $H_0$. When there are no tied failure times (all $d_j = 1$) then

\[
U(0) = \sum_{j=1}^{k} \{Z_{(j)} - \bar{Z}_j(0)\}
\]

where

\[
\bar{Z}_j(0) = \frac{\sum_{\ell \in R_j} Z_{\ell}}{\sum_{\ell \in R_j} 1}
\]

which equals the average value of the $Z$ in $R_j$. 

78
3.6 Prediction

3.6.1 Previous Estimates

A model is considered a good fit for prediction if the computed prediction errors are small. The prediction errors are computed as the squared (or absolute) difference between a future response and its prediction from the fitted model. In an ideal experiment, the future response would be taken from new data independent from the data used for fitting the model. However, there are not always new data available. A widely popular technique to handle the availability of the new data problem include bootstrapping and cross-validation (Gerds and Schumacher, 2006). Both the cross-validation and the bootstrap estimator estimate prediction errors in different ways by using the available data to repeatedly simulate two subsets: a training sample to fit the model and a validation set to validate the model.

Gerds and Schumacher (2007) extended the applicability of these estimators of prediction errors to survival data. Based on the definition of the Brier score (Brier, 1950), they computed prediction errors as the squared difference between the true survival status (taken from the validation set) and the estimated survival probabilities under the model (fitted with the training sample). The Brier score is one of the main approaches to estimate prediction errors in survival models and measures the accuracy of probabilistic predictions for models with binary or categorical outcomes. This method was later adopted for the study of estimation errors from survival models (Graf et al., 1999; Gerds and Schumacher, 2006; Gerds and Schumacher, 2007). The Brier Score for a model that predicts \( P_i \) for patient \( i \) out of \( N \) patients is:

\[
BS = \frac{1}{N} \sum_{i=1}^{N} (Y_i - P_i)^2
\]

For a given model it estimates the expected squared difference between patient status and predicted probability, and is formulated as the usual concept of residuals, that is, the squared deviation between observed and estimated outcomes under the model. The lower
the Brier score of a model, the better the predictive performance. For the purposes of this thesis the outcome is the survival endpoint, $S(t)$. To account for the loss of information due to censoring, Graf et al. (1999) formulated this measure as a weighted average of squared deviations, where the weight for an individual contributes to the Brier score according to their censoring information. Gerds and Schumacher (2006) extended these weights to allow for non-random censoring, $W_c(t, \hat{G}, X_i)$.

$$SSR_{br} = \frac{1}{n} \sum_{i=1}^{n} (S_i(t) - \hat{S}(t))^2 W_c(t, \hat{G}, X_i)$$

where

$$S_i(t) = \begin{cases} 
1 & \text{if person } i \text{ did not have an event up to time } t \\
0 & \text{otherwise}
\end{cases}$$

and

$$W_c(t, \hat{G}, X_i) = \frac{I(t_i \leq t) \delta_i}{\hat{G}(t_i - |X_i|)} + \frac{I(t_i > t)}{\hat{G}(t |X_i|)}$$

$W_c(t, \hat{G}, X_i)$ is a the weight for person $i$ determined only by the empirical estimate of the survival function for censoring ($\hat{G}$). In other words, where censoring can be assumed not to depend on the individual’s survival chances, the weights incorporate the Kaplan-Meier estimator $G$ of the censoring or potential follow-up distribution, which is obtained from $(Y, 1 - \delta)$ by exchanging the roles of censored and uncensored observations. $W_c(t, \hat{G}, X_i)$ does not involve the use of $\hat{\beta}$ estimates, or any other estimate based on the Cox model. Individuals whose survival status at $t_i$ is known receive weights $> 1$, so that they represent the contributions of individuals whose Brier score is unobservable in addition to their own contribution. A drawback to this method is that the Brier score assigns weights equal to 0 to individuals censored before time $t$, and therefore information of these censored individuals are left out from the estimate (these individuals contribute indirectly to the empirical Brier
score, because they are used in the calculation of G).

3.6.2 Proposed Method

Define

\[ Y_i(t) = I(t \leq U_i); \quad \bar{Y}(t) = \sum_{i=1}^{n} Y_i(t) \]

(16)

\[ N_i(t) = I(t \geq U_i, \delta_i = 1) \text{ or } \delta_i I(t \geq U_i); \quad \bar{N}(t) = \sum_{i=1}^{n} N_i(t) \]

(17)

\[ A_i(t) = \int_{0}^{t} Y_i(u) \lambda(u|Z_i) du \]

(18)

\[ S_i(t) = 1 - N_i(t) \]

(19)

\[ M_i(t) = N_i(t) - A_i(t) \]

(20)

\[ = N_i(t) - \int_{0}^{t} Y_i(u) \lambda(u|Z_i) du \]

(21)

\[ = N_i(t) - \int_{0}^{t} Y_i(u) d\Lambda(u|Z_i) \text{ since } \lambda(u|Z_i) = \frac{d\Lambda(u|Z_i)}{du} \]

(22)

We propose a measure of model fit similar in flavor to the estimate developed by Gerds and Schumacher (2006), however distinct differences exist. To show the relationship between the two approaches, note that \( (S_i(t) - \hat{S}_i(t))^2 \) can be written as

\[ ((1 - N_i(t)) - e^{-\Lambda_i(t)})^2. \]

Using a first-order Taylor series approximation this is

\[ (1 - N_i(t)) - (1 - \Lambda_i(t))^2 = (N_i(t) - \Lambda_i(t))^2 \]
When assuming the Cox model is properly specified, it follows that

\[
E[N_i(t)] = E[A_i(t)]
\]

(23)

\[
= E \left[ \int_0^t Y_i(v) d\Lambda_0(v) e^{\beta Z_i} \mid Z_i \right]
\]

(24)

\[
= \int_0^t E \left[ (Y_i(v) d\Lambda_0(v) e^{\beta Z_i}) \mid Z_i \right]
\]

(25)

\[
= \int_0^t P(C_i \geq v, T_i \geq v \mid Z_i) d\Lambda_0(v) e^{\beta Z_i}
\]

(26)

\[
= \mu(t)
\]

(27)

For the proposed method, we use a Brier type score having the common property of expected squared difference between the observed and expected outcome and we evaluate differences through time for each individual.

Consider the following statistic, \( D \), which represents the prediction error estimate of a survival model:

\[
D(\Lambda_0(\cdot), \beta, t) = D(t) = \frac{1}{n} \sum_{i=1}^n \int_0^t Y_i(u) [N_i(u) - \mu(t)]^2 du
\]

(28)

\[
= \frac{1}{n} \sum_{i=1}^n A_i^*(t)
\]

(29)

We summarize by using \( D(\infty) \equiv D = \sum A_i^*(\infty) \equiv \sum A_i^* \)

To estimate \( D \), consider

\[
\hat{\mu}(t) = \int_0^t \hat{P}(C_i \geq v, T_i \geq v \mid Z_i) d\hat{\Lambda}_0(v) e^{\hat{\beta} Z_i} dv
\]
where \( \hat{\Lambda}_0(t) \) is Breslow’s (1974) estimator and equals:

\[
\hat{\Lambda}_0(t) = \sum_{\tau_j \leq t} \frac{d_j}{\sum_{i=1}^n Y_i(t) e^{\beta Z_i}} = \int_0^t \frac{d\bar{N}(u)}{\sum_{i=1}^n Y_i(u) e^{\beta Z_i}}
\]

and \( \hat{\beta} \) is the maximum partial-likelihood estimator (MPLE).

Assume \( T \perp C|Z \) (non-informative censoring), then

\[
\hat{\mu}(t) = \int_0^t \hat{P}(C_i \geq v|Z_i) \hat{P}(T_i \geq v|Z_i) d\hat{\Lambda}_0(v) e^{\hat{\beta} Z_i}
\]

where \( \hat{P}(T_i \geq v|Z_i) = e^{-\hat{\Lambda}_0(u) e^{\hat{\beta} Z_i}} \), and, if \( C \) is censored by \( T \) (i.e. \( T \) is time to death), \( \hat{P}(C_i \geq v|Z_i) = e^{-\Lambda^C_0(u) e^{\gamma Z_i}} \), where \( \Lambda^C_0(\cdot) \) and \( \gamma \) are the parameters of a proportional hazards model for \( C \), and \( \hat{\Lambda}^C_0(\cdot) \) and \( \hat{\gamma} \) are Breslow’s estimate and the MPLE, respectively. On the other hand if \( C \) is observed, \( \forall \ i = 1, \ldots, n \) we do not necessarily need a censored data regression model such as the Cox model. For example, a linear regression model can be used, with error distribution specified; see below.

If \( C \) is censored by \( T \), then for subject \( i \):

\[
\hat{A}^*_i = \int_0^\infty Y_i(t) \left( N_i(t) - \int_0^t \hat{P}(C_i \geq v|Z_i) \hat{P}(T_i \geq v|Z_i) d\hat{\Lambda}_0(v) e^{\hat{\beta} Z_i} \right)^2 \tag{30}
\]

\[
= \int_0^\infty Y_i(t) \left( N_i(t) - \int_0^t e^{-\Lambda^C_0(v) e^{\gamma Z_i}} e^{-\hat{\Lambda}_0(v) e^{\hat{\beta} Z_i}} d\hat{\Lambda}_0(v) e^{\hat{\beta} Z_i} \right)^2 \tag{31}
\]

so that \( \text{var}(A^*_i) = \frac{1}{n} \sum_{i=1}^n A^*_i - (\frac{1}{n} \sum_{i=1}^n A^*_i)^2 \).
needs to be selected and estimated. Here $\hat{\epsilon} = C_i - \hat{\gamma}Z_i$, where $\hat{\gamma}$ is obtained by finding the root of $\frac{1}{n} \sum_{i=1}^{n} Z_i [C_i - \hat{\gamma}Z_i]$. One way to do this is to use the EasyFit Professional Microsoft Excel add-on which allows a user to fit probability distributions to sample data. For example, in our data results, we chose the Generalized Extreme Value distribution based on goodness of fit (GOF) test statistics, such as Kolmogorov Smirnov.

If $C_i \sim$ Generalized Extreme Value, then $F_i(x) = e^{-(1+k(\frac{x-\mu_i}{\sigma}))^{-1/k}}$, and $\epsilon \sim F(k, \mu, \sigma)$.

Then,

$$\hat{P}(C_i \geq v|Z_i) = 1 - e^{-(1+k(\frac{v-\mu_i}{\sigma}))^{-1/k}}$$

gets substituted in (30), where $\mu_i = \gamma Z_i + \mu$. See Section 9 for the results of fitting this model to our dataset.

### 3.7 Model Checking

Common questions after fitting a cox model include: Do we have the correct functional form for continuous covariates? Are there any significant interactions? Is the proportional hazards assumptions met? If not, what are the options?

#### 3.7.1 Martingale Residuals

Often we assume continuous covariates have a linear form. However, this assumption should always be checked, often by graphical methods based on the analysis of residuals for Cox regression models (Cox and Snell, 1968). For this model, partial likelihood methods are commonly used to estimate the vector of regression parameters and the baseline hazard. An approach for forming residuals involves using the fact that the counting process martingales (Grambsch and Therneau, 1994) we use in analyzing Cox’s model have zero
mean, i.e.

\[ E\{M_i(t)\} = E\{N_i(t) - A_i(t)\} = 0 \]

Note that checking functional form of a variable requires it is included in the model in some form.

Recall that if (15) were the true model, then the true compensator for the counting process \(N_i(t) = 1(U_i \leq t, \delta_i = 1)\), i.e. that variable that satisfies \(E\{N_i(t)\} = E\{A_i(t)\}\) would be

\[ A_i(t) = \int_0^t Y_i(u)\lambda_0(u)e^{\beta Z_i}du = e^{\beta Z_i} \int_0^t Y_i(u)\lambda_0(u)du = e^{\beta Z_i} \int_0^t Y_i(u)d\Lambda_0(u) \]

This suggests that the process \(\hat{M}_i(\cdot)\), a subject-specific martingale, defined at time \(t\) by

\[ \hat{M}_i(t) = N_i(t) = e^{\hat{\beta} Z_i} \int_0^t Y_i(u)d\hat{\Lambda}_0(u) \quad i = 1, 2, \ldots, n, \]

where \(\hat{\beta}\) is the maximum partial likelihood estimator and \(\hat{\Lambda}_0(\cdot)\) is the Breslow estimator of the cumulative baseline hazard, should have (approximately) uncorrelated increments and a zero mean function if the Cox model (15) were the correct model. These "martingale residuals" have been studied by several authors (see, for example, Schoenfeld, 1982), and satisfy \(\sum_{i=1}^n \hat{M}_i(t) = 0\), regardless of whether the fitted model in (15) is correct. The residual can be interpreted, at each \(t\), as the difference over \([0, t]\) in the observed number of events minus the expected number given the model, or as excess events.

For a Cox model with no time-dependents covariates and where \(t = U_i\), the residual reduces to the simple form

\[ M_i^* = N_i(U_i) - e^{\hat{\beta} Z_i}\int_0^{U_i} Y_i(u)d\hat{\Lambda}_0(u) = \delta_i - \hat{\Lambda}_0(U_i)e^{\hat{\beta} Z_i} = \delta_i - U_i^* \]
3.7.2 Cumulative Martingale Residuals

Suppose we want to focus on the fit of a continuous covariate, say $Z_1$, in our fitted model. Another, more useful way to visualize the martingale residuals is to plot their cumulative sums against this covariate’s values. That is, consider

$$Q(z) = \sum_{i=1}^{n} 1(Z_{1i} \leq z)M_i^*$$

Note that the above sum arises when the data are sorted in ascending order corresponding to $(Z_{11}, \ldots, Z_{1n})$, and then the correspondingly re-ordered martingale residuals $M_i^*$ are cumulatively summed.

Then we plot $Q(z)$ versus $z$. If the fitted model were correct, then each term in the above sum, and hence the sum itself, has expectation that is approximately zero. That is, each

$$E(M_i^*) = 0$$

if the working model is correctly compensating for $N_i(t), i = 1, \ldots, n$. Since the terms $(M_1^*, \ldots, M_n^*)$ are approximately uncorrelated, the resulting process $Q(z)$, viewed as a function of $z$, would have (approximately) a zero mean function and uncorrelated increments, and thus an increasing variance function, if the fitted model were correct. Thus, when the working Cox PH model is correct, one should not expect to see a trend in the mean function in a plot of $Q(z)$ versus $z$, rather there should be a flat line at zero.

Given the correlation in such graphs between successive values of $z$, it is not simple to visually judge whether evidence of departure from the model versus random fluctuations is seen. One way to assess this is to augment plots by adding some realizations from the process $Q(\cdot)$ that we would expect to see if the model were correct, denoted $\tilde{Q}_1(\cdot), \tilde{Q}_2(\cdot), \ldots$, where, for the random sample of $n$ i.i.d. $N(0, 1)$ variables $(G_1^{(m)}, \ldots, G_n^{(m)})$,

$$\tilde{Q}_m(z) = \sum_{i=1}^{n} a(Z_{1i} \leq z)N_i(U_i)G_i^{(m)} = \sum_{i=1}^{n} 1(Z_{1i} \leq z)\delta_i G_i^{(m)}$$
One could then compare the observed realization of $Q(\cdot)$ to the realizations $\tilde{Q}_1(\cdot), \tilde{Q}_2(\cdot), \ldots$, that are actually obtained under the hypothesis that the Cox model is correctly specified, and then judge if the observed realization of $Q(\cdot)$ seems like an “outlier” compared to the $\tilde{Q}_1(\cdot), \tilde{Q}_2(\cdot), \ldots$. This approach was suggested by Lin, Wei & Ying (1993), by introducing a type of Multiplier Central Limit Theorem (CLT) technique.

### 3.7.3 Checking for Proportional Hazards

A key assumption for many types of survival models, including the Cox model, is that the hazard ratios are proportional to one another and that proportionality is maintained over time. Checking for PH can be done by graphical methods such as plotting $\ln(-\ln(S(t)))$ vs. $t$ or $\ln(t)$ and looking for parallelism; plotting the observed and predicted $S(t)$ and looking for a close fit; and using PH graphs in statistical packages such as SAS®. PH assumption is supported by parallel lines and refuted by lines that cross or nearly cross.

Another way to check for PH would be to add time-dependent covariates (interactions) such as time*covariate to the model to fit non-PH. If the coefficient for the time-dependent variable is significantly different from zero, non-PH is present. If significant non-PH is found, this model can be kept to fit and interpret the non-PH. Lastly, similar to checking the functional form of a covariate, we can use tests based on re-sampling to check if a covariate has a proportional effect on the hazard of $T$. Note that the Cox score equation can be written as:

$$
D_n(t; \beta) = \sum_{i=1}^{n} \left( Z_i - \frac{\sum_{j=1}^{n} Y_j(U_i)e^{\beta Z_j}}{\sum_{i=1}^{n} Y_j(U_i)e^{\beta Z_j}} \right) 1(U_i \leq t)
$$

and when the Cox model is properly specified, that is, the true conditional (on $Z_i$) popula-
tion hazard is \( h(t|Z_i) = \lambda_0(t)e^{\beta Z_i} \), then

\[
D_n(t; \beta) = \sum_{i=1}^{n} \left( Z_i - \frac{\sum_{i=1}^{n} Y_j(U_i)e^{\beta Z_j}}{\sum_{i=1}^{n} Y_j(U_i)e^{\beta Z_j}} \right) dM_i(u)
\]

(34)

\[
= \sum_{i=1}^{n} Z_i M_i(t)
\]

(35)

When we plug in the MPLE \( \hat{\beta} \) and the Breslow estimator \( \hat{\lambda}_0(\cdot) \) the cumulative baseline hazard, we get

\[
Q_1(t) = D_n(t; \hat{\beta}) = \sum_{i=1}^{n} Z_i \hat{M}_i(t)
\]

Note that \( Q_1(\infty) = 0 \), since \( \hat{\beta} \) is the value that sets \( D_n(\infty; \hat{\beta}) = 0 \) (the derivative of the log partial likelihood = 0). We can thus plot \( Q_1(t) \) versus \( t \) along with multiple realizations of the process under the null hypothesis that the Cox PH model is correctly specified, namely,

\[
\hat{Q}_1(t) = \sum_{i=1}^{n} Z_i N_i(t) G_i
\]

one for each independent random sample from \( N(0, 1) \), given by \( (G_1, \ldots, G_n) \).

These techniques can be utilized in the SAS® PHREG procedure. While this option is a useful tool, this technique should be used in conjunction with other checks for functional form and PH. Furthermore, the cumulative martingale residual plots are not very sensitive for fine-tuning functional form and can show grossly incorrect forms.
3.7.4 Calculating p-values

A nice feature of the above mentioned simulation approach is that it easily lends itself to formal hypothesis tests of goodness of fit. For example, one can compute

\[ s_m = \sup_z |\tilde{Q}_m(z)| \quad m = 1, \ldots, B, \]

and then compare the observed value \( \sup_z |Q(z)| \) to the distribution of the \( (s_1, \ldots, s_B) \), so that a (two-sided) p-value is

\[ \frac{1}{B} \sum_{m=1}^{B} 1\{s_m \geq \sup_z |Q(z)|\} \]

3.7.5 Schoenfeld Residuals

Schoenfeld (1982) proposed the first set of residuals for use with Cox regression packages. Instead of a single residual for each individual, there is a separate residual for each individual for each covariate based on the individual contributions to the derivative of the log partial likelihood (Hosmer and Lemeshow, 1999). It is important to note that Schoenfeld residuals are not defined for censored individuals.

Assume \( p \) covariates and \( n \) independent observations of time, covariates and censoring, which are represented as \((t_i, z_i, c_i)\), where \( i = 1, 2, \ldots, n \), and \( c_i = 1 \) for uncensored observations and zero otherwise. To derive the Schoenfeld residuals, one takes the derivative for the \( k^{th} \) covariate,

\[
\frac{\partial L_p(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} c_i \left\{ z_{ik} - \frac{\sum_{i \in R(t_i)} z_{jk} e^{\beta z_j}}{\sum_{i \in R(t_i)} z_{jk} e^{\beta z_j}} \right\}
= \sum_{i=1}^{n} c_i \left\{ z_{ik} - \bar{z}_{w_i,k} \right\}
\]
where

\[ \bar{z}_{wik} = \frac{\sum_{i \in R(t_i)} z_{jk} e^{\beta z_j'}}{\sum_{i \in R(t_i)} z_{jk} e^{\beta z_j'}} \]

Hosmer and Lemeshow (1999) show, the estimator of the Schoenfeld residual for the \( i \)th subject on the \( k \)th covariate are then obtained by substituting the partial likelihood estimator of the coefficient, \( \hat{\beta} \):

\[ \hat{r}_{si,k} = c_i (z_{ik} - \bar{z}_{wik}), \]

where

\[ \hat{z}_{wik} = \frac{\sum_{i \in R(t_i)} z_{jk} e^{\beta z_j'}}{\sum_{i \in R(t_i)} z_{jk} e^{\beta z_j'}} \]

Schoenfeld residuals can be thought of as the observed minus the expected values of the covariates at each failure time. If the residual exhibits a random, (i.e. unsystematic) pattern at each failure time, then this gives evidence the covariate effect is not changing with respect to time, which is precisely the PH assumption. If it is systematic, it suggests that as time passes, the covariate effect is changing. This is because if the PH assumption holds, then we would expect the difference between covariate values at failure times versus a weighted average of the covariate values to display no temporal trends. In residual plots, we might expect the slope of the (re-scaled) Schoenfeld residuals with respect to time to be zero.

The SAS® macro, SCHOEN, produces graphical check for PH. Consider the possibility that the \( \beta \) coefficient for a given covariate, \( \beta_k \), changes over time, thus giving a non-constant hazard ratio. Macro SCHOEN uses a scaled Schoenfeld residual, multiplying the vector of Schoenfeld residuals by the inverse of their covariance matrix. This scaled residual, \( r_{ik}^* \), added to \( \beta_k \), is an estimate of the time-dependent \( \beta \) coefficient: \( r_{ik} + \beta_k \approx \beta_k(t_i) \), where \( r_{ik} + \beta_k \) is plotted against time, or a function of time. PH is indicated by a flat pattern around \( Y = 0 \). Non-PH is indicated by any deviation from a flat line at \( Y = 0 \). The Schoenfeld residuals can be plotted against any function of time. The pattern shown over time indicates the form of non-PH.
3.8 Model Comparisons

Traditional approaches for model comparisons include Akaike’s Information Criteria (AIC), Schwarz Bayesian Criteria (SC), and likelihood ratio tests (-2 Log L) which are all likelihood based measures that estimate a measure of the difference between a given model and the true underlying model (Agresti, 2002). The model with the smallest AIC, SC, or LRT among all competing models is deemed the best model. The likelihood ratio test statistic is defined as:

\[-2 \log L = -2 \sum w_j f_j \log (\hat{p}_j),\]

where \(w_j\) and \(f_j\) are the weight and frequency values of \(j^{th}\) observation and \(\hat{p}_j\) is the estimated event probability.

AIC:

\[-2 \log L = -2 \log L + 2(k + s),\]

where \(k\) is the total number of response levels minus one, and \(s\) is the number of explanatory effects.

SC:

\[-2 \log L = -2 \log L + (k + s) \log \left( \sum f_j \right).\]

The -2 Log Likelihood statistic has a chi-square distribution under the null hypothesis (that all the explanatory effects in the model are zero). The AIC and SC statistics give two different ways of adjusting the 2 Log Likelihood statistic for the number of terms in the model and the number of observations used. These statistics are used when comparing different models for the same data. Lower values of the statistic indicate a more desirable model.

Our goal is to conduct model comparisons that do not require proper specification of like-
likelihoods for valid inference. The methods use a 2-step Delta method to derive consistent variance estimates of prediction error estimates. Multiple hypothesis testing procedures are employed that maintain a chosen error rate of some pre-specified level for arbitrary data generating estimates, or so-called strong control.

Multiple testing refers to any instance that involves the simultaneous testing of several hypotheses. Multiple testing procedures adjust p-values derived from multiple statistical tests to correct for the occurrence of false positives. Suppose there are several model comparisons of interest denoted \( H_1 : \Phi_0^{(1)} = \Phi_1^{(1)}, \ldots, H_Q : \Phi_0^{(Q)} = \Phi_1^{(Q)} \). To obtain a test statistic for each hypothesis, define \( W_q = \frac{\hat{\Phi}_1^{(q)} - \hat{\Phi}_0^{(q)}}{\sqrt{\text{var}(\hat{\Phi}_1^{(q)} - \hat{\Phi}_0^{(q)})}}, q = 1, \ldots, Q \) where \( \text{var}(\hat{\Phi}_1^{(q)} - \hat{\Phi}_0^{(q)}) \) is the sample variance of \( \{A_{i,0}^{(q)} - A_{i,1}^{(q)}\}, i = 1, \ldots, n \) and \( A_{i,j}^{(q)} \) is defined for \( j = (0,1) \), \( q = 1, \ldots, Q \) as taking the form shown in (17). The marginal p-values \( \hat{\pi}_q = P(W_q < Z) \) can be obtained where \( Z \sim N(0,1) \). With \( \hat{\pi}_1, \ldots, \hat{\pi}_Q \) we can simultaneously test \( H_1, \ldots, H_Q \) using the following multiple testing procedures (MTP).

Accounting for the multiplicity of individual hypothesis tests can be attained by controlling an appropriate error rate. A classical approach to examining the multiplicity problem is to restrict to procedures that control the familywise error rate (FWER), defined to be the probability of one or more false rejections.

Romano and Shaikh (2006) proposed step-up procedures that provide strong control, that is, for arbitrary \( P(Y, X) \), of either the number or proportion of true null hypotheses among the rejected hypotheses at level \( \alpha \). If \( PFP \) represents the proportion of false positivities of among the rejected hypotheses (which equals 0 if there are no rejections), then control of the \( PFP \) satisfies \( P(PFP > \phi) \leq \alpha \), where \( \phi \in [0,1] \) is user defined. Note that choosing \( \phi = 0 \) corresponds to the usual FWER. Although use of the FWER leads to a simple interpretation of the covariate screening result, it can be considerably less powerful.
compared to choosing $\phi > 0$. For simplicity, only procedures for controlling the FWER are presented below. Note that it is not necessary to estimate the covariance between p-values to execute any of these testing procedures (whether or not $\phi$ is taken to equal 0). To control the FWER, Romano and Shaikh (2006) propose the following step-up algorithm. Given p-value $\hat{\pi}(q)$ for testing $H(q)$, order the observed p-values in descending order to obtain $\hat{\pi}(Q) \geq \cdots \geq \hat{\pi}(1)$ with corresponding $H(Q), \ldots, H(1)$. If $\hat{\pi}(Q) \leq \frac{\alpha}{a(Q)}$, then reject all $Q$ hypotheses. Otherwise, if $\hat{\pi}(Q) > \frac{\alpha}{a(Q)}$, do not reject $H_Q$ and next consider if $\hat{\pi}(Q-1) > \frac{\alpha}{a(Q-1)}$.

In general, reject hypotheses $H_{(q^*)}, \ldots, H(1)$, where $q^*$ is the smallest index satisfying

$$\hat{\pi}(Q) > \frac{\alpha}{a(Q)}, \ldots, \hat{\pi}(q^*+1) > \frac{\alpha}{a(q^*+1)}.$$ 

If, for all $q$, $\hat{\pi}(q) > \frac{\alpha}{a(q)}$, then reject no hypotheses. That is, a step-up procedure begins with the least significant p-value and continues accepting hypotheses as long as their corresponding p-values are large. The formula for the constant $a(Q)$ is provided in Romano and Shaikh (2006) and equals 2.13 for $Q \geq 25$.

Holm (1979) introduced a step-down procedure that provides strong control of the FWER. The algorithm is as follows. A step-down procedure begins with the most significant test statistic and determines which null hypotheses to reject. If $\hat{\pi}(1) > \frac{\alpha}{Q}$, then reject no hypotheses. Otherwise, reject hypotheses $H(1), \ldots, H_{(q^*)}$, where $q^*$ is the largest index such that

$$\hat{\pi}(1) \leq \frac{\alpha}{Q}, \ldots, \hat{\pi}(q^*) \leq \frac{\alpha}{Q - q^* + 1}.$$ 

That is, a step-down procedure begins with the most significant p-value and continues rejecting hypotheses as long as their corresponding p-values are small. Holm’s method is based on the Bonferroni method and is valid regardless of the joint distribution of the test statistics (strong control).
Lehmann and Romano (2005) proposed step-down procedures that provide strong control over the number or proportion of false positives. Romano and Shaikh (2004) proposed step-down methods that provide strong control of the proportion of false positives that can be considerably more powerful than those of Lehmann and Romano (2005).

3.9 Application

The data used are from AIDS Clinical Trials Group (ACTG) 398, a cohort of 481. ACTG 398 was a randomized, double-blind, placebo-controlled study of saquinavir, indinavir, or nelfinavir added as a second protease inhibitor to the 4-drug class regimen of amprenavir, abacavir, efavirenz, and adefovir dipivoxil in patients with virologic failure. Patients were randomized between October 31, 1998 and April 14, 1999, with a study completion date of April 2000 (Hammer, 2002). Virologic failure was defined as a viral load above 1000 copies/mL while receiving saquinavir, nelfinavir, indinavir, or ritonavir (Hammer, 2002). The main objective of ACTG 398 was to assess whether the addition of a second protease inhibitor to a new 4-drug class regimen including amprenavir would improve virologic response in patients failing protease inhibitor containing regimen. The study incorporated patients recruited from 31 AIDS Clinical Trial Units that were 13 years of age or older, had laboratory documentation of an HIV-1 infection, had prior exposure to a maximum of 3 protease inhibitors from among saquinavir, ritonavir, indinavir, and/or nelfinavir for a cumulative period of protease inhibitor therapy of at least 16 weeks, receiving the failing protease inhibitor containing regimen at time of screening, and having certain laboratory parameter levels (Hammer, 2002). Patients were also required to be treatment-naive to amprenavir, abacavir, and adefovir dipivoxil. Participants received follow-up visits (including clinical assessment, routine laboratory monitoring, and viral load measurement) at weeks 2, 4, 8, 16, and every 8 weeks until the last participant completed 48 weeks on study (Ham-
The CD4 cell counts of participants were measured twice at baseline and at weeks 4, 8, 16 and every 8 weeks until the study end (Hammer, 2002). The study found that antiretroviral-experienced patients with advanced immunodeficiency could achieve viral suppression with regimens containing four or five new drugs, with 148 (31%) of the 481 participants achieving a viral load below 200 copies/mL at week 24 (Hammer, 2002).

In this paper, the response variable Y is the HIV-1 time to off-track measure, an important measure of the patients’ response to treatment from baseline administration.

As previously described in Paper 1: *Statistical Methods for Identifying and Modeling Patterns of HIV-1 Genotype Evolution*, conditional independence tests were utilized to predict week 24 RNA measures based on significant codons found using MTP for 1-way, 2-way, and 3-way changes in HIV-1 RNA genotype. Codons found to have a statistically significant change between baseline and post-baseline values included \( PR_5 \), \( RT_{44} \), \( RT_{74} \), \( RT_{100} \), \( RT_{101} \), \( RT_{103} \), \( RT_{108} \), \( RT_{181} \), and \( RT_{190} \). In an attempt to improve the accuracy of predicting week 24 RNA measures, additional week 4 genotype variables were included in a polychotomous logistic regression model (Paper 2: *New Statistical Methods to Evaluate and Compare Generalized Logit Models*). With these results various working models \( M_0 - M_3 \) were posited to investigate if the addition of week 4 RNA genotype information improves model fit. From this analysis, it was concluded that models incorporating week 4 clinical and genotypic information had a significant effect on the prediction of week 24 RNA measures, potentially allowing for early identification of virologic failure. Models \( M_0 - M_3 \) used for this analysis are described in Table 12.

During this current analysis, previously posited models \( M_0 - M_3 \) were utilized with a focus on the right-censored survival endpoint of time to virologic failure since not all participants were observed to fail at the same time. Predictors for these models included continuous
baseline CD4 and RNA, continuous week 4 RNA, and categorical baseline codons RT100, RT103, and week 4 RT103.

In addition to fitting survival models for M0-M3, the censored distribution was also fit as described in Section (3.6). For this analysis the censored distribution was fit to the generalized extreme value distribution with parameters equal to: $k = -0.47886$, $\sigma = 66.781$, and $\mu = -15.936$. A PDF plot of the residuals is shown in Figure 17 and indicates an adequate fit for the censored errors.

Next, we plotted the cumulative sums of martingale residuals against our covariates to check functional form. Additionally, we plotted the observed score process against time to check for proportional hazards for all covariates. All models met functional form requirements and proportionality except model $M_3$ (containing both genotype and clinical information) which showed that while baseline CD4 and baseline RNA had proper covariate form, week 4 RNA did not. Overall, covariates have proportional hazards with the exception of week 4 RNA. The observed residual was too large relative to the randomly generated sample processes (p-value = < 0.0001) indicating the proportional hazards assumption did not hold. This is a Brownian process, or Brownian bridge, meaning that the values always start and end at zero. Random paths are generated under PH, where the path from the actual data is compared to the randomly-generated paths under PH. If the actual path is within the cloud of random paths, this indicates PH. In an attempt to address the misspecification of functional form, we transformed week 4 RNA into a binary variable based on log RNA level ($1 = RNA < 4$, 0 otherwise). Adding this variable to the model in place of continuous week 4 RNA improved model fit and resulted in proportional hazards across all covariates. Plots for assessing functional form and proportional hazards for model $M_3$ are displayed in Figures 18–25. Figures 18 and 19 show the observed curves for baseline CD4 and baseline RNA to be within the distribution of the simulated cumulative
martingale residuals curves, indicating acceptable fit with a linear form (p-values 0.1982 and 0.2496, respectively). Resulting supremum test for checking functional form and proportional hazards are displayed in Tables 13 and 14.

After fitting appropriate Cox models, prediction error was assessed for all models. As week 4 information was added into the baseline model, the average squared prediction error decreased. Model $M_3$ had the smallest average squared prediction error of 0.75 (Table 15). Although model $M_3$ had the lowest prediction error, can we conclude that model $M_3$ has statistically lower prediction error than $M_0$, $M_1$, and $M_2$?

After comparing selected models via MTP, rejected hypotheses due to highly significant p-values are displayed in Table 16. The Holm step-down method showed that hypotheses 2, 3, 4, 5 and 6 were rejected (indicated with a single *). No additional hypotheses were rejected using the step-up method or the Lehmann and Romano (2005) step-down method. Model $M_3$ had the smallest prediction error and was statistically better than models with less week 4 information. In other words, we concluded that compared to $M_0$ (baseline model), $M_1$ (baseline plus week 4 RT103), and $M_2$ (baseline plus week 4 RNA), model $M_3$ is statistically better in the sense that baseline information alone or the addition of one piece of week 4 information is not the best model. The addition of week 4 genotype alone ($M_1$) or week 4 RNA alone ($M_2$) were not statistically better than the baseline model, but the model with the combination of week 4 genotype, week 4 RNA, and baseline information ($M_3$) was statistically better than all models.

Hazard ratios for model $M_3$ were calculated and are presented in Table 17. The hazard ratio of virologic failure for week 4 RT103 among mutant versus wildtype was 2.4 (CI 1.61, 3.69) (Table 17). Furthermore, the hazard ratio of virologic failure for those identified with week 4 RNA $>10,000$ copies/mL versus $<10,000$ copies/mL was 6.35 (CI 4.56, 8.84).
showing that as week 4 RNA levels increase, the hazard for time to virologic failure increases. The median predicted survival time was about 30 days for week 4 RT103 mutant. Week 4 RT103 predicted survival time did not reach the median, but approximately 40% reach 380 days. Presented in Figures 26 and 27 are the predicted survival functions for week 4 RNA and week 4 RT103, respectively. Patients with week 4 RNA levels >10,000 copies/mL are predicted to experience virologic failure at a faster rate than those with <10,000 copies/mL (Figure 26). Additionally, patients with mutant RT103 at week 4 are predicted to experience virologic failure at a faster rate than those with wildtype RT103 (Figure 27). The median estimated survival time was about 130 days for RT103 mutant and about 350 days for RT103 wildtype, showing a better predicted survival experience for RT103 wildtype at week 4.

3.10 Simulation

In order to demonstrate performance of results from the previous section, we performed a limited simulation study to evaluate properties of the derived test statistic for calculating prediction error in survival models.

We calculated the empirical distribution and variance of $D$ for a specific set of fixed covariates by repeated sampling of $n$ observations from a Cox model, where the conditional hazard function is given by

$$\lambda(t|Z) = \lambda_0(t)e^{\beta_1 Z_1 + \beta_2 Z_2},$$

where $\lambda_0(t) = 0.1$, $\beta_1 = 0$, $\beta_2 = \ln(2)$, $Z_1$ and $Z_2$ are uncorrelated N(0,1).

Therefore we know that the true model is that with covariate $Z_2$ only. The simulated survival times were subjected to random censoring, therefore we also simulated data from a
censoring distribution, where its conditional hazard function is given by

$$\lambda_c(t|Z) = \lambda_0(t)e^{\gamma_1 Z_1 + \gamma_2 Z_2},$$

where $\lambda_0(t) = 0.04$, $\gamma_1 = 0.5$, $\gamma_2 = 0.75$, $Z_1$ and $Z_2$ are uncorrelated N(0,1).

At each iteration, we calculated the average prediction error ($D$) and variance. In practice, consistent estimates of the variance are desired in order to make inferences about the prediction error estimate no matter if the working model is correctly specified. To examine the effects of sample size on estimation, we conducted simulations with $n = 200, 500, 800$, and 2000. These parameter settings correspond to about 30% censoring. In order to make the simulation study feasible, we considered five working Cox models with covariates as follows:

\begin{align*}
S_1 &= Z_1 \\
S_2 &= Z_2 \\
S_3 &= \{Z_1, Z_2\} \\
S_4 &= \{Z_1, Z_2, Z_1Z_2\} \\
S_5 &= \{Z_1, Z_2, Z_1^2, Z_2^2\}
\end{align*}

where we know that model $S_2$ is the correct model, $S_1$ omits $Z_2$ corresponding to a mis-specified model, and $S_3$, $S_4$, and $S_5$ are inefficient. A total of 500 simulation iterations were performed to investigate if the derived analytical results are supported numerically. This was chosen due to computing efficiency and the ability to still serve as an acceptable comparison of the before mentioned methods. Also, we investigate consequences of not considering the stochastic nature of the MPLE in calculating the variance of
Results of the simulation are presented in Table 18. Across all sample sizes model S2 presented with the lowest prediction error. Additionally, model S2 consistently had the smallest average prediction error as well as the smallest prediction error for each iteration and, as $N$ gets larger, this proportion approaches 1. For assessing the variance estimates for $D$, across all values on $N$ the estimated variance is between 0.5 and 0.6. This indicated that measure derived for prediction error has a conservative variance estimate resulting in wide confidence intervals. Ideally the variance would be close to 1. If we wanted improved estimates of variance we could consider the expansions method as in Paper 2. For example, we could consider the Weibull model for cumulative baseline hazards. If we wish to obtain bootstrap estimates we would also need to use this type of method. However, as a starting point in obtaining prediction error estimates for survival models, the approach of using consistent, sample-based variance estimates are used. Furthermore, future work should consider investigating other methods such as the bootstrap, and parametrically modeling the cumulative baseline hazard.

### 3.11 Discussion

The early identification of biomarkers associated with treatment failure is vital due to the high proportion of individuals who do not respond to treatment effectively. Timely identification of these biomarkers may have a profound effect on the selection of appropriate care and management of patients. As found previously through polychotomous models, the addition of week 4 information in a survival model was found to contribute significantly to the prediction of time to off track. This information is vital and can be utilized during clinical management of HIV-1 patients.
This research project utilized former models to assess if previous findings, yielding improved model fit due to the addition of week 4 clinical and genetic factors, hold true in a survival model. Survival analysis yielded similar results as the previously performed polychotomous modeling. In both polychotomous modeling and survival modeling, inclusion of more week 4 information resulted in a smaller prediction error, meaning the most accurate prediction of the outcome, time to off track, can be better identified utilizing week 4 information. However, survival models found model $M_3$ to have the best model fit, which included week 4 clinical and genotypic information, compared to the baseline model and models with less week 4 information. In addition to model $M_3$, it was also apparent that the model with the week 4 clinical predictor fit better than the baseline model. Overall, as week 4 information was added, the prediction error decreased.

In addition to model $M_3$ having the best fit, functional form for continuous covariates and the PH assumption was met for all covariates after necessary variable transformations. Baseline CD4 and baseline RNA had appropriate functional form and PH across time. All other covariates, excluding week 4 RNA, had PH as well. Week 4 RNA did not have good functional form or PH. After multiple graphical checks it was clear that a transformation of week 4 RNA was a necessary part of the model. After week 4 RNA was transformed from continuous into a binary variable and added to the model, the functional form was appropriate for baseline CD4 and RNA and all covariates had PH. These nonlinear effects arise when the effect of a covariate differs across values of that covariate over time. Fitting the incorrect functional form to a covariate is a form of model mis-specification and leads to statistical errors, such as, bias and decreased power of tests for statistical significance (Lagakos and Schoenfeld, 1984; Struthers and Kalbfleisch, 1986; Therneau, Grambsch and Fleming, 1990; Anderson and Fleming, 1995). In general, the effects of incorrect functional form are similar to that of non-PH, which is itself a form of model mis-specification. Incorrect functional form for a covariate is one model failure that can lead to a diagnosis of
non-PH (Grambsch and Therneau, 2000), which was apparent here and resolved appropriately.

Through simulation, we identified that this new method of calculating prediction error for survival models performed well in practice without needing to derive variance estimates through Taylor series expansion. It was shown that the variance estimates were a bit too large since the standardized versions have variances less than 1. If we wanted to get improved estimates of variance we could consider the variance expansions as described in Paper 2 for survival models by parameterizing the baseline hazard. We could also consider bootstrap estimates of error, which would additionally require such parameterization techniques.

In addition, a large contribution of this paper was the adaptation of a distribution free model comparison methodology by extending previous prediction error techniques for survival models to allow for the inclusion of information about censored subjects. The implementation of this technique has far reaching uses beyond the scope of this paper and can be used in other survival outcome settings.

Clinically, the information found during these research efforts are essential to clinicians monitoring treatment regimens for patients with HIV-1. The addition of week 4 information is vital to predicting week 24 RNA measures, thus allowing for the identification of virologic failure quickly. Using this information, clinicians may alter the treatment course leading to a better outcome for patients and potentially preventing virologic failure. The findings of this research, utilizing a survival model, support the current NIH treatment guidelines, Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents, which propose viral load testing at 2-8 weeks after initiating or changing antiretroviral therapy, with HIV-1 RNA genotyping performed if viral load remains non-
suppressed. The early identification of virologic failure may assist in placing patients on the appropriate treatment regimens quickly.
4 Future Work

It is important to validate that the model fits well on data independent from the data used in the fitting procedure. A fitted model may prove to be a good fit for the data sample. However, to prove if a model is a good prognostic model, it should be validated. Validation data is not always available and it may be too costly to collect new data. Techniques such as bootstrapping and cross-validation partition the data at hand into training and validation datasets in order to get predictions separately without acquiring new data. When validating a working model, we need to consider its ability to predict not only the outcomes for the particular individuals in a cohort, but also the outcome for any other individual from the study population. Using the prediction errors we just computed above, we can evaluate the validity of the model for a population and not only for the data at hand. With bootstrapping techniques we can repeatedly simulate training samples and validation sets, so that we can estimate prediction errors with the availability of two different data sets.

Given the available data $Z$, the simplest bootstrap approach to estimate prediction errors has been described previously (Efron and Tibshirani, 1993). The following method contains four steps. The first, draw at random and with replacement a sample of size $n$ from $Z$, obtaining bootstrap sample $Z^*$. Second, estimate the prediction rule using the bootstrap sample $Z^*$. Then, test the rule on the original data $Z$ and estimate the prediction error. Therefore, $Z^*$ is the training sample and $Z$ is the validation set. Third, repeat steps one and two $B$ times to obtain a set of $B$ estimates of prediction errors. The forth and final step is to compute the average over the $B$ estimates to obtain the bootstrap estimate of prediction errors.

However, obtaining these bootstrap estimates is complicated for the same reason as the sample variance estimates are used instead of using expansion techniques to obtain better variance estimates. The reason is that in order to do the expansion, we would need to pa-
rameterize the baseline hazard which then results in specific formulae. Another approach is to accommodate the infinite dimensional parameter $\Lambda_0(\cdot)$, by using empirical process theory. We will not consider this approach at this time.

When using a parametric working model for $\Lambda(t)$, we can compute

$$D = \int \hat{M}_i^2(u)du$$

where $\hat{M}_i(t) = N_i(t) - e^{\hat{\beta}Z_i} \int_0^t Y_i(u)$

$$D(\Lambda_0, \beta) = \frac{1}{n} \sum_{i=1}^{n} \int_{\tau}^{t} \hat{M}_i^2(u)du$$

Using a Taylor series expansion,

$$D(\hat{\Lambda}_0, \hat{\beta}) = D(\Lambda_0, \beta) + \left( \begin{array}{c} \Lambda_0 - \hat{\Lambda}_0 \\ \beta - \hat{\beta} \end{array} \right) D'(\hat{\Lambda}_0, \hat{\beta})$$

where

$$D'(\hat{\Lambda}_0, \hat{\beta}) = \frac{\partial D(\Lambda_0, \beta)}{\partial \beta} = \frac{1}{n} \sum_{i=1}^{n} \hat{M}_i^2(\tau)$$

such that

$$\text{var} \left[ D(\Lambda_0, \beta) \right] = \frac{1}{n} \text{var} \int_{\tau}^{t} M_i^2(u)du$$

Taking the derivative with respect to $\beta$: 

105
\[
\frac{\partial}{\partial \beta} \int_0^\tau M_i^2(u) du = \int_0^\tau \frac{\partial}{\partial \beta} M_i^2(u) du
\]

(36)

\[
= \int_0^\tau \frac{\partial}{\partial \beta} \left[ \hat{M}_i(u) \right]^2
\]

(37)

\[
= \int_0^\tau 2\hat{M}_i(u) \frac{\partial}{\partial \beta} \hat{M}_i(u)
\]

(38)

\[
= \int_0^\tau 2\hat{M}_i(u) \left[ -Z_i e^{\hat{\beta}Z_i} \int_0^u Y_i(s) d\Lambda_0(s) \right]
\]

(39)

If we use the Weibull distribution, the derivative with respect to \( \gamma, \alpha \):

\[
\frac{\partial}{\partial (\gamma, \alpha)} \int_0^\tau M_i^2(u) du = \int_0^\tau \frac{\partial}{\partial (\gamma, \alpha)} M_i^2(u) du
\]

(40)

\[
= \int_0^\tau 2\hat{M}_i(u) \frac{\partial}{\partial (\gamma, \alpha)} \hat{M}_i(u)
\]

(41)

\[
= \int_0^\tau 2\hat{M}_i(u) \left[ -e^{\hat{\beta}Z_i} \int_0^u Y_i(s) d\Lambda_0(s) \right]
\]

(42)

Recall the Weibull Distribution:

\[
f(t) = \frac{\beta}{\alpha} \left( \frac{t}{\alpha} \right)^{\beta-1} e^{-\left( \frac{t}{\alpha} \right)^\beta}
\]

\[
S(t) = e^{-\left( \frac{t}{\alpha} \right)^\beta}
\]

\[
\Lambda(t) = \left( \frac{t}{\alpha} \right)^\beta
\]

\[
\lambda(t) = \frac{\beta}{\alpha} \left( \frac{t}{\alpha} \right)^{\beta-1}
\]

NOTE: when \( \alpha = 1 \), this is the exponential distribution.

In order to get parameter estimates for equation (42) the derivative of Weibull with respect
to $\alpha$ is:

$$\frac{d}{d\alpha} \left[ \frac{\beta}{\alpha} \right] = -\frac{\beta}{\alpha^2}$$

$$\frac{d}{d\alpha} \left[ \left( \frac{t}{\alpha} \right)^{\beta-1} \right] = -\frac{\beta - 1}{\alpha} \left( \frac{t}{\alpha} \right)^{\beta-1}$$

Using the product rule and simplifying/combining terms:

$$\frac{d\lambda}{d\alpha} = -\frac{\beta}{\alpha} \left( \frac{\beta - 1}{\alpha} \right) \left( \frac{t}{\alpha} \right)^{\beta-1} - \left( \frac{t}{\alpha} \right)^{\beta-1} \left( \frac{\beta}{\alpha^2} \right)$$

Derivative of Weibull with respect to $\beta$:

$$\frac{d\lambda}{d\beta} \left[ \frac{\beta}{\alpha} \right] = \frac{1}{\alpha}$$

$$\frac{d\lambda}{d\beta} \left[ \left( \frac{t}{\alpha} \right)^{\beta-1} \right] = \left( \frac{t}{\alpha} \right)^{\beta-1} \ln \left( \frac{t}{\alpha} \right)$$

Using the product rule and simplifying/combining terms:

$$\frac{d\lambda}{d\beta} = \ln \left( \frac{t}{\alpha} \right) \left( \frac{\beta}{\alpha} \right) \left( \frac{t}{\alpha} \right)^{\beta-1} + \left( \frac{1}{\alpha} \right) \left( \frac{t}{\alpha} \right)^{\beta-1}$$

It is also possible that we do not need to assume non-informative censoring ($T \perp C|Z$) when constructing a measure of model fit. In this case, there are three other options to
explore. Define

\[ E[A(t)] = E \left( \int Y_i(t) d\Lambda_i(t) \right) \]

\[ = \int E(Y_i(t)) d\Lambda_i(t) \]

\[ = \int P(Y_i(t) = 1) d\Lambda_i(t) \]

\[ = \int P(U_i \geq t) d\Lambda_i(t) \]

First, we could replace \( \hat{P}(C_i \geq v|Z_i) \times \hat{P}(T_i \geq v|Z_i) \) in equation (30) with \( P(U_i \geq t) \) which is equal to \( \sum_{i=1}^{n} I(v_i \geq t) \). Second, since we have \( U \) observed for everyone, we could use the generalized extreme value distribution, or some other parametric distribution, for \( P(U_i \geq t|Z_i) \), a "finite-dimensional" parametric approach that incorporated covariates. We also can extend the second approach to also include a parametric for the conditional variance of \( E \left( (U_i - \hat{\theta}Z_i)^2 | Z_i \right) \) and estimate the parameters of the mean and variance model simultaneously.

To consider obtaining an alternate estimate of \( \text{var}(\hat{D}) \), we will consider a re-sampling approach. For working model

\[ \hat{\Lambda}_i(t) = \hat{\Lambda}(t|Z_i) = \hat{\Lambda}_0(t)e^{\hat{\beta}Z_i} \]

we can write

\[ \hat{D} = \frac{1}{n} \sum_{i=1}^{n} B_i \left( (\hat{k}, \hat{\mu}, \hat{\sigma}), \hat{\Lambda}_0(\cdot), \hat{\beta} \right) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} B_i \left( \hat{\gamma}, \hat{\Lambda}_0(\cdot) \right) \]

where \( \hat{\gamma} \) represents all finite dimensional parameters.
We can re-write $\hat{D}$ to be equal to

$$
\begin{align*}
&= \frac{1}{n} \sum_{i=1}^{n} \left( B_i(\hat{\gamma}, \hat{\Lambda}_0(\cdot)) - B_i(\hat{\gamma}, \Lambda_0(\cdot)) + B_i(\hat{\gamma}, \hat{\Lambda}_0(\cdot)) \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left( B_i(\hat{\gamma}, \hat{\Lambda}_0(\cdot)) - B_i(\hat{\gamma}, \Lambda_0(\cdot)) \right) + \frac{1}{n} \sum_{i=1}^{n} B_i(\hat{\gamma}, \Lambda_0(\cdot))
\end{align*}
$$

The second term on the right hand side above can easily be shown to be asymptotically equivalent to a sum of i.i.d. terms using Taylor series approximation with finite dimensional parameters as shown similarly in Paper 2.

The first term on the right hand side above can be written as a function involving

$$
d \left( \hat{\Lambda}_0(\cdot) - \Lambda_0(\cdot) \right)
$$

Since

$$Q(t) = \sqrt{n} \left( \hat{\Lambda}_0(t) - \Lambda_0(t) \right), \quad 0 < t < \tau$$

can be shown to converge weakly to a martingale process with mean function zero and variance function estimated by

$$
\int_{0}^{t} \frac{d\bar{N}(u)}{\left( \sum_{i=1}^{n} Y_i(u)e^{\hat{\beta}Z_i} \right)^2}
$$

As a result, we can easily simulate from this process using the powerful results of Lin, Wei, Ying (1993). This will enable us the estimate the variance of $\hat{D}(\hat{\gamma}, \hat{\Lambda}_0(\cdot))$ by invoking the double variance formula in a simulation scheme. That is, since

$$
\text{var} \left[ B(\hat{\gamma}, \hat{\Lambda}_0(\cdot)) \right] = E_{\hat{\Lambda}_0(\cdot)}E \left[ B(\hat{\gamma}, \hat{\Lambda}_0(\cdot))^2 | \hat{\Lambda}_0(\cdot) \right] + E_{\hat{\Lambda}_0(\cdot)} \left( E \left[ B(\hat{\gamma}, \hat{\Lambda}_0(\cdot)) | \hat{\Lambda}_0(\cdot) \right] \right)^2
$$

109
we can use the fact that the unconditional distribution of $\hat{\Lambda}_0(\cdot)$ can be approximated by the conditional distribution of $\bar{\Lambda}_0(\cdot)$ given the data, where $\bar{\Lambda}_0(\cdot)$ is a simulated version of $\hat{\Lambda}_0(\cdot)$. 
References


Breslow NE. Covariance analysis of censored survival data. Biometrics 1974;30:89-100

Brier GW. The statistical theory of turbulence and the problem of diffusion in the atmo-
sphere. J. Meteor. 1950;7:283290


Gerds T, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biometrical Journal 2006; 48:10291040

Gerds T, Schumacher M. Efron-type measures of prediction error for survival analysis. Biometrics 2007;63:12831287


Grambsch PM, Therneau TM. Proportional hazards tests in diagnostics based on weighted residuals. Biometrika 1994;81:515-526


Hosmer DW, Lemeshow S. A goodness-of-fit test for multiple logistic regression model.


Ireland CT, Ku HH, Kullback S. Symetry and marginal homogeneity of an r x r contingency table. J Amer Statist Assoc;64:1323-1341


Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale residuals. Biometrika 1993;80:557-572


McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153-157


Pallela F, Chimel JS, Moorman AC, Holmberg SD. Durability and predictors of success of


Schumacher M, Binder H, Gerds T. Assessment of survival prediction models based on microarray data, Bioinformatics 2007;23(14):1768-1774


Struthers CA, Kalbfleisch JD. Misspecified proportional hazard models. Biometrika 1986;73: 363-369


Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals for survival mod-

Tian L, Cai T, Goetghebeur E, Wei LJ. Model evaluation based on the sampling distribution of estimated absolute prediction error. Biometrika 2007;92:297-311


Appendix I

Conditional independence null hypotheses for fourth-order combinations

Consider the fourth-order variable/level combination $J_{r_1;\ell_1}, r_2;\ell_2, r_3;\ell_3, r_4;\ell_4$, with $r_i = 1, \ldots, R$, $r_i \neq r_j$, $i \neq j$ and $\ell_i = 1, \ldots, L_{r_i}$. There are seven conditional independence null hypotheses in each group $W = 0, 1$. To ease notation, define

$r_1;\ell_1, r_2;\ell_2, r_3;\ell_3, r_4;\ell_4 \equiv \text{comb}$.

$Q_{\text{comb}}^{W,1} : P(J_{\text{comb}} = 1|W) = P(J_{r_1;\ell_1} = 1|W) P(J_{r_2;\ell_2, r_3;\ell_3, r_4;\ell_4} = 1|W)$,

$Q_{\text{comb}}^{W,2} : P(J_{\text{comb}} = 1|W) = P(J_{r_2;\ell_2} = 1|W) P(J_{r_1;\ell_1, r_3;\ell_3, r_4;\ell_4} = 1|W)$,

$Q_{\text{comb}}^{W,3} : P(J_{\text{comb}} = 1|W) = P(J_{r_3;\ell_3} = 1|W) P(J_{r_1;\ell_1, r_2;\ell_2, r_4;\ell_4} = 1|W)$,

$Q_{\text{comb}}^{W,4} : P(J_{\text{comb}} = 1|W) = P(J_{r_4;\ell_4} = 1|W) P(J_{r_1;\ell_1, r_2;\ell_2, r_3;\ell_3} = 1|W)$,

$Q_{\text{comb}}^{W,5} : P(J_{\text{comb}} = 1|W) = P(J_{r_1;\ell_1, r_2;\ell_2} = 1|W) P(J_{r_3;\ell_3, r_4;\ell_4} = 1|W)$,

$Q_{\text{comb}}^{W,6} : P(J_{\text{comb}} = 1|W) = P(J_{r_1;\ell_1, r_3;\ell_3} = 1|W) P(J_{r_2;\ell_2, r_4;\ell_4} = 1|W)$,

$Q_{\text{comb}}^{W,7} : P(J_{\text{comb}} = 1|W) = P(J_{r_1;\ell_1, r_4;\ell_4} = 1|W) P(J_{r_2;\ell_2, r_3;\ell_3} = 1|W)$.
5 Tables

<table>
<thead>
<tr>
<th></th>
<th>Post-Base $RT_{74}$</th>
<th></th>
<th>Post-Base $RT_{100}$</th>
<th></th>
<th>Post-Base $RT_{103}$</th>
<th></th>
<th>Post-Base $RT_{108}$</th>
<th></th>
<th>Post-Base $RT_{190}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base $RT_{74}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>240</td>
<td>24</td>
<td>262</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>43</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>240</td>
<td>65</td>
<td>305</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base $RT_{101}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>252</td>
<td>25</td>
<td>277</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>25</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>255</td>
<td>50</td>
<td>305</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base $RT_{103}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>101</td>
<td>133</td>
<td>234</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>67</td>
<td>71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>276</td>
<td>29</td>
<td>305</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base $RT_{108}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>264</td>
<td>30</td>
<td>294</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>9</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>266</td>
<td>39</td>
<td>305</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base $RT_{190}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>228</td>
<td>39</td>
<td>267</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>35</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>231</td>
<td>74</td>
<td>305</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: 2-way Joint Changes for RT108 and RT190

<table>
<thead>
<tr>
<th>Post-Base $RT_{108}$ &amp; $RT_{190}$</th>
<th>(0,0)</th>
<th>(0,1)</th>
<th>(1,0)</th>
<th>(1,1)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0)</td>
<td>63</td>
<td>19</td>
<td>112</td>
<td>2</td>
<td>206</td>
</tr>
<tr>
<td>(0,1)</td>
<td>0</td>
<td>19</td>
<td>2</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>Base $RT_{108}$ &amp; $RT_{190}$</td>
<td>(1,0)</td>
<td>2</td>
<td>0</td>
<td>51</td>
<td>61</td>
</tr>
<tr>
<td>(1,1)</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>65</td>
<td>40</td>
<td>166</td>
<td>34</td>
<td>305</td>
</tr>
</tbody>
</table>

Table 3: 2-way Joint Changes for RT181 and RT190

<table>
<thead>
<tr>
<th>Post-Base $RT_{181}$ &amp; $RT_{190}$</th>
<th>(0,0)</th>
<th>(0,1)</th>
<th>(1,0)</th>
<th>(1,1)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0)</td>
<td>184</td>
<td>18</td>
<td>2</td>
<td>4</td>
<td>208</td>
</tr>
<tr>
<td>(0,1)</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>Base $RT_{181}$ &amp; $RT_{190}$</td>
<td>(1,0)</td>
<td>4</td>
<td>1</td>
<td>38</td>
<td>59</td>
</tr>
<tr>
<td>(1,1)</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>22</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>191</td>
<td>30</td>
<td>40</td>
<td>44</td>
<td>305</td>
</tr>
</tbody>
</table>

Table 4: 3-way Joint Changes

<table>
<thead>
<tr>
<th>Base RT190, Base RT44, Base PR5</th>
<th>(0,0,0)</th>
<th>(0,0,1)</th>
<th>(0,1,0)</th>
<th>(0,1,1)</th>
<th>(1,0,0)</th>
<th>(1,0,1)</th>
<th>(1,1,0)</th>
<th>(1,1,1)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0)</td>
<td>201</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>235</td>
</tr>
<tr>
<td>(0,0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0,1,0)</td>
<td>1</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>(0,1,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,0,0)</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>(1,0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>(1,1,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>205</td>
<td>0</td>
<td>26</td>
<td>0</td>
<td>64</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>305</td>
</tr>
</tbody>
</table>
Table 5: Predictors Included in Polychotomous Models M0-M4

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_0$</td>
<td>$CD_4(0), RNA(0), RT_{100}(0), RT_{103}(0)$</td>
</tr>
<tr>
<td>$M_1^*$</td>
<td>$CD_4(0), RNA(0), RT_{100}(0), RT_{103}(0), RNA(4)^*RT_{103}(4), CD_4(0)^*RT_{103}(4)$</td>
</tr>
<tr>
<td>$M_2$</td>
<td>$CD_4(0), RNA(0), RT_{100}(0), RT_{103}(0), RNA(4)$</td>
</tr>
<tr>
<td>$M_3$</td>
<td>$CD_4(0), RNA(0), RT_{100}(0), RT_{103}(0), RNA(4), RT_{103}(4)$</td>
</tr>
<tr>
<td>$M_4$</td>
<td>$CD_4(0), RNA(0), RT_{100}(0), RT_{103}(0), RNA(4), RT_{103}(4), RT_{74}(0), RT_{100}(0), RT_{101}(0), RT_{103}(0), RT_{108}(0), RT_{190}(0), RT_{103}(0)^*RT_{190}(0), RT_{181}(0)^*RT_{190}(0), PR_5(0)^*RT_{144}(0)^*RT_{190}(0), RT_{103}(4)^*RT_{190}(4), RT_{181}(4)^*RT_{190}(4)$, and $PR_5(4)^*RT_{144}(4)^*RT_{190}(4)$</td>
</tr>
</tbody>
</table>

Note: Model $M_1^*$ had best model fit with interaction terms based on Hosmer-Lemeshow test statistic.

Table 6: P-values for Checking Functional Form of Continuous Covariates and the Link Function for Polychotomous Models M0-M3

<table>
<thead>
<tr>
<th>Model</th>
<th>var/link</th>
<th>LEVEL 1</th>
<th>LEVEL 2</th>
<th>LEVEL 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0</td>
<td>$CD_4_0$</td>
<td>0.63</td>
<td>0.17</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>$RNA_0$</td>
<td>0.10</td>
<td>0.46</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>LINK</td>
<td>0.84</td>
<td>0.25</td>
<td>0.01</td>
</tr>
<tr>
<td>M1</td>
<td>$CD_4_0$</td>
<td>0.65</td>
<td>0.22</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>$RNA_0$</td>
<td>0.16</td>
<td>0.34</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>$CD_4_0^{*week_4} RT_{103}$</td>
<td>0.44</td>
<td>0.30</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>$RNA_0^{*week_4} RT_{103}$</td>
<td>0.13</td>
<td>0.16</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>LINK</td>
<td>0.59</td>
<td>0.15</td>
<td>0.26</td>
</tr>
<tr>
<td>M2</td>
<td>$CD_4_0$</td>
<td>0.75</td>
<td>0.18</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>$RNA_0$</td>
<td>0.10</td>
<td>0.41</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>LINK</td>
<td>0.61</td>
<td>0.40</td>
<td>0.21</td>
</tr>
<tr>
<td>M3</td>
<td>$CD_4_0$</td>
<td>0.75</td>
<td>0.23</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>$RNA_0$</td>
<td>0.10</td>
<td>0.43</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>LINK</td>
<td>0.61</td>
<td>0.49</td>
<td>0.34</td>
</tr>
</tbody>
</table>
Table 7: P-values for Checking Functional Form of Continuous Covariates and the Link Function for Polychotomous Model M4

<table>
<thead>
<tr>
<th>Variable</th>
<th>Level (1,2) vs. (3,4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4</td>
<td>0.25</td>
</tr>
<tr>
<td>RNA</td>
<td>0.50</td>
</tr>
<tr>
<td>LINK</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 8: Prediction Error (PE) Results for Polychotomous Models

<table>
<thead>
<tr>
<th>Model</th>
<th>PE (squared)</th>
<th>PE (abs. value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_0$</td>
<td>0.8772</td>
<td>0.6757</td>
</tr>
<tr>
<td>$M_1$</td>
<td>0.8130</td>
<td>0.6427</td>
</tr>
<tr>
<td>$M_2$</td>
<td>0.8287</td>
<td>0.6453</td>
</tr>
<tr>
<td>$M_3$</td>
<td>0.7847</td>
<td>0.6281</td>
</tr>
<tr>
<td>$M_4$</td>
<td>0.1921</td>
<td>0.3363</td>
</tr>
</tbody>
</table>
Table 9: Selected Model Results for $M_3$

<table>
<thead>
<tr>
<th>Variable</th>
<th>Week 24 RNA Level</th>
<th>OR</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 4 RNA: Level 1</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td>3.4</td>
<td>0.6, 19.7</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>3.1</td>
<td>1.0, 9.8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>1.1</td>
<td>0.4, 3.1</td>
</tr>
<tr>
<td>Week 4 RNA: Level 2</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td>13.0</td>
<td>0.9, 180.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>14.0</td>
<td>2.0, 100.3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>13.8</td>
<td>2.2, 86.9</td>
</tr>
<tr>
<td>Week 4 RNA: Level 3</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td>13.8</td>
<td>2.1, &gt;100</td>
</tr>
<tr>
<td>Week 4 RT103: Wild Type</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
<td>Ref</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td>3.5</td>
<td>0.1, 236.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>19.1</td>
<td>0.8, 478.1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>50.4</td>
<td>2.1, &gt;100</td>
</tr>
</tbody>
</table>

Table 10: Model Comparisons for Polychotomous Models

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Hypothesis</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1**</td>
<td>$M_0$ vs. $M_1$</td>
<td>0.0120</td>
</tr>
<tr>
<td>2</td>
<td>$M_0$ vs. $M_2$</td>
<td>0.0548</td>
</tr>
<tr>
<td>3*</td>
<td>$M_0$ vs. $M_3$</td>
<td>0.0047</td>
</tr>
<tr>
<td>4*</td>
<td>$M_0$ vs. $M_4$</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>5</td>
<td>$M_1$ vs. $M_2$</td>
<td>0.6277</td>
</tr>
<tr>
<td>6</td>
<td>$M_1$ vs. $M_3$</td>
<td>0.1780</td>
</tr>
<tr>
<td>7*</td>
<td>$M_1$ vs. $M_4$</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>8**</td>
<td>$M_2$ vs. $M_3$</td>
<td>0.0159</td>
</tr>
<tr>
<td>9*</td>
<td>$M_2$ vs. $M_4$</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>10*</td>
<td>$M_3$ vs. $M_4$</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

* Rejected using Holm step-down multiple testing correction
**Rejected using Lehman and Ramano step-down multiple testing correction
Table 11: Polychotomous Model Simulation Results

<table>
<thead>
<tr>
<th>Simulation Models, N=200</th>
<th>Simulation Models, N=500</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S1</strong></td>
<td><strong>S2</strong></td>
</tr>
<tr>
<td>Prop. Min.</td>
<td>0</td>
</tr>
<tr>
<td>Mean $D$</td>
<td>1.23</td>
</tr>
<tr>
<td>Mean $\left(\frac{D}{\sigma}\right)$</td>
<td>16.58</td>
</tr>
<tr>
<td>SD $\left(\frac{D}{\sigma}\right)$</td>
<td>1.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation Models, N=800</th>
<th>Simulation Models, N=2000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S1</strong></td>
<td><strong>S2</strong></td>
</tr>
<tr>
<td>Prop. Min.</td>
<td>0</td>
</tr>
<tr>
<td>Mean $D$</td>
<td>1.22</td>
</tr>
<tr>
<td>Mean $\left(\frac{D}{\sigma}\right)$</td>
<td>32.22</td>
</tr>
<tr>
<td>SD $\left(\frac{D}{\sigma}\right)$</td>
<td>1.15</td>
</tr>
</tbody>
</table>

Table 12: Predictors Included in Survival Models M0-M4

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_0$</td>
<td>CD4$<em>{(0)}$, RNA$</em>{(0)}$, RT100$<em>{(0)}$, RT103$</em>{(0)}$</td>
</tr>
<tr>
<td>$M_1$</td>
<td>CD4$<em>{(0)}$, RNA$</em>{(0)}$, RT100$<em>{(0)}$, RT103$</em>{(0)}$, RT103$_{(4)}$</td>
</tr>
<tr>
<td>$M_2$</td>
<td>CD4$<em>{(0)}$, RNA$</em>{(0)}$, RT100$<em>{(0)}$, RT103$</em>{(0)}$, RNA$_{(4)}$</td>
</tr>
<tr>
<td>$M_3$</td>
<td>CD4$<em>{(0)}$, RNA$</em>{(0)}$, RT100$<em>{(0)}$, RT103$</em>{(0)}$, RT103$<em>{(4)}$, RNA$</em>{(4)}$</td>
</tr>
</tbody>
</table>

Table 13: Supremum Test for Functional Form

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pr &gt; Max Abs Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline CD4</td>
<td>0.1982</td>
</tr>
<tr>
<td>Baseline RNA</td>
<td>0.2496</td>
</tr>
</tbody>
</table>

126
Table 14: Supremum Test for Proportional Hazards Assumption

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pr &gt; Max Abs Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline CD4</td>
<td>0.9896</td>
</tr>
<tr>
<td>Baseline RNA</td>
<td>0.9496</td>
</tr>
<tr>
<td>Baseline RT100</td>
<td>0.7822</td>
</tr>
<tr>
<td>Baseline RT103</td>
<td>0.9842</td>
</tr>
<tr>
<td>Week 4 RT103</td>
<td>0.9932</td>
</tr>
<tr>
<td>Week 4 RNA</td>
<td>0.6228</td>
</tr>
</tbody>
</table>

Table 15: Average Squared Prediction Error Estimates for Survival Models

<table>
<thead>
<tr>
<th>Model</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_0$</td>
<td>1.82</td>
</tr>
<tr>
<td>$M_1$</td>
<td>2.05</td>
</tr>
<tr>
<td>$M_2$</td>
<td>1.56</td>
</tr>
<tr>
<td>$M_3$</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 16: Model Comparisons for Survival Models

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Hypothesis</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$M_0$ vs. $M_1$</td>
<td>0.9963</td>
</tr>
<tr>
<td>2*</td>
<td>$M_0$ vs. $M_2$</td>
<td>0.0111</td>
</tr>
<tr>
<td>3*</td>
<td>$M_0$ vs. $M_3$</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>4*</td>
<td>$M_1$ vs. $M_2$</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>5*</td>
<td>$M_1$ vs. $M_3$</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>6*</td>
<td>$M_2$ vs. $M_3$</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

* Rejected using Holm step-down multiple testing correction
### Table 17: Survival Model Results for $M_3$

<table>
<thead>
<tr>
<th>Variable</th>
<th>Hazard Ratio</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline CD4</td>
<td>1.02</td>
<td>0.94, 1.11</td>
<td>0.59</td>
</tr>
<tr>
<td>Baseline RNA</td>
<td>1.02</td>
<td>0.88, 1.18</td>
<td>0.79</td>
</tr>
<tr>
<td>Baseline RT100</td>
<td>&lt;0.01</td>
<td>&lt;0.01, 0.01</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Baseline RT103</td>
<td>0.75</td>
<td>0.48, 1.17</td>
<td>0.20</td>
</tr>
<tr>
<td>Week 4 RT103</td>
<td>2.44</td>
<td>1.61, 3.69</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Week 4 RNA</td>
<td>6.35</td>
<td>4.56, 8.84</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

### Table 18: Survival Model Simulation Results

<table>
<thead>
<tr>
<th>Simulation Models, N=200</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. Min.</td>
<td>0</td>
<td>0.65</td>
<td>0.02</td>
<td>0</td>
<td>0.13</td>
</tr>
<tr>
<td>Mean $D$</td>
<td>8.64</td>
<td>6.72</td>
<td>7.75</td>
<td>7.93</td>
<td>7.22</td>
</tr>
<tr>
<td>Mean ($\hat{D}$)</td>
<td>12.52</td>
<td>12.81</td>
<td>13.24</td>
<td>13.29</td>
<td>13.03</td>
</tr>
<tr>
<td>SD ($\hat{D}$)</td>
<td>0.49</td>
<td>0.56</td>
<td>0.42</td>
<td>0.40</td>
<td>0.32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation Models, N=500</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. Min.</td>
<td>0</td>
<td>0.78</td>
<td>0</td>
<td>0</td>
<td>0.22</td>
</tr>
<tr>
<td>Mean $D$</td>
<td>21.06</td>
<td>16.40</td>
<td>18.86</td>
<td>19.31</td>
<td>17.65</td>
</tr>
<tr>
<td>Mean ($\hat{D}$)</td>
<td>19.25</td>
<td>19.54</td>
<td>20.32</td>
<td>20.41</td>
<td>19.92</td>
</tr>
<tr>
<td>SD ($\hat{D}$)</td>
<td>0.51</td>
<td>0.55</td>
<td>0.38</td>
<td>0.38</td>
<td>0.31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation Models, N=800</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. Min.</td>
<td>0</td>
<td>0.81</td>
<td>0</td>
<td>0</td>
<td>0.19</td>
</tr>
<tr>
<td>Mean $D$</td>
<td>33.37</td>
<td>25.98</td>
<td>28.87</td>
<td>30.60</td>
<td>27.66</td>
</tr>
<tr>
<td>Mean ($\hat{D}$)</td>
<td>24.17</td>
<td>24.51</td>
<td>25.53</td>
<td>25.62</td>
<td>24.97</td>
</tr>
<tr>
<td>SD ($\hat{D}$)</td>
<td>0.47</td>
<td>0.51</td>
<td>0.40</td>
<td>0.39</td>
<td>0.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation Models, N=2000</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. Min.</td>
<td>0</td>
<td>0.91</td>
<td>0</td>
<td>0</td>
<td>0.09</td>
</tr>
<tr>
<td>Mean $D$</td>
<td>82.53</td>
<td>63.94</td>
<td>73.76</td>
<td>75.58</td>
<td>67.97</td>
</tr>
<tr>
<td>Mean ($\hat{D}$)</td>
<td>37.91</td>
<td>38.37</td>
<td>39.99</td>
<td>40.15</td>
<td>39.09</td>
</tr>
<tr>
<td>SD ($\hat{D}$)</td>
<td>0.45</td>
<td>0.55</td>
<td>0.36</td>
<td>0.35</td>
<td>0.31</td>
</tr>
</tbody>
</table>
6 Figures

Figure 1: Model M0, Level 1

Figure 2: Model M0, Level 2

Figure 3: Model M0, Level 3

Figure 4: Model M1, Level 1
Figure 11: Model M3, Level 2

Figure 12: Model M3, Level 3

Figure 13: Predicted Probabilities Against Baseline RNA

Figure 14: Predicted Probabilities Against Week 4 RNA

Figure 15: Predicted Probabilities Against Baseline RT103

Figure 16: Predicted Probabilities Against Week 4 RT103
Figure 17: Generalized Extreme Plot of Errors

Figure 18: Functional Form: Baseline CD4

Figure 19: Functional Form: Baseline RNA
Figure 20: Proportional Hazards: Baseline CD4

Figure 21: Proportional Hazards: Baseline RNA

Figure 22: Proportional Hazards: Week 4 RNA

Figure 23: Proportional Hazards: Baseline RT100

Figure 24: Proportional Hazards: Baseline RT103

Figure 25: Proportional Hazards: Week 4 RT103
Figure 26: Predicted Survival Function - Week 4 RNA

Figure 27: Predicted Survival Function - Week 4 RT103