Novel Methods for the Design, Monitoring, and Analysis of Clinical Trials

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Ross Lo Peterson

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advisers: David M. Vock and Joseph S. Koopmeiners

July, 2020

Acknowledgements

I would like to thank David Vock, who has been with me for all five years and without whom none of this research would be possible. David once told me that some students get a lot of help with their dissertations, and I was one of those students. He never gave up on me. I would also like to thank Joe Koopmeiners, who for the past three years has been remarkably positive. Like David, he stayed with me even when I was struggling hard on the software for Chapter 3. Additionally, I would like to thank Jim Neaton. Under his tutelage in both research meetings and journal club, I have grown much as a biostatistician specializing in clinical trials. Lastly, I would like to thank Rich Maclehose for serving on my committee and contributing to Chapter 4.

Dedication

To the PhD students who dream of going the distance, keep going.

Abstract

Clinical trials can be divided into three phases conducted before, during, and after the trial: 1) trial design; 2) data monitoring; and 3) statistical analysis, respectfully. The scientific accuracy of any clinical trial is inherently tied to the validity of the statistical assumptions made in each phase. I present methods that seek to both measure and account for violations of such statistical assumptions. For 1), a novel ordinal endpoint was recently proposed for evaluating new treatments for patients hospitalized by influenza. I investigate the power of the ordinal endpoint under model misspecification and compare it to the power of other endpoints derived from the same information. For 2), in some trials participants are not obligated to comply with the treatment protocol, and compliance may be unobserved. Biomarker measurements, however, can provide objective information on compliance. I propose a novel method that models a longitudinal biomarker as a longitudinal mixture density to estimate various probabilities of compliance. For 3), in completed trials participant noncompliance may create a difference between the intention-to-treat estimate and the treatment effect if all participants had complied (i.e., the causal effect), as noncompliant participants may systematically differ from compliant participants on some confounding variable. The G-computation algorithm can estimate the causal effect for longitudinal trials, but typically assumes that both compliance is observed and that the confounder models can be correctly specified. I propose a modified G-computation algorithm estimator in the scenario where both assumptions are violated.

Contents

Acknowledgements i

Dedication ii

Abstract iii

List of Tables viii

List of Figures xii

1 Introduction 1

2 Analysis and Comparison of an Ordinal Endpoint to Time-to-Event, Longitudinal, and Binary Endpoints for Use in Evaluating Treatments for Severe Influenza Requiring Hospitalization 5

2.1 Introduction 5

2.2 Methods 7

2.2.1 The INSIGHT FLU-IVIG Trial 7

2.2.2 Case Study 1: Proportional Odds Model Power 9

2.2.2.1 Simulation Study Design 9

2.2.2.2 Factor 1: Treatment Effect 9

2.2.2.3 Factor 2: Distribution of the Placebo Group 11

2.2.2.4 Factor 3: Misclassification 12

2.2.2.5 Factor 4: Number of Categories 13

2.2.2.6 Interactions 13

2.2.3 Case Study 2: Comparison to Other Types of Endpoints 14
2.2.3.1 Analytic Comparison of Ordinal Endpoint to Time-to-Event Endpoint 14
2.2.3.2 Simulation Comparison of Ordinal Endpoint to Other Types of Endpoints 16
2.3 Results 21
2.3.1 Case Study 1: Proportional Odds Model Power 21
2.3.1.1 Main Effects 21
2.3.1.2 Interactions 27
2.3.2 Case Study 2: Comparison to Other Types of Endpoints 33
2.3.2.1 Analytic Comparison of Ordinal Endpoint to Time-to-Event Endpoint 33
2.3.2.2 Simulation Comparison of Ordinal Endpoint to Other Types of Endpoints 36
2.4 Discussion 42

3 Detecting Participant Noncompliance Across Multiple Time Points by Modeling a Longitudinal Biomarker 47
3.1 Introduction 47
3.2 Methods 50
3.2.1 Overview 50
3.2.2 Modeling the Mixture Density 50
3.2.3 Deriving Compliance Probabilities of Interest 52
3.2.4 Simulation Study Design 53
3.2.5 Application to CENIC-P1S2 55
3.3 Results 57
3.3.1 Simulation Study 57
3.3.2 Application to CENIC-P1S2 62
3.4 Discussion 65

4 Estimating Longitudinal Causal Effects with Unobserved Participant Noncompliance and Confounders with Densities that May Be Difficult to Parametrically Specify 69

4.1 Introduction . . . 69
4.2 Methods . . . 72
4.2.1 Dataset and Target of Inference . . . 72
4.2.2 Identifying Assumptions . . . 73
4.2.3 The G-computation Algorithm . . . 74
4.2.4 The G-computation Algorithm with Unobserved Participant Non-compliance . . . 76
4.2.5 The G-computation Algorithm with Predictive Mean Matching . . . 78
4.3 Simulation Study Design . . . 79
4.3.1 Data Generation . . . 79
4.3.2 Proposed Estimators . . . 81
4.3.3 Simulation Results . . . 82
4.4 Application to CENIC-P2 . . . 85
4.4.1 Trial Design . . . 85
4.4.2 CENIC-P2 Results . . . 88
4.5 Discussion . . . 90

5 Conclusion **93**

6 Bibliography **96**

A Chapter 2 Appendix **104**
A.1 Deviations from Proportional Odds while Maintaining the Same Overall Treatment Effect . . . 104
A.2 Derivation of the Cumulative Log Odds for the Distribution of the Placebo Group . . . 107
A.3 Misclassification Among Categories of the Ordinal Endpoint . . . 108
A.4 Derivation of the Non-Centrality Parameter for the Accelerated Failure Time Model . . . 110
A.5 In-depth Explanation of the Longitudinal Data Generating Process . . . 113

B Chapter 3 Appendix **132**
B.1 EM Algorithm Step Updates for $\boldsymbol{\theta}$ . . . 132

B.2 Misspecification of the Covariance Structure 134

C Chapter 4 Appendix **137**

C.1 Proof that the Expected Value of (4.4) is Equal to the Expected Value of (4.6) 137

C.2 EM Algorithm Step Updates 138

List of Tables

2.1 Models fitted to endpoints for the simulated data. 19
2.2 FLU-IVIG placebo group distribution of the ordinal endpoint at randomization and across the 7 day follow-up period according to the cohort study.[18,19] 19
2.3 Distributions of the FLU-IVIG placebo group and IVIG group on day 7 of follow-up. 20
2.4 Main effect of the treatment effect on power (factor 1). 23
2.5 Main effect of the distribution of the placebo group on power (factor 2). 24
2.6 Main effect of misclassification on power (factor 3). 25
2.7 Main effect of the number of categories on power (factor 4). 26
2.8 Effect of interacting the treatment effect (factor 1) and placebo group distribution (factor 2) on power. 29
2.9 Effect of interacting the treatment effect (factor 1) and misclassification (factor 3) on power. 30
2.10 Effect of interacting the treatment effect (factor 1) and number of categories (factor 4) on power. 31
2.11 Simulated power (%)[a] results for the FLU-IVIG placebo group on day 7 of follow-up (p_{07}[b]$= 0.63$).[c] 39
2.12 Simulated power (%)[a] results for the more severe placebo group on day 7 of follow-up (p_{07}[b]$= 0.42$).[c] 40
2.13 Simulated power (%)[a] results for the less severe placebo group on day 7 of follow-up (p_{07}[b]$= 0.82$).[c] 41

3.1 Estimated bias and Monte Carlo standard deviation (MC SD) of parameter estimates for $\boldsymbol{\theta}$ where $n = 100$, the number of time points $K \in \{6, 5\}$, and the compliance effect $\alpha_1 \in \{-1.5, -1.8, -2.1\}$. For both compliance and the conditional biomarker, the models fit are either correctly specified with compound symmetric covariance structures, or incorrectly specified where compound symmetric covariance structures are erroneously fit to data that is generated as AR(1). 59

3.2 True AUC values for $P(C_{i6} = 1|B_{i1}, \ldots, B_{i6})$ and $P(C_{i1} = 1, \ldots, C_{i6} = 1|B_{i1}, \ldots, B_{i6})$, and average AUC values for $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i6})$, $\hat{P}(C_{i1} = 1, \ldots, C_{i6} = 1|B_{i1}, \ldots, B_{i6})$, $\hat{P}(C_{i6} = 1|B_{i6})$, and $\prod_{j=1}^{6} \hat{P}(C_{ij} = 1|B_{ij})$, where the compliance effect $\alpha_1 \in \{-1.5, -1.8, -2.1\}$. For both compliance and the conditional biomarker, the models fit are either correctly specified with compound symmetric covariance structures, or incorrectly specified where compound symmetric covariance structures are erroneously fit to data that is generated as AR(1). 60

3.3 Parameter estimates and standard errors (SE) for $\boldsymbol{\theta}$ using all six time points ($n = 114$) and the first five time points ($n = 114$) of CENIC-P1S2 for participants who had at least one anatabine value. 63

4.1 Simulation results with $R^2 = 0.3$ for the confounding in the model of the study outcome across time points $j \in \{0, \ldots, 5\}$. For both $n = 500$ and $n = 1000$, approximately 80% of observations across simulations had self-reported compliance and thus contributed to parameter estimation. 83

4.2 Simulation results with $R^2 = 0.5$ for the confounding in the model of the study outcome across time points $j \in \{0, \ldots, 5\}$. For both $n = 500$ and $n = 1000$, approximately 80% of observations across simulations had self-reported compliance and thus contributed to parameter estimation. 84

4.3 Simulation results with $R^2 = 0.7$ for the confounding in the model of the study outcome across time points $j \in \{0, \ldots, 5\}$. For both $n = 500$ and $n = 1000$, approximately 80% of observations across simulations had self-reported compliance and thus contributed to parameter estimation. 85

4.4 CENIC-P2 results. For the ITT and Per Protocol estimators, the sample size comprises 297 and 195 participants, respectively. For all other causal estimators, the sample size comprises the 287 participants who self-reported compliance for at least one time point after baseline, where time points $j \in \{0, \ldots, 5\}$. 89
A.1 Power to detect a significant treatment effect. 115
A.2 Power for detecting a significant treatment effect for all possible ways of dividing the ordinal endpoint into a binary endpoint. 115
A.3 Altering the distribution of the placebo group by changing its cumulative log odds. 116
A.4 Effect of all possible interactions of the treatment effect (factor 1) and placebo group distribution (factor 2) on power. 117
A.5 Effect of all possible interactions of the treatment effect (factor 1) and misclassification (factor 3) on power. 118
A.6 Effect of all possible interactions of the treatment effect (factor 1) and number of categories (factor 4) on power. 119
A.7 Day 7 distributions of the ordinal endpoint for the FLU-IVIG placebo group (P1) and its corresponding treatment effects. 120
A.8 Day 7 distributions of the ordinal endpoint for the more severe placebo group (P2) and its corresponding treatment effects. 121
A.9 Day 7 distributions of the ordinal endpoint for the less severe placebo group (P3) and its corresponding treatment effects. 122
A.10 Average coefficient values across the 10,000 trials for the FLU-IVIG placebo group on day 7 of follow-up (${p_{07}}^{\text{a}} = 0.63$).[b] 123
A.11 Average coefficient values across the 10,000 trials for the more severe placebo group on day 7 of follow-up (${p_{07}}^{\text{a}} = 0.42$).[b] 124
A.12 Average coefficient values across the 10,000 trials for the less severe placebo group on day 7 of follow-up (${p_{07}}^{\text{a}} = 0.82$).[b] 125
A.13 FLU-IVIG placebo group transition matrix from day 0 to day 1. 126
A.14 FLU-IVIG placebo group transition matrix from day 1 to day 2. 126
A.15 FLU-IVIG placebo group transition matrix from day 2 to day 3. 127
A.16 FLU-IVIG placebo group transition matrix from day 3 to day 4. 127

A.17 FLU-IVIG placebo group transition matrix from day 4 to day 5. 128
A.18 FLU-IVIG placebo group transition matrix from day 5 to day 6. 128
A.19 FLU-IVIG placebo group transition matrix from day 6 to day 7. 129
A.20 Mathematical descriptions of treatment effects designed to simulate potential IVIG data. 130
A.21 FLU-IVIG placebo group transition matrix from day 0 to day 1 additionally including distributions on both days. 131
A.1 Estimated bias and Monte Carlo standard deviation (MC SD) of parameter estimates for $\boldsymbol{\theta}$ where $n = 500$, the number of time points $K \in \{6, 5\}$, and the compliance effect $\alpha_1 \in \{-1.2, -1.5, -1.8, -2.1\}$. For both compliance and the conditional biomarker, the models fit are either correctly specified with compound symmetric covariance structures, or incorrectly specified where compound symmetric covariance structures are erroneously fit to data that is generated as AR(1). 136

List of Figures

2.1 Comparison of the ordinal endpoint and time-to-event endpoint by follow-up period and $1 - \sum_{i=1}^{k} \bar{q}_i^3$ according to their respective non-centrality parameters which uniquely determine power. 34

2.2 The right-center plot of Figure 2.1 for FLU-IVIG with contours of power added to indicate by how much either the ordinal endpoint or the time-to-event endpoint surpassed the other in power. 35

3.1 Calibration lines (in gray) for $\hat{P}(C_{i6}|B_{i1}, \ldots, B_{i5})$ corresponding to the different sets of parameter estimates where the compliance effect $\alpha_1 \in \{-1.5, -1.8, -2.1\}$. The black lines mark the average calibration line and the 45 degree red lines mark perfect prediction. For both compliance and the conditional biomarker, the models fit are either correctly specified with compound symmetric covariance structures, or incorrectly specified where compound symmetric covariance structures are erroneously fit to data that is generated as AR(1). 61

3.2 Spaghetti plot of the longitudinal biomarker for the time points when participants had data for CENIC-P1S2. Lines are shaded by the probability of compliance at all time points. Lighter lines indicate a probability closer to 0 while darker lines indicate a probability closer to 1. The x-axis refers to the time points when participants had data. For example, a participant with data only at time points 1 and 4 would have a line from 1 to 2. 64

3.3 Averaged estimated prediction probability of $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i5})$ plotted against the average estimated compliance probability of $\hat{P}(C_{i6} = 1|B_{i6})$ by quintile for the 97 participants of CENIC-P1S2 who had anatabine data at the last time point. The 45 degree red line marks perfect prediction and the green lines mark the bounds of the 95% prediction interval. 65

4.1 One plausible directed acyclic graph for CENIC-P2. The time-invariant confounders X_i affect all L_{ij}, C_{ij} and B_{ij}, but not D_{ij}. 76

Chapter 1

Introduction

Clinical trials have been long been viewed as the gold standard for validating new treatments for target populations of interest. The practice of clinical trials can be divided into three phases conducted before, during, and after the trial: trial design, data monitoring, and statistical analysis, respectfully. The execution of each phase depends on the validity of the statistical assumptions made. My dissertation comprises three papers that each address important statistical assumptions of each phase.

For trial design, the choice of endpoint has vast implications for evaluating the treatment effect, and any novel endpoint should be thoroughly examined. In my first paper, I analyze a novel ordinal endpoint that was recently proposed for evaluating new treatments for patients hospitalized by influenza. First, I investigate the statistical assumptions of the model fit to the ordinal endpoint. Second, I compare the ordinal endpoint to other endpoints that can be derived from the same information.

For data monitoring, in some trials participants must self-administer the treatment, and as such may break from their assigned treatment protocol. Methods are thus needed to identify noncompliant participants during the trial. In my second paper, I propose a novel method for detecting participant noncompliance by modeling a longitudinal biomarker.

For statistical analysis, in completed trials where participant noncompliance is thought

to have occurred, the treatment effect from the trial may not match the treatment effect if all participants had complied (i.e., the causal effect). Specifically, confounding may create a difference between the treatment effect and causal effect. In the longitudinal setting, the G-computation algorithm can adjust for such confounding, but typically assumes that both compliance is observed and that the confounders come from known densities. For my third paper, I propose a modified G-computation algorithm estimator in the scenario where both assumptions are violated.

My first paper, given in Chapter 2, comprises two published case studies of the novel ordinal endpoint that was used to evaluate intravenous hyperimmune immunoglobulin (IVIG) as a treatment for hospitalized influenza patients.[1,2] To achieve a feasible sample size and power relative to a binary endpoint (e.g., the proportion of hospitalized patients), the trial to evaluate IVIG, FLU-IVIG (NCTO2287467),[3] used a 6-category ordinal endpoint of patient health status as the primary endpoint. Patients were randomized to either IVIG or placebo and were evaluated 7 days after randomization.

The use of a novel ordinal endpoint inherently invites a number of statistical considerations regarding power. In my first case study of the FLU-IVIG ordinal endpoint, which was published in *Clinical Trials*, we conduct a simulation study to investigate four trial design factors that may affect power under the pre-specified proportional odds model: 1) deviations from the proportional odds assumption which result in the same overall treatment effect as specified in the FLU-IVIG protocol and which result in a diminished overall treatment effect, 2) deviations from the distribution of the placebo group assumed in the FLU-IVIG design, 3) the effect of patient misclassification among the six categories, and 4) the number of categories of the ordinal endpoint. We also consider interactions between the treatment effect (i.e. factor 1) and each other factor.

In my second case study of the FLU-IVIG ordinal endpoint, which was published in *Contemporary Clinical Trials Communications*, we compare the power of the ordinal endpoint to other clinically relevant endpoints that were also considered. In an analytic comparison, we compare the ordinal endpoint to a time-to-event endpoint of time-to-hospital discharge according to the granularity of the ordinal endpoint, follow-up period of the time-to-event endpoint, and the distribution of the placebo group. In a simulation study, we compare the ordinal endpoint to time-to-event, longitudinal, and

binary endpoints according to different expressions of the treatment effect across the follow-up period and the distribution of the placebo group.

The second paper of my dissertation, given in Chapter 3, aims to develop a method that can detect unobserved participant noncompliance in longitudinal studies. Recently, a number of regulatory tobacco trials to evaluate nicotine reduction as a potential regulatory strategy for cigarettes were completed. In these trials, participants were randomized to smoke very low nicotine content (VLNC) cigarettes among other cigarettes. However, participants were not obligated to smoke the VLNC cigarettes and may have instead used commercial cigarettes with normal nicotine content. Participants' total nicotine equivalents, a biomarker of nicotine exposure, changes in response to cigarette consumption and can suggest exposure to normal nicotine content cigarettes.

We propose a longitudinal method that uses biomarker data collected across multiple time points to estimate longitudinal compliance when the true compliance status may be unobserved. Our method can estimate: 1) the probability that a participant complied at a single time point of the trial; 2) the probability that a participant complied at all time points; and 3) the prediction probability that a participant will comply at a future time point. In simulation, we evaluate the accuracy of our method with respect to two factors: 1) the strength of the association between compliance and the biomarker; and 2) the misspecification of the covariance structures of the models fit. Additionally, we apply our method to the data collected in the Center for the Evaluation of Nicotine in Cigarettes Project 1 Study 2 (CENIC-P1S2) trial.

Chapter 4 gives the third paper of my dissertation, which seeks to develop a longitudinal causal estimator when compliance is unknown. In regulatory tobacco trials, the treatment effect may differ from the causal effect due to unobserved noncompliance, as noncompliant participants may systematically differ from compliant participants on some confounding variable. The G-computation algorithm can adjust for such confounding under the assumption that the models of the confounders are correctly specified. Regulatory tobacco trials, however, include confounders like survey data whose densities may substantially deviate from parametric densities.

Our goal is thus to implement a G-computation algorithm estimator that can adjust

for both 1) unobserved noncompliance; and 2) confounders with densities that may be difficult to parametrically specify. For 1), we insert probabilities of compliance in place of the unknown indicator of compliance. The weights are derived from a biomarker associated with compliance. For 2), we implement predictive mean matching, in which the predictive values of the generated data are matched with similar observed values. This constrains the data generation to the densities of the observed values. In simulation, we evaluate how robust our method is to differences in the levels of confounding and sample size. We also apply our method to the results of the Center for the Evaluation of Nicotine in Cigarettes Project 2 (CENIC-P2) trial.

Chapter 2

Analysis and Comparison of an Ordinal Endpoint to Time-to-Event, Longitudinal, and Binary Endpoints for Use in Evaluating Treatments for Severe Influenza Requiring Hospitalization

2.1 Introduction

Influenza causes 226,000 excess hospitalizations and more than 500,000 deaths worldwide.[4] In spite of the large disease burden, no study has definitively demonstrated substantial clinical efficacy of an antiviral drug in hospitalized influenza patients.[5] For this subpopulation, the proportion of patients dying is small, increasing the challenge

to demonstrate treatment effects with all-cause mortality as the sole endpoint. Therefore, the United States Food and Drug Administration (FDA) recommends that the primary endpoint of randomized controlled trials evaluating new treatments include any of the following measures: clinical signs and symptoms, duration of hospitalization, time to normalization of vital signs and oxygenation, requirements for supplemental oxygen or assisted ventilation, and mortality. FDA guidance further states that no single best endpoint has been identified for studying treatments in patients hospitalized by influenza.[5]

Primary endpoints in randomized trials of treatments for hospitalized influenza patients have included continuous measures of virologic activity, time to event outcomes (e.g., time to clinical stability) and binary outcomes (e.g., proportion of patients returning to premorbid status).[6–10] Following the successful completion of a pilot study of intravenous hyperimmune immunoglobulin (IVIG),[11] the International Network for Strategic Initiatives in Global HIV Trials (INSIGHT) initiated a trial of IVIG (FLU-IVIG) to evaluate its efficacy in patients hospitalized with influenza (NCTO2287467).[3] To conduct a study with a feasible sample size and to improve the likelihood of demonstrating benefit relative to a binary outcome, a novel ordinal outcome of patient status serves as the primary endpoint of FLU-IVIG. The ordinal endpoint constructs categories of various outcome assessments ranked in order of patient status. A recent report by a working group from the U.S. Department of Health and Human Services advocates the use of an ordinal endpoint like the one in FLU-IVIG.[12] Two other randomized trials are using an ordinal scale as either the primary or a secondary endpoint (NCT02572817 & NCT03376321).[13,14]

Many factors must be considered when selecting an endpoint for a trial, including clinical relevance, potential bias in ascertainment, and statistical power. In our first case study of the FLU-IVIG ordinal endpoint, we consider the impact of four factors on power under the pre-specified proportional odds model: 1) deviations from the proportional odds assumption which result in the same overall treatment effect as specified in the FLU-IVIG protocol and which result in a diminished overall treatment effect; 2) deviations from the distribution of the placebo group that researchers expect to observe in the FLU-IVIG protocol; 3) the effect of patient misclassification among the 6 categories; and 4)

the number of categories of the ordinal endpoint. In addition to examining these factors separately, we also consider the effect of interactions between the treatment effect (i.e., factor 1) and each of the other factors. In our second case study, we compare the power of the ordinal endpoint to other clinically relevant endpoints that were also considered for FLU-IVIG, including time-to-event, longitudinal, and binary endpoints. We draw comparisons to other endpoints analytically and by simulation.

2.2 Methods

2.2.1 The INSIGHT FLU-IVIG Trial

The FLU-IVIG study is a multicenter, double-blind randomized trial comparing IVIG versus placebo in hospitalized patients with locally confirmed influenza A or B who have a National Early Warning Score of two or higher.[15] Patients receive IVIG or placebo in addition to standard of care treatment. The primary endpoint for FLU-IVIG is the following 6-category ordinal outcome evaluated 7 days after randomization:

1. death;
2. intensive care unit hospitalization (In ICU);
3. non-ICU hospitalization, requiring supplemental oxygen;
4. non-ICU hospitalization, not requiring supplemental oxygen;
5. discharged from the hospital, but unable to resume normal activities;
6. discharged from the hospital with resumption of normal activities.

The categories were defined to delineate clinically relevant change in patient health status due to IVIG. Day 7 was chosen for evaluation because a pilot study had established that differences in influenza antibody titer levels between IVIG and placebo were highest in the first few days following randomization.[11] The ordinal endpoint was chosen over a binary endpoint (e.g., proportion of patients discharged by day 7) because it was thought to provide more power and clinical information about patient recovery. Note that the ordinal endpoint ignores patient health trajectory across follow-up (e.g., hospitalized patients are counted as equal to re-admitted patients), only evaluating status

on day 7.

The FLU-IVIG protocol specifies that the estimated odds ratio from fitting a proportional odds cumulative logistic model will be used to evaluate the effect of IVIG. Under the proportional odds assumption of the model, the odds ratios for any of the five better versus worse divisions of the ordinal endpoint (e.g., discharged versus hospitalized or dead) are constant. That is, the effect of IVIG is no more likely to benefit patients in one category versus another. Even if the proportional odds assumption is violated, the estimated odds ratio is still a valid measure of treatment efficacy for hypothesis testing and can be interpreted as the average shift across the ordinal endpoint due to IVIG, or alternatively as the odds of having a more favorable outcome due to IVIG compared to placebo. Note that the estimated odds ratio is not the arithmetic mean of the odds ratio for every possible binary division of the ordinal endpoint but is instead a nonlinear function of the probabilities of each category of the ordinal endpoint in the placebo and IVIG groups. The score test of the odds ratio is equivalent to the Wilcoxon rank-sum test.[16]

With consideration for the anticipated distribution of the ordinal endpoint in the placebo group, the FLU-IVIG trial had a sample size of 320 patients to detect an odds ratio of 1.77 (or a log odds ratio of 0.57) with 80% power at the 0.05 (two-sided) level of significance. An odds ratio (log odds ratio) greater than 1 (0) indicates a more favorable outcome due to IVIG. For reference, Supplementary Table A.1 gives the power for detecting different log odds ratios with a sample size of 320 patients. Supplementary Table A.2 gives the power for detecting a significant treatment effect under each possible way of dividing the ordinal endpoint into a binary endpoint. Power calculations were derived from formulas provided by Whitehead.[17] The power from each binary endpoint is substantially less than the power from the ordinal endpoint.

To estimate the distribution of the ordinal endpoint in the placebo group at day 7, the FLU-IVIG trial used data from a cohort study of patients hospitalized with influenza at many of the same sites participating in FLU-IVIG.[18,19] As the cohort study is still in progress, we derived more recent placebo group distributions for the first and second case studies updated as of September 1st, 2015 and September 1st, 2016, respectively. For the category percentages of the placebo group specified in the FLU-IVIG protocol,

see the second footnote of Table 2.4.

2.2.2 Case Study 1: Proportional Odds Model Power

2.2.2.1 Simulation Study Design

We first derived different distributions of the ordinal endpoint in the placebo and IVIG groups under different scenarios of each factor which we describe below. For each scenario, we ran 10,000 simulations of the clinical trial assuming that 320 patients were sampled from the corresponding placebo and IVIG group distributions. For each simulated trial, we analyzed the data assuming a proportional odds cumulative logistic model and computed a Wald test statistic for the treatment effect. The empirical power is the proportion of the 10,000 simulations for which the Wald test statistic was significant. With this approach, estimates of power do not require any large sample approximations.

The reference level of the factors in our simulation experiment corresponds to the assumptions used in the sample size calculation for the FLU-IVIG design; that is, the proportional odds assumption holds, the distribution of the placebo group is determined from the cohort study, no misclassification of patients among the categories of the ordinal endpoint occurs, and the full 6-level ordinal endpoint is used.

2.2.2.2 Factor 1: Treatment Effect

We first sought to derive distributions of the IVIG group that deviated from proportional odds while maintaining the same overall treatment effect as specified in the FLU-IVIG protocol. By overall treatment effect, we mean the average (across repeated experimentation) estimated log odds ratio for the effect of IVIG relative to placebo from fitting the proportional odds model to the data. We refer to this as the average log odds ratio but note that this is not the arithmetic mean of the log odds ratio for every possible binary division of the ordinal endpoint and is a nonlinear function of the probabilities of each category in the ordinal endpoint in the placebo and IVIG groups.

For large samples, the average log odds ratio of a misspecified proportional odds model is the value for which the expected score function equals zero. Therefore, we can

constrain the distribution of the IVIG group such that the average log odds ratio is maintained across deviations from proportional odds (see Appendix A.1 for a derivation). We created a novel algorithm which, given the desired average log odds ratio, the distribution of the ordinal endpoint in the placebo group and the proportions of observations in all but two categories of the treatment group, returns the proportions in the final two categories of the treatment group to maintain the desired average log odds ratio. Code to implement our algorithm in the programming language R is available as a GitHub repository (https://github.com/RPeterson4/Supplementary-Code-for-Evaluating-the-Ordinal-Endpoint-for-FLU-IVIG).

We considered three treatment effect scenarios (T1–T3 below) that deviate from proportional odds while maintaining an average log odds ratio of 0.57 assuming the other three factors are not altered. The deviation from proportional odds is strong enough in each of these scenarios to yield, on average, a significant p-value for the test of the proportional odds assumption (at the 0.05 level) across the samples. We also considered two treatment effect scenarios (T4 and T5) with a log odds ratio of 0.57 across a subset of all possible binary divisions of the categories of the ordinal endpoint but zero elsewhere. In these scenarios, the overall treatment effect is diminished. The six treatment effect scenarios are:

- T0: Proportional odds is satisfied and the average log odds ratio is 0.57 (FLU-IVIG design assumption).
- T1: The treatment effect constantly weakens across the ordinal endpoint. The log odds ratio is 2.6 between the binary outcome of alive and dead patients, and then constantly decreases by 0.6 with each successive binary division of the ordinal endpoint (e.g., the log odds ratio is 2.0 for Hospitalized, not in ICU, on oxygen or better versus death or in ICU).
- T2: The treatment effect is constant and positive across the most severe categories of the ordinal endpoint. Specifically, the log odds ratio of 1.16 for the first four binary divisions of the ordinal endpoint (ordering the scale from most severe outcome to least severe). There is no treatment effect for the last binary division (Discharged, back to normal activities or worse versus Discharged, not back to

normal activities).

- T3: The treatment only benefits patients in the discharged categories. That is, the log odds ratio is 1.16 for the last binary division and 0 for all other binary divisions.
- T4: The log odds ratio is 0.57 for the first four binary divisions and 0 for the last binary division.
- T5: The log odds ratio is 0.57 for the last binary division and 0 for the first four binary divisions.

2.2.2.3 Factor 2: Distribution of the Placebo Group

To systematically alter the distribution of the placebo group, note that the cumulative log odds of being in a more versus less severe category for each possible binary split of the ordinal endpoint (see Supplementary Table A.3) uniquely determines the placebo group distribution. To derive different distributions of the placebo group, we added or subtracted a constant from each of the cumulative log odds of being in a more versus less severe category from the placebo group design estimate (see Appendix A.2 for a derivation). Adding (subtracting) a constant increases the proportion of patients with more (less) severe outcomes of the ordinal endpoint. Note that 62.9% of subjects are in the discharged categories of the ordinal endpoint for the placebo group design estimate. Therefore, having more (fewer) patients in more severe categories will yield a less (more) skewed distribution. The five distributions of the placebo group are:

- P0: The placebo group distribution for the FLU-IVIG design.
- P1: Add 0.5 to the cumulative log odds of P0 (less skewed distribution).
- P2: Add 1 to the cumulative log odds of P0 (less skewed distribution).
- P3: Subtract 0.5 from the cumulative log odds of P0 (more skewed distribution).
- P4: Subtract 1 from the cumulative log odds of P0 (more skewed distribution).

2.2.2.4 Factor 3: Misclassification

For our purposes, we studied misclassification among adjacent pairs of categories that may be difficult to distinguish between for significant numbers of patients. This misclassification may result from a combination of the subjective nature of the categories, inconsistent clinician judgment, and patients' memory of their recovery. To study the effect of misclassification, we considered scenarios investigated by Whitehead who supposed 20% misclassification between two pairs of categories.[17] Whitehead represented misclassification by exchanging certain percentages of patients between categories that could be misclassified (see Appendix A.3 for a derivation).

Here, we assumed that 20% and 40% of patients in the non-ICU hospitalized categories and the discharged categories could be misclassified. We chose the non-ICU hospitalized categories because use of oxygen during the day can be variable, and the discharged categories for depending on the patient's memory of when they resumed normal activities. We assumed the misclassification rate to be constant across both randomized groups because the study is double-blind (i.e., nondifferential misclassification). We also considered scenarios in which 20% misclassification affected either the non-ICU hospitalized categories or the discharged categories but not the other. The type I error rate does not change under the nondifferential misclassification we assumed. The five levels of misclassification are:

- M0: No misclassification (FLU-IVIG design assumption).
- M1: 20% misclassification between the non-ICU hospitalized categories and the discharged categories.
- M2: 40% misclassification between the non-ICU hospitalized categories and the discharged categories.
- M3: 20% misclassification between the non-ICU hospitalized categories.
- M4: 20% misclassification between the discharged categories

2.2.2.5 Factor 4: Number of Categories

Misclassification between adjacent pairs of categories can be eliminated by collapsing each into a single category. Thus, we examined collapsing the non-ICU hospitalized categories and the discharged categories. Furthermore, the discharged categories contain the largest percentage of patients in each scenario on average, implying that collapsing them may have an outsize effect on power. Conversely, we collapsed the four most severe categories because they contain the smallest percentage of patients. We also collapsed the ordinal endpoint into a binary hospitalized or dead versus discharged endpoint, which is a clinically relevant cut-point. The six levels of collapsing categories are:

- C0: The full 6-category ordinal endpoint (FLU-IVIG design assumption).
- C1: Collapse the non-ICU hospitalized categories and the discharged categories.
- C2: Collapse the non-ICU hospitalized categories.
- C3: Collapse the discharged categories.
- C4: Collapse the hospitalization or death categories.
- C5: Collapse the hospitalization or death categories and the discharged categories to make for a binary endpoint.

2.2.2.6 Interactions

As detection of the treatment effect is of primary interest for the FLU-IVIG trial, we explored the effect on power if the treatment effect (factor 1) and the placebo group distribution, misclassification, or number of categories also deviated from the levels assumed in the design of FLU-IVIG. This yielded three groups of two-way interactions: 1) treatment effect scenarios and distributions of the placebo group; 2) treatment effect scenarios and levels of misclassification; and 3) treatment effect scenarios and number of categories. Due to the interacting factors, the overall treatment effect may differ from the 0.57 log odds ratio specified in the FLU-IVIG protocol.

2.2.3 Case Study 2: Comparison to Other Types of Endpoints

2.2.3.1 Analytic Comparison of Ordinal Endpoint to Time-to-Event Endpoint

We analytically compare the power of an ordinal endpoint to a time-to-event endpoint both derived from the same information. For concreteness, we refer to the time-to-event endpoint as time-to-hospital discharge, where deaths are censored at the end of follow-up and first hospital discharge counts as the event. The power for most hypothesis tests at the 0.05 (two-sided) level of significance is approximately equal to:

$$\text{Power} \approx \Phi(-1.96 + c),$$

where Φ denotes the cumulative density function of the standard normal distribution and c is the non-centrality parameter which varies based on the type of endpoint chosen and the data generating mechanism.

We assume that time-to-hospital discharge follows an accelerated failure time (AFT) model with an exponential distribution and constant hazard ratio between treatment groups.[20] The AFT model assumes that the treatment proportionally increases or decreases the quantiles (e.g., median) of the duration of hospitalization. The non-centrality parameter for the AFT model, c_a, is given by (see Appendix A.4 for derivation):

$$c_a = \sqrt{n} * \log\left[\frac{\log(1-p_{1t})}{\log(1-p_{0t})}\right] * \left[\frac{1}{\sqrt{2}}\left(\frac{1}{p_{0t}} + \frac{1}{p_{1t}}\right)^{-1/2}\right],$$

where n denotes the total sample size assuming equal randomization to both groups, log denotes the natural logarithm, and p_{it} denotes the probability of discharge by the end of follow-up (i.e., categories 5 and 6 of the ordinal endpoint combined) for the ith randomized group (0 denotes placebo, 1 denotes treatment) assuming a follow-up period of t days. We assume the same follow-up period for all patients because in influenza trials which motivate this work, follow-up is typically short and endpoints such as survival and hospital discharge are frequently assessed with very little missing data. Because the FLU-IVIG trial protocol expects minimal missing data on day 7 and

time-to-hospital discharge is easily recorded, we assumed no missing data in both the analytic and simulation settings.

We assume that the ordinal endpoint follows a proportional odds model, as specified in the FLU-IVIG trial. For the proportional odds model, the non-centrality parameter, c_p, under the proportional odds assumption on day 7 is provided by Whitehead[17] and is approximately equal to:

$$c_p \approx \sqrt{n} * \log\left[\frac{p_{17}(1-p_{07})}{p_{07}(1-p_{17})}\right] * \sqrt{\frac{(1-\sum_{i=1}^{k}\bar{q}_i^3)}{12}},$$

where $\bar{q}_i^3$ denotes the average categorical probability between both randomized groups for the ith category of an ordinal endpoint with k categories, and the fraction $\frac{p_{17}(1-p_{07})}{p_{07}(1-p_{17})} = 1.77$ from FLU-IVIG. The term $1-\sum_{i=1}^{k}\bar{q}_i^3$ measures the granularity of the ordinal endpoint with larger values indicating an ordinal endpoint whose category proportions are more evenly spread. For example, assume that we have an ordinal endpoint with three categories. If patients in the placebo group are evenly spread about the categories with proportions (1/3, 1/3, 1/3), and the distribution of patients in the treatment group meets proportional odds (odds ratio = 1.77 from FLU-IVIG) with category proportions (0.22, 0.31, 0.47), then $1-\sum_{i=1}^{k}\bar{q}_i^3 = 0.88$. Conversely, if patients in the placebo group tend to fall towards the last category with (1/10, 1/10, 4/5), and the treatment group again meets proportional odds (odds ratio = 1.77) with (0.06, 0.06, 0.88), then $1-\sum_{i=1}^{k}\bar{q}_i^3 = 0.41$.

Because both non-centrality parameters are proportional to $\sqrt{n}$, this implies that relative comparisons of the two models according to power do not depend on sample size for sufficiently large n. We compared the power of both endpoints according to different values of $1-\sum_{i=1}^{k}\bar{q}_i^3$ for the ordinal endpoint and different follow-up periods for the time-to-event endpoint. We varied $1-\sum_{i=1}^{k}\bar{q}_i^3$ as it measures the granularity of the ordinal endpoint, which is what a trial designer would want to maximize (e.g., by splitting or collapsing categories to be more evenly spread) to raise power. We fixed follow-up at day 7 for the ordinal endpoint in accordance with the FLU-IVIG trial and because in a trial with longer follow-up than 7 days, the two hospital discharge categories may be more difficult to ascertain. We extended the follow-up period of the time-to-event

endpoint from 7 to 14 days to allow for more time for the event of hospital discharge to occur. Thus, we aimed to make a comparison between the ordinal and time-to-event endpoints under follow-up periods considered optimal for each. For both endpoints, we additionally varied the probability of discharge by day 7 (p_{07}) as a measure of the underlying risk of the population enrolled in the study. All other factors (e.g., treatment effect odds ratio of 1.77) were held constant. Note that we assume that hospital discharge constitutes at least one category of the ordinal endpoint; thus, values of $1 - \sum_{i=1}^{k} \bar{q}_i^3$ below $1 - \left(\frac{p_{07}+p_{17}}{2}\right)^3$ are not possible.

2.2.3.2 Simulation Comparison of Ordinal Endpoint to Other Types of Endpoints

Closed form expressions for (approximate) power only exist for a small number of endpoints. Furthermore, these expressions typically assume that the analysis model is consistent with the data generating mechanism. To broaden the scope of our study, we used simulation to compare the ordinal endpoint to other types of endpoints for data that does not follow a constant hazard ratio between treatment groups. We considered six different endpoints each evaluated on day 7, the pre-specified time point for FLU-IVIG:

- E1: Proportion of patients hospitalized or dead on day 7.
- E2: Proportion of patients moving to less severe categories from day 0 to day 7.
- E3: Winners versus losers between IVIG and placebo on day 7.
- E4: Day 7 ordinal endpoint.
- E5: Longitudinal measures of the ordinal endpoint over days 1–7 of follow-up.
- E6: Time-to-first hospital discharge.

We considered seven models to fit to these six endpoints. We fitted a simple logistic regression model to both E1 and E2, but given that E2 compares patient status for two different time points, we refer to its fitted model as the sliding dichotomy.[21] The endpoint E3 considers all possible comparisons of patients in the IVIG group to those in the placebo group according to their given ordinal endpoint category on day 7.[22]

The summary measure for this endpoint, called the win ratio, calculates the number of comparisons of greater health status for IVIG relative to placebo (wins) divided by the number of comparisons of worse health status (losses). Without any stratification, the win ratio performs similar to the well-known nonparametric Wilcoxon rank-sum test.[23]

We fitted the proportional odds model to E4 and a longitudinal ordinal outcome model to E5 using generalized estimating equations assuming an independent working correlation matrix.[24] The longitudinal ordinal outcome model includes a term for treatment group, day of assessment (treated as a continuous variable), and day by treatment group interaction to model the distributions of the ordinal endpoint in both randomized groups over days 1–7 of follow-up. The day by treatment group interaction captures the treatment effect as the average multiplicative change in the odds ratio from the proportional odds model across days. For example, if the treatment effect proportionally increases by day up to an odds ratio of 1.77 on day 7, then the coefficient for the day by treatment group interaction would be 1.085 such that $1.085^7 = 1.77$.

We fitted both the Cox proportional hazards model and the AFT model to E6.[25] Across the follow-up period, the Cox model calculates the hazard ratio of hospital discharge between groups. To allow for more flexibility in modeling time-to-hospital discharge for the AFT model, we assumed a Weibull distribution in addition to an exponential distribution. All seven models with their corresponding six endpoints are displayed in Table 2.1. In addition to models which only include the treatment effect, we considered models which adjusted for baseline health status. Baseline health status was defined as which of the three categories the patient was in at enrollment (i.e., ICU; hospitalized, not in ICU, on oxygen; and hospitalized, not in ICU, not on oxygen). For the win ratio, we stratified comparisons between randomized groups according to baseline health status.

To generate longitudinal data comparable to the data that was expected in the FLU-IVIG trial, we simulated from the distributions of the ordinal endpoint in the cohort study at randomization and across the follow-up period (see Table 2.2). Table 2.3 gives the category percentages of the ordinal endpoint at day 7 for the placebo and IVIG groups under proportional odds. Table 2.3 additionally includes the five odds ratios

for the five better versus worse divisions of the ordinal endpoint, with a demonstration below of how to derive the five odds ratios.

For the simulation, we first randomly sampled 320 patients from the day 0 distribution of the cohort study to create day 0 data for both groups (i.e., no treatment effect at randomization). We generated longitudinal data using a discrete-time Markov model whereby the day-to-day transition probabilities for the placebo group were estimated from the cohort study. For the treatment group, we varied on which days and for which groups of patients (e.g., hospitalized or dead versus discharged) the transition probabilities differed from the placebo group. Using a method from Peterson et al.,[1] data were generated such that the average odds ratio on day 7 for the ordinal endpoint approximated 1.77, the pre-specified value of FLU-IVIG. Additionally, we varied the transition probabilities for the placebo group to simulate different target populations. Unlike the analytic setting, the proportional odds assumption may not hold on day 7 and time-to-first hospital discharge may not be exponentially distributed. Appendix A.5 explains the data generating process in greater detail.

We ran 10,000 simulations of the clinical trial for each treatment effect and placebo group combination. For each simulated trial, we fitted the seven models to the corresponding six endpoints and computed the corresponding Wald test statistics for the treatment effect. For each model, the empirical power is the proportion of the 10,000 simulations for which the Wald test statistic meets significance. Code to run our simulation in the R programming language can be downloaded as a GitHub repository (https://github.com/RPeterson4/Comparative_FLU_IVIG_Code).

Table 2.1: Models fitted to endpoints for the simulated data.

Model	Endpoint	Endpoint Variable Type	Coefficient Interpretation
Simple Logistic	E1: Proportion of patients hospitalized or dead on day 7	Binary	Odds ratio of discharged from the hospital versus not discharged on day 7
Sliding Dichotomy	E2: Proportion of patients moving to less severe categories from day 0 to day 7	Binary	Odds ratio of moving versus not moving to a less severe category from day 0 to day 7
Win Ratio	E3: Winners versus losers between IVIG and placebo on day 7	Binary	For all possible comparisons of patients in IVIG versus placebo, the number of IVIG winners divided by the number of IVIG losers
Proportional Odds[a]	E4: Day 7 ordinal endpoint	Ordinal	Average odds ratio of being in a less versus more severe category on day 7
Longitudinal Ordinal Outcome	E5: Distribution of the ordinal endpoint over the seven days of follow-up	Ordinal Longitudinal	Average multiplicative increase in the odds ratio of being in a less versus more severe category across the follow-up period
Cox Proportional Hazards	E6: Number of days to first hospital discharge	Time-to-Event	Hazard ratio of time-to-hospital discharge
Accelerated Failure Time (Exponential[a] and Weibull distributions)	E6: Number of days to first hospital discharge	Time-to-Event	Reduction in quantiles of time-to-hospital discharge

[a]Model was used in the analytic setting, displayed in Figures 2.1 and 2.1.

Table 2.2: FLU-IVIG placebo group distribution of the ordinal endpoint at randomization and across the 7 day follow-up period according to the cohort study.[18,19]

Day of Evaluation	Death	ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
Day 0 (%)[a]	0	7.6	46.1	46.3	0	0
Day 1 (%)	0.2	6.4	40.9	43.1	8.1	1.2
Day 2 (%)	0.2	6.7	33.7	37.9	18.5	3.0
Day 3 (%)	0.2	6.7	28.8	33.0	23.2	8.1
Day 4 (%)	0.5	6.4	21.2	26.8	31.8	13.3
Day 5 (%)	0.5	5.9	19.2	21.9	33.3	19.2
Day 6 (%)	1.0	4.9	17.7	18.5	34.7	23.2
Day 7 (%)	1.0	4.9	16.3	14.5	36.2	27.1

[a](%) percentage of patients in the placebo group for the given ordinal endpoint category.

Table 2.3: Distributions of the FLU-IVIG placebo group and IVIG group on day 7 of follow-up.

	Death	ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
% Placebo[a]	1.0	4.9	16.3	14.5	36.2	27.1
% IVIG[b]	0.6	2.9	10.4	10.8	35.6	39.7
Odds ratio[c]		1.77	1.77	1.77	1.77	1.77

The FLU-IVIG placebo group distribution was estimated from the cohort study.[18,19]

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]Odds ratio: The odds of having the given ordinal endpoint category or less severe versus more severe between the IVIG and placebo groups.

Derivation of the five odds ratios

Under the proportional odds assumption of the model, the odds ratio of having a given category or less severe versus more severe for any binary split of the ordinal endpoint is fixed at 1.77. For example, for the binary split of discharged, back to normal activities versus discharged, not back to normal activities or worse, the odds ratio can be derived as:

$$\frac{39.7/(1-39.7)}{27.1/(1-27.1)} = 1.77.$$

Similarly, for the binary split of discharged versus not discharged or dead, the odds ratio can be derived as:

$$\frac{(39.7+35.6)(1-39.7-35.6)}{(27.1+36.2)(1-27.1-36.2)} = 1.77.$$

2.3 Results

2.3.1 Case Study 1: Proportional Odds Model Power

2.3.1.1 Main Effects

Provided that the average log odds ratio was maintained, treatment effect scenarios that violated proportional odds only marginally reduced power (see Table 2.4). For example, under treatment effect scenario T2 in which the treatment benefit is only evident over the most severe categories of the ordinal endpoint (hospitalization or death categories), power declined from 80% to 77.3%. However, both scenarios in which the log odds ratio was 0.57 for some binary divisions but 0 for the rest greatly reduced power, mainly due to the decline in the average log odds ratio from 0.57 (0.31 and 0.25 under T4 and T5, respectively).

Changes in the distribution of the placebo group from the FLU-IVIG design led to moderate differences in power. Less skewed placebo group distributions yielded slightly higher power, from 80% to 80.9% and 81.9% in scenarios P1 and P2, respectively (see Table 2.5). Conversely, distributions of the placebo group which were more skewed led to modest declines in power, with the most skewed distribution returning the largest loss of power from 80% to 73.8%.

Table 2.6 shows that misclassification among the categories always reduced power by lowering the average log odds ratio. Scenarios in which there was greater misclassification, the misclassification involved more categories, or the misclassification was between categories containing many patients decreased power the most. For example, limiting the misclassification to the discharged categories, which comprise 62.9% of patients in the distribution of the placebo group, reduced power from 80% to 70.1%. Expanding the 20% misclassification to include the non-ICU hospitalized categories, which together contain 30.6% of patients, only additionally reduced power from 70.1% to 69.7%.

Reducing the number of categories always lowered power (see Table 2.7). Generally, power declined more when multiple categories or categories with many patients were combined. For example, collapsing the discharged categories reduced power from 80% to 65.6%, while having a binary hospitalized or dead versus discharged endpoint reduced

power from 80% to 63.9%. Collapsing the non-ICU hospitalized categories or the four most severe categories did not substantially reduce power, mainly due to the small percentage of patients in those categories.

Table 2.4: Main effect of the treatment effect on power (factor 1).

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
T0: Proportional odds holds (FLU-IVIG design assumption)	% Placebo[a]	1.2	5.3	16.2	14.4	36.4	26.5	Power (%)[d]	80.0
	% IVIG[b]	0.7	3.1	10.5	10.8	36.0	39.0		
	logOR[c]	0.57	0.57	0.57	0.57	0.57	NA	Avg. logOR[e]	0.57
T1: Treatment effect constantly weakens	% Placebo	1.2	5.3	16.2	14.4	36.4	26.5	Power (%)	79.1
	% IVIG	0.1	0.8	5.8	14.2	48.5	30.5		
	logOR	2.60	2.00	1.40	0.80	0.20	NA	Avg. logOR	0.57
T2: Treatment effect limited to the hospitalization or death categories	% Placebo	1.2	5.3	16.2	14.4	36.4	26.5	Power (%)	77.3
	% IVIG	0.4	1.7	6.3	7.2	57.9	26.5		
	logOR	1.16	1.16	1.16	1.16	0	NA	Avg. logOR	0.57
T3: Treatment effect limited to the discharged categories	% Placebo	1.2	5.3	16.2	14.4	36.4	26.5	Power (%)	78.7
	% IVIG	1.2	5.3	16.2	14.4	9.3	53.6		
	logOR	0	0	0	0	1.16	NA	Avg. logOR	0.57
T4: Smaller treatment effect limited to the hospitalization or death categories	% Placebo	1.2	5.3	16.2	14.4	36.4	26.5	Power (%)	33.1
	% IVIG	0.7	3.1	10.5	10.8	48.5	26.5		
	logOR	0.57	0.57	0.57	0.57	0	NA	Avg. logOR	0.31
T5: Smaller treatment effect limited to the discharged categories	% Placebo	1.2	5.3	16.2	14.4	36.4	26.5	Power (%)	23.9
	% IVIG	1.2	5.3	16.2	14.4	23.9	39.0		
	logOR	0	0	0	0	0.57	NA	Avg. logOR	0.25

Under factor 1, the treatment effect deviates from proportional odds assuming the distribution of the placebo group is as specified in the FLU-IVIG design, no misclassification, and the full 6-level ordinal endpoint is used.

The placebo group distribution specified in the FLU-IVIG design has been updated from the FLU-IVIG protocol using data from the cohort study as of 2015. For reference, the category percentages in the FLU-IVIG protocol are 1.8, 3.6, 15.6, 14.1, 39.0, and 25.8% for Death through Discharged, back to normal activities categories, respectively.

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Table 2.5: Main effect of the distribution of the placebo group on power (factor 2).

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
P1: Less skewed placebo group distribution	% Placebo[a]	2.0	8.3	22.4	16.6	32.7	18.0	Power (%)[d]	80.9
	% IVIG[b]	1.1	5.0	15.4	13.9	36.6	27.9		
	logOR[c]	0.57	0.57	0.57	0.57	0.57	NA	Avg. logOR[e]	0.57
P2: Even less skewed placebo group distribution	% Placebo	3.2	12.7	28.5	17.1	26.7	11.7	Power (%)	81.9
	% IVIG	1.8	7.8	21.5	16.4	33.4	19.0		
	logOR	0.57	0.57	0.57	0.57	0.57	NA	Avg. logOR	0.57
P3: More skewed placebo group distribution	% Placebo	0.7	3.3	11.1	11.2	36.3	37.3	Power (%)	78.6
	% IVIG	0.4	1.9	6.8	7.7	31.9	51.3		
	logOR	0.57	0.57	0.57	0.57	0.57	NA	Avg. logOR	0.57
P4: Even more skewed placebo group distribution	% Placebo	0.4	2.0	7.3	8.1	32.6	49.5	Power (%)	73.8
	% IVIG	0.3	1.2	4.3	5.2	25.6	63.5		
	logOR	0.57	0.57	0.57	0.57	0.57	NA	Avg. logOR	0.57

Under factor 2, the placebo group distribution deviates from that specified in the FLU-IVIG design assuming the treatment effect follows proportional odds, no misclassification, and the full 6-level ordinal outcome is used.

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Table 2.6: Main effect of misclassification on power (factor 3).

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
M1: 20% misclassification between the non-ICU hospitalized categories and discharged categories	% Placebo[a]	1.2	5.3	15.8	14.8	34.4	28.5	Power (%)[d]	69.7
	% IVIG[b]	0.7	3.1	10.6	10.7	36.6	38.4		
	logOR[c]	0.57	0.57	0.54	0.57	0.45	NA	Avg. logOR[e]	0.50
M1: 40% misclassification between the non-ICU hospitalized categories and discharged categories	% Placebo	1.2	5.3	15.5	15.1	32.4	30.5	Power (%)	57.7
	% IVIG	0.7	3.1	10.6	10.7	37.2	37.8		
	logOR	0.57	0.57	0.52	0.57	0.33	NA	Avg. logOR	0.44
M3: 20% misclassification between the non-ICU hospitalized categories	% Placebo	1.2	5.3	15.8	14.8	36.4	26.5	Power (%)	79.6
	% IVIG	0.7	3.1	10.6	10.7	36.0	39.0		
	logOR	0.57	0.57	0.54	0.57	0.57	NA	Avg. logOR	0.57
M4: 20% misclassification between the discharged categories	% Placebo	1.2	5.3	16.2	14.4	34.4	28.5	Power (%)	70.1
	% IVIG	0.7	3.1	10.5	10.8	36.6	38.4		
	logOR	0.57	0.57	0.57	0.57	0.45	NA	Avg. logOR	0.51

Under factor 3, patients are misclassified between the non-ICU hospitalized categories and the discharged categories assuming the treatment effect (without misclassification) follows proportional odds, the distribution of the placebo group is as specified in the FLU-IVIG design, and the full 6-level ordinal endpoint is used.

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Table 2.7: Main effect of the number of categories on power (factor 4).

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
C1: Collapse the non-ICU hospitalized categories and discharged categories	% Placebo[a]	1.2	5.3	30.6		62.9		Power (%)[d]	65.1
	% IVIG[b]	0.7	3.1	21.2		75.0			
	logOR[c]	0.57	0.57	0.57		NA		Avg. logOR[e]	0.57
C2: Collapse the non-ICU hospitalized categories	% Placebo	1.2	5.3	30.6		36.4	26.5	Power (%)	79.7
	% IVIG	0.7	3.1	21.2		36.0	39.0		
	logOR	0.57	0.57	0.57		0.57	NA	Avg. logOR	0.57
C3: Collapse the discharged categories	% Placebo	1.2	5.3	16.2	14.4	62.9		Power (%)	65.6
	% IVIG	0.7	3.1	10.5	10.8	75.0			
	logOR	0.57	0.57	0.57	0.57	NA		Avg. logOR	0.57
C4: Collapse the four most severe categories	% Placebo	37.1				36.4	26.5	Power (%)	78.8
	% IVIG	25.0				36.0	39.0		
	logOR	0.57				0.57	NA	Avg. logOR	0.57
C5: Collapse the four most severe categories and discharged categories (binary endpoint)	% Placebo	37.1				62.9		Power (%)	63.9
	% IVIG	25.0				75.0			
	logOR	0.57				NA		Avg. logOR	0.57

Under factor 4, categories of the ordinal endpoint are collapsed assuming that the treatment effect follows proportional odds, the distribution of the placebo group is as specified in the FLU-IVIG design, and no misclassification.

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

2.3.1.2 Interactions

The power for all possible combinations of the interacting factors for the three groups of two-way interactions considered is given in Supplementary Tables A.4 to A.6. From these, we selected a subset from each group of interactions for further investigation based on their effect on power and clinical relevance. We present our findings in Tables 2.8 to 2.10.

Many of the interactions between the treatment effect and each of the other factors were qualitative, that is the direction of the main effect on power changed when an additional factor was altered. For example, Table 2.8 demonstrates that the effect of deviations from proportional odds on power may change with different placebo group distributions. Power substantially increased (decreased) when treatment effect scenarios were paired with distributions of the placebo group that had more (fewer) patients in categories influenced by the treatment. For example, under treatment effect scenario T2 in which the treatment benefit is only evident for the hospitalization or death categories, having a less skewed placebo group distribution (i.e., more hospitalized or dead patients) raised power from 77.3% to 99.7% (see Tables 2.4 and 2.8). On the other hand, having a more skewed placebo group distribution reduced power from 77.3% to 16.8%.

Similarly, Table 2.9 shows that misclassification may not reduce power when coupled with deviations from proportional odds. In some cases, it may even raise power. Under treatment effect scenario T2, 20% and 40% misclassification between the non-ICU hospitalized categories and the discharged categories raised power from 77.3% to 86.2% and 92.7%, respectively (see Tables 2.4 and 2.9). This is likely because without misclassification the scenario assumes no treatment effect for patients discharged from the hospital. Misclassification, though, evens the proportions between patients who have and have not resumed normal activities, creating the illusion that the treatment has shifted patients into resuming normal activities. Consequently, the log odds ratio for Not Normal or worse versus Normal increased from 0 to raise the average log odds ratio and power.

Additionally, Table 2.10 demonstrates that for scenarios in which the treatment effect

was absent across a range of the ordinal endpoint, collapsing the corresponding categories raised power by increasing the average log odds ratio. Under treatment effect scenarios T1, T2, and T4 (scenarios in which the intervention primarily shows benefit for the hospitalization or death categories), collapsing the discharged categories increased power from 79.1% to 95.8%, 77.3% to 99.5%, and 33.1% to 65.6%, respectively (see Tables 2.4 and 2.10). Conversely, collapsing categories over ranges of the ordinal endpoint with a discernible treatment effect reduced power. For example, under treatment effect scenarios T1, T2, and T4, collapsing the four most severe categories reduced power from 79.1% to 67.0%, 77.3% to 76.0%, and 33.1% to 32.0%, respectively.

Comparing Tables 2.5 and 2.6, collapsing the non-ICU hospitalized categories and the discharged categories to eliminate potential misclassification yielded greater power than using the 6-level ordinal endpoint when misclassification between both pairs of categories was 40% (65.1% versus 57.7%). When misclassification was limited to the non-ICU hospitalized categories at the 20% level, collapsing those categories yielded approximately equal power compared to using the 6-level ordinal endpoint. Limiting 20% misclassification to the discharged categories generated greater power for the 6-level endpoint relative to collapsing those categories (70.1% versus 65.6%).

Table 2.8: Effect of interacting the treatment effect (factor 1) and placebo group distribution (factor 2) on power.

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
T1: Treatment effect	% Placebo[a]	3.2	12.7	28.5	17.1	26.7	11.7	Power (%)[d]	99.6
constantly weakens	% IVIG[b]	0.2	2.3	14.0	25.4	44.1	13.9		
P2: Even less skewed	logOR[c]	2.60	2.00	1.40	0.80	0.20	NA	Avg. logOR[e]	0.94
T2: Treatment effect limited									
to the hospitalization or	% Placebo	3.2	12.7	28.5	17.1	26.7	11.7	Power (%)	99.7
death categories	% IVIG	1.0	4.6	14.4	13.4	54.9	11.7		
P2: Even less skewed	logOR	1.16	1.16	1.16	1.16	0	NA	Avg. logOR	0.96
T3: Treatment effect limited	% Placebo	3.2	12.7	28.5	17.1	26.7	11.7	Power (%)	19.6
to the discharged categories	% IVIG	3.2	12.7	28.5	17.1	8.6	29.8		
P2: Even less skewed	logOR	0	0	0	0	1.16	NA	Avg. logOR	0.22
T1: Treatment effect	% Placebo	0.4	2.0	7.3	8.1	32.6	49.5	Power (%)	33.1
constantly weakens	% IVIG	0.0	0.3	2.3	6.3	36.7	54.4		
P4: Even more skewed	logOR	2.60	2.00	1.40	0.80	0.20	NA	Avg. logOR	0.33
T2: Treatment effect limited									
to the hospitalization or	% Placebo	0.4	2.0	7.3	8.1	32.6	49.5	Power (%)	16.8
death categories	% IVIG	0.1	0.7	2.5	3.1	44.1	49.5		
P4: Even more skewed	logOR	1.16	1.16	1.16	1.16	0	NA	Avg. logOR	0.21
T3: Treatment effect limited	% Placebo	0.4	2.0	7.3	8.1	32.6	49.5	Power (%)	97.4
to the discharged categories	% IVIG	0.4	2.0	7.3	8.1	6.3	75.9		
P4: Even more skewed	logOR	0	0	0	0	1.16	NA	Avg. logOR	0.91

Under factors 1 and 2, the treatment effect deviates from proportional odds and the distribution of the placebo group deviates from that specified in the FLU-IVIG design assuming no misclassification and the full 6-level ordinal endpoint is used. The treatment effects were paired with placebo group distributions that had more or fewer patients in categories affected by the treatment.

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Table 2.9: Effect of interacting the treatment effect (factor 1) and misclassification (factor 3) on power.

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
T1: Treatment effect	% Placebo[a]	1.2	5.3	16.2	14.4	36.4	26.5	Power (%)[d]	79.1
constantly weakens	% IVIG[b]	0.1	0.8	7.5	12.6	44.9	34.1		
M1: 20% misclassification	logOR[c]	2.60	2.00	1.14	0.80	0.26	NA	Avg. logOR[e]	0.57
T1: Treatment effect	% Placebo	1.2	5.3	15.5	15.1	32.4	30.5	Power (%)	79.6
constantly weakens	% IVIG	0.1	0.8	9.2	10.9	41.3	37.7		
M2: 40% misclassification	logOR	2.60	2.00	0.92	0.80	0.32	NA	Avg. logOR	0.57
T2: Treatment effect limited to the hospitalization or death categories	% Placebo	1.2	5.3	16.2	14.4	36.4	26.5	Power (%)	86.2
	% IVIG	0.4	1.7	6.5	7.0	51.7	32.8		
M1: 20% misclassification	logOR	1.16	1.16	1.12	1.16	0.20	NA	Avg. logOR	0.64
T2: Treatment effect limited to the hospitalization or death categories	% Placebo	1.2	5.3	15.5	15.1	32.4	30.5	Power (%)	92.7
	% IVIG	0.4	1.7	6.6	6.8	45.4	39.1		
M2: 40% misclassification	logOR	1.16	1.16	1.08	1.16	0.38	NA	Avg. logOR	0.71
T3: Treatment effect limited	% Placebo	1.2	5.3	15.8	14.8	36.4	26.5	Power (%)	38.0
to the discharged categories	% IVIG	1.2	5.3	15.8	14.8	18.2	44.7		
M1: 20% misclassification	logOR	0	0	0	0	0.71	NA	Avg. logOR	0.33
T3: Treatment effect limited	% Placebo	1.2	5.3	15.5	15.1	32.4	30.5	Power (%)	7.5
to the discharged categories	% IVIG	1.2	5.3	15.5	15.1	27.0	35.9		
M2: 40% misclassification	logOR	0	0	0	0	0.24	NA	Avg. logOR	0.24

Under factors 1 and 3, the treatment effect deviates from proportional odds and patients are misclassified by % between the non-ICU hospitalized categories and the discharged categories. We assume the distribution of the placebo group is as specified in the FLU-IVIG design and the full 6-level ordinal endpoint is used.

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Table 2.10: Effect of interacting the treatment effect (factor 1) and number of categories (factor 4) on power.

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
T1: Treatment effect	% Placebo[a]	1.2	5.3	16.2	14.4	62.9		Power (%)[d]	95.8
constantly weakens	% IVIG[b]	0.1	0.8	5.8	14.2	79.0			
C3: Discharged collapsed	logOR[c]	2.60	2.00	1.40	0.80	NA		Avg. logOR[e]	0.89
T1: Treatment effect	% Placebo	37.1				36.4	26.5	Power (%)	67.0
constantly weakens	% IVIG	21.0				48.5	30.5		
C4: Four most severe collapsed	logOR	0.80				0.20	NA	Avg. logOR	0.50
T2: Treatment effect limited to the hospitalization or death categories	% Placebo	1.2	5.3	16.2	14.4	62.9		Power (%)	99.5
	% IVIG	0.4	1.7	6.3	7.2	84.4			
C3: Discharged collapsed	logOR	1.16	1.16	1.16	1.16	NA		Avg. logOR	1.16
T2: Treatment effect limited to the hospitalization or death categories	% Placebo	37.1				36.4	26.5	Power (%)	76.0
	% IVIG	15.6				57.9	26.5		
C4: Four most severe collapsed	logOR	1.16				0	NA	Avg. logOR	0.56
T4: Smaller treatment effect limited to the hospitalization or death categories	% Placebo	1.2	5.3	16.2	14.4	62.9		Power (%)	65.6
	% IVIG	0.7	3.1	10.5	10.8	75.0			
C3: Discharged collapsed	logOR	0.57	0.57	0.57	0.57	NA		Avg. logOR	0.57
T4: Smaller treatment effect limited to the hospitalization or death categories	% Placebo	37.1				36.4	26.5	Power (%)	32.0
	% IVIG	25.0				48.5	26.5		
C4: Four most severe collapsed	logOR	0.57				0	NA	Avg. logOR	0.31

Under factors 1 and 4, the treatment effect deviates from proportional odds and categories of the ordinal endpoint are collapsed assuming the distribution of the placebo group is as specified in the FLU-IVIG design and no misclassification. Categories were collapsed according to whether or not they contained the treatment effect.

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

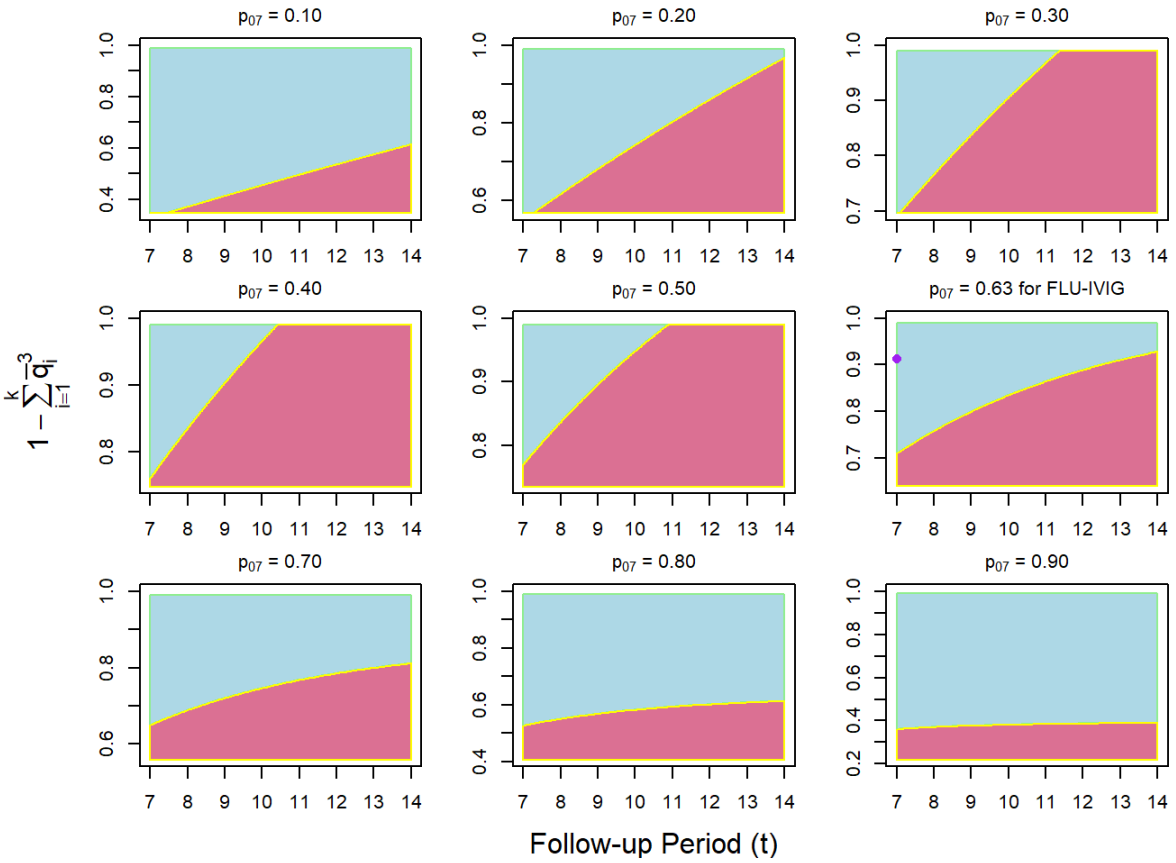
2.3.2 Case Study 2: Comparison to Other Types of Endpoints

2.3.2.1 Analytic Comparison of Ordinal Endpoint to Time-to-Event Endpoint

Figure 2.1 compares the power of the ordinal endpoint to the time-to-event endpoint as a function of the granularity of the ordinal endpoint (i.e., $1 - \sum_{i=1}^{k} \bar{q}_i^3$), the number of days of follow-up for the time-to-event endpoint, and the probability of discharge by day 7 in the placebo group (p_{07}). For fixed p_{07}, the time-to-event endpoint tends to perform better with additional days of follow-up. For p_{07} values of 0.10 or lower and 0.63 or higher, ordinal endpoints of high granularity with values of 0.9 or higher for $1 - \sum_{i=1}^{k} \bar{q}_i^3$ have an almost universal advantage over the time-to-event endpoint, including for the parameter values specified in the power calculations of the FLU-IVIG trial. For p_{07} values ranging from 0.20 to 0.50, longer follow-up periods (i.e., from 8 to 14 days) grant the time-to-event endpoint a near universal advantage over the ordinal endpoint.

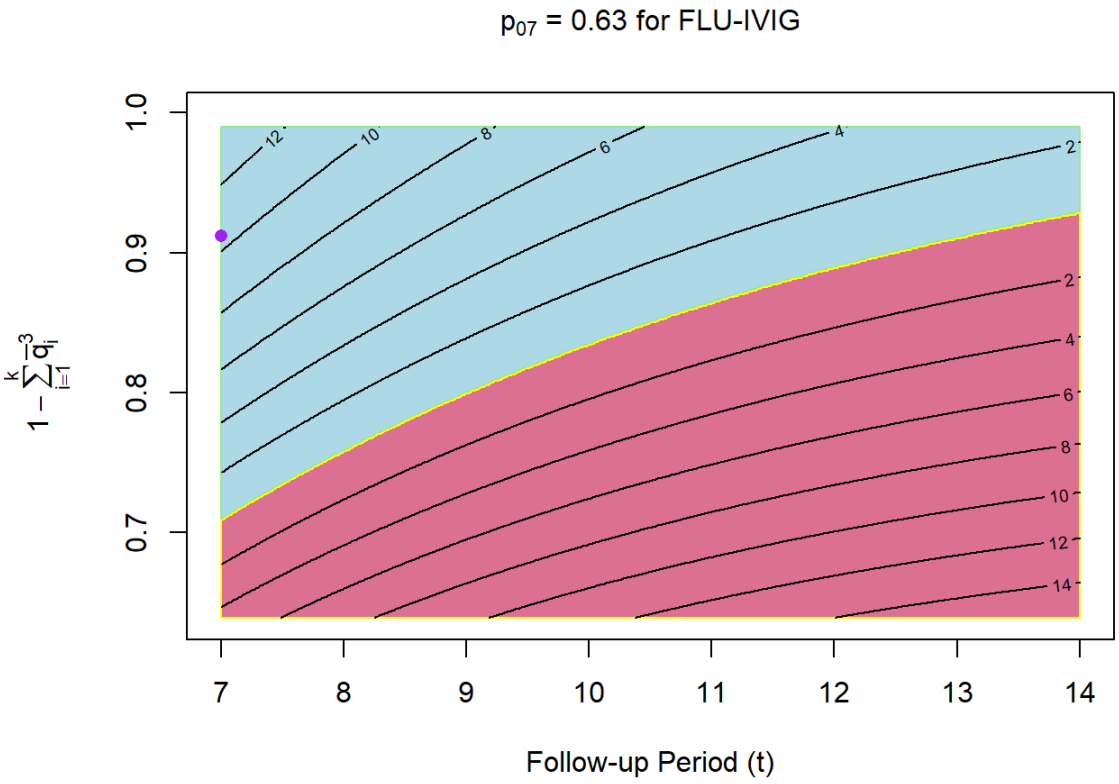
Figure 2.2 displays the right-center plot of Figure 2.1 for FLU-IVIG but with contours of power added to further clarify the difference in performance between the time-to-event and ordinal endpoints. From day 7 to day 14 of follow-up, the time-to-event endpoint gains about 10% power over the ordinal endpoint. As the value of $1 - \sum_{i=1}^{k} \bar{q}_i^3$ increases from the lower bound of $1 - \left(\frac{p_{07}+p_{17}}{2}\right)^3$ to the upper bound of 1, the ordinal endpoint gains about 18% power over the time-to-event endpoint in that span.

Figure 2.1: Comparison of the ordinal endpoint and time-to-event endpoint by follow-up period and $1 - \sum_{i=1}^{k} \bar{q}_i^3$ according to their respective non-centrality parameters which uniquely determine power.



Plots are reproduced across different probabilities of discharge by day 7 (p_07) for the placebo group. The ordinal endpoint is evaluated on day 7, while the time-to-event endpoint is evaluated across days 7–14. The blue region indicates that the ordinal endpoint yielded higher power, while the red region indicates that the time-to-event endpoint yielded higher power. The purple dot in the right-center plot marks the value for FLU-IVIG. As we assume that hospital discharge constitutes at least one category of the ordinal endpoint, the y-axes are bounded below by $1 - \left(\frac{p_{07}+p_{17}}{2}\right)^3$.

Figure 2.2: The right-center plot of Figure 2.1 for FLU-IVIG with contours of power added to indicate by how much either the ordinal endpoint or the time-to-event endpoint surpassed the other in power.



The blue region indicates that the ordinal endpoint yielded higher power, while the red region indicates that the time-to-event endpoint yielded higher power. The purple dot marks the value for FLU-IVIG. As we assume that hospital discharge constitutes at least one category of the ordinal endpoint, the y-axis is bounded below by $1 - \left(\frac{p_{07}+p_{17}}{2}\right)^3$.

2.3.2.2 Simulation Comparison of Ordinal Endpoint to Other Types of Endpoints

We investigated seven different treatment effects each with different transition probabilities across the seven days of follow-up. Each treatment effect is defined from days 1 to 7 of follow-up and arrives at an odds ratio of 1.77 on day 7:

- T1: The treatment effect remains constant for all categories across each day of follow-up.
- T2: The treatment effect only benefits patients for days 1 to 3 for all categories.
- T3: The treatment effect constantly decreases with each day with no additional benefit on day 7 for all categories.
- T4: The treatment effect remains constant across each day of follow-up but is 33.3% more effective for hospitalized patients on the log odds ratio scale.
- T5: The treatment effect only benefits patients in the "ICU" and "non-ICU, on oxygen categories" across each day of follow-up.
- T6: The treatment effect only benefits patients for days 5 to 7 for all categories.
- T7: The treatment effect constantly increases with each day, with benefit starting on day 2 for all categories.

We considered placebo group distributions corresponding to consistently more or less underlying risk throughout the follow-up period:

- The FLU-IVIG placebo group distribution based on the cohort study.17,18
  - Category percentages on day 7 from death to hospital discharge with resumption of normal activities were: 1.0%, 4.9%, 16.3%, 14.5%, 36.2%, 27.1%.
- More severe placebo group (with more severe cases on each day of follow-up).
  - Category percentages on day 7 from death to hospital discharge with resumption of normal activities were: 2.0%, 9.3% 28.0%, 19.2%, 27.2%, 14.3%.
- Less severe group (with fewer severe cases on each day of follow-up).

- Category percentages on day 7 from death to hospital discharge with resumption of normal activities were: 0.5%, 2.2%, 7.1%, 7.7%, 37.4%, 45.0%.

Supplementary Tables A.7 to A.9 display the respective placebo group distributions on day 7 and their corresponding IVIG group distributions for each treatment effect.

Tables 2.11 to 2.13 give the estimated power for the six endpoints considered under different treatment effect scenarios and relative acuity in the patient population. Supplementary Tables A.10 to A.12 give the average coefficient values corresponding to each power estimate (for how to interpret each coefficient estimate, see Table 2.1). For the proportional odds model, adjusting for baseline health status (the bottom half of each table) universally raised power by about 4–9 percentage points across all distributions of the placebo group and treatment effect scenarios considered (see Tables 2.11 to 2.13). Across the three distributions of the placebo group, only the Cox and AFT models assuming a Weibull distribution yielded higher power for certain treatment effects (i.e., T2–4) relative to the proportional model, a finding that generally held with adjustment for baseline health status. Both time-to-event endpoints yielded low power under treatment effects T6 and T7 which assume delay of benefit. Note that the overall proportion of patients discharged by day 7 under each treatment effect scenario is relatively constant (see Supplementary Tables A.7 to A.9), but the discharge times in the treatment group differ significantly across scenarios. Overall, the Weibull AFT model performed better than the exponential AFT model.

Excluding the Cox and AFT models, the proportional odds model generally yielded the highest power across all treatment effects and distributions of the placebo group considered, including after adjustment for baseline health status (see Tables 2.11 to 2.13). Moreover, the proportional odds model with adjustment for baseline health status consistently returned power close the pre-specified level of 80% or higher, ranging from 79.94% to 91.66%. No other model was able to maintain power at the desired level of 80% across all treatment effects and placebo group distributions. However, for each treatment effect under the more severe placebo group distribution, the longitudinal ordinal outcome and sliding dichotomy models generally yielded greater power than the proportional odds model (see Table 2.12). These advantages mostly failed to hold after adjustment for baseline health status. Across all treatment effects and placebo group

distributions, the simple logistic model and win ratio generally yielded lower power relative to the proportional odds model including after adjustment for baseline health status.

Table 2.11: Simulated power (%)[a] results for the FLU-IVIG placebo group on day 7 of follow-up (p_{07}[b]$= 0.63$).[c]

Model	T1: Constant benefit[d]	T2: Benefit only for first three days	T3: Linear decrease in benefit	T4: Larger benefit for hospitalized patients	T5: Benefit only for ICU and non-ICU, on oxygen patients	T6: Benefit only for last three days	T7: Linear increase in benefit
No adjustment for baseline status							
Simple Logistic	71.65	75.07	75.29	76.70	83.20	60.79	67.87
Sliding Dichotomy	69.55	78.04	75.71	74.81	**93.04**	59.11	64.11
Win Ratio	**79.54**	79.58	79.38	79.61	79.66	**79.86**	**80.66**
Proportional Odds	**79.71**	79.69	79.45	79.68	79.86	**79.94**	**80.77**
Longitudinal Ordinal Outcome	**82.98**	71.52	77.96	**83.35**	67.21	**79.45**	**84.59**
Cox Proportional Hazards	76.73	**96.46**	**92.66**	**81.92**	**87.47**	36.39	53.81
Accelerated Failure Time (Exponential)	68.43	**94.24**	**89.17**	74.72	81.44	27.45	42.49
Accelerated Failure Time (Weibull)	78.83	**96.64**	**93.62**	**83.78**	**88.55**	37.09	55.83
Adjustment for baseline status							
Simple Logistic	78.70	81.24	82.28	83.68	**86.08**	68.84	75.55
Sliding Dichotomy	71.96	79.82	77.92	77.49	**93.29**	61.88	67.09
Win Ratio[e]	83.67	82.90	82.59	83.94	72.09	**86.26**	**86.14**
Proportional Odds	**86.25**	86.50	85.95	**86.50**	**87.26**	**86.93**	**86.70**
Longitudinal Ordinal Outcome	70.45	23.81	42.36	70.12	53.20	**84.90**	**85.63**
Cox Proportional Hazards	**84.93**	**98.92**	**96.89**	**89.27**	81.81	40.94	60.28
Accelerated Failure Time (Exponential)	74.41	**97.37**	**93.94**	80.50	75.69	26.67	43.51
Accelerated Failure Time (Weibull)	**86.23**	**98.98**	**97.21**	**90.48**	82.51	41.08	61.99

[a]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the estimated treatment effect of the fitted model was significant at the two-sided 0.05 level. Power for the top three models for each treatment effect with or without adjustment for baseline health status is printed in bold. [b]p_{07} denotes the probability of discharge (i.e., categories 5 and 6 of the ordinal endpoint combined) by day 7 of follow-up for the placebo group.

[c]Supplementary Table 1 displays the category percentages for the FLU-IVIG placebo group on day 7 and each of the seven IVIG groups.

[d]Benefit refers to the differences in transition probabilities between randomized groups across treatment effects.

[e]Unlike all other models which include a covariate to adjust for baseline status, the win ratio stratifies by only comparing patients who started at the same baseline category between randomized groups.

Table 2.12: Simulated power (%)[a] results for the more severe placebo group on day 7 of follow-up (p_{07}[b]$= 0.42$).[c]

Model	T1: Constant benefit[d]	T2: Benefit only for first three days	T3: Linear decrease in benefit	T4: Larger benefit for hospitalized patients	T5: Benefit only for ICU and non-ICU, on oxygen patients	T6: Benefit only for last three days	T7: Linear increase in benefit
No adjustment for baseline status							
Simple Logistic	82.12	80.31	81.22	83.64	65.36	80.66	**82.11**
Sliding Dichotomy	**82.88**	86.93	85.24	**84.25**	**90.98**	79.40	80.78
Win Ratio	80.85	81.52	80.71	81.15	80.61	**81.28**	81.04
Proportional Odds	80.97	81.59	80.89	81.31	**80.77**	**81.43**	**81.21**
Longitudinal Ordinal Outcome	**91.73**	87.42	**90.20**	**92.16**	**87.42**	**85.37**	**90.88**
Cox Proportional Hazards	82.14	**92.36**	**89.57**	84.21	68.21	62.28	71.95
Accelerated Failure Time (Exponential)	78.48	**90.77**	87.13	80.57	63.60	56.69	67.01
Accelerated Failure Time (Weibull)	**82.70**	**92.53**	**90.08**	**84.90**	68.96	61.33	72.09
Adjustment for baseline status							
Simple Logistic	87.82	85.40	86.69	88.66	69.21	86.46	87.88
Sliding Dichotomy	85.13	88.85	87.48	86.52	**91.51**	82.00	83.25
Win Ratio[e]	**89.45**	88.78	88.41	88.64	**88.84**	**89.51**	**89.62**
Proportional Odds	**90.92**	91.00	90.35	**90.86**	**89.69**	**91.66**	**90.76**
Longitudinal Ordinal Outcome	84.67	49.41	65.97	84.07	82.32	**92.25**	**92.45**
Cox Proportional Hazards	88.46	**96.79**	**94.64**	**90.08**	64.76	66.60	77.72
Accelerated Failure Time (Exponential)	83.84	**95.45**	**92.43**	86.31	60.35	57.61	70.32
Accelerated Failure Time (Weibull)	**88.92**	**96.85**	**94.92**	**90.61**	65.07	65.65	77.73

[a]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the estimated treatment effect of the fitted model was significant at the two-sided 0.05 level. Power for the top three models for each treatment effect with or without adjustment for baseline health status is printed in bold. [b]p_{07} denotes the probability of discharge (i.e., categories 5 and 6 of the ordinal endpoint combined) by day 7 of follow-up for the placebo group.

[c]Supplementary Table 1 displays the category percentages for the FLU-IVIG placebo group on day 7 and each of the seven IVIG groups.

[d]Benefit refers to the differences in transition probabilities between randomized groups across treatment effects.

[e]Unlike all other models which include a covariate to adjust for baseline status, the win ratio stratifies by only comparing patients who started at the same baseline category between randomized groups.

Table 2.13: Simulated power (%)[a] results for the less severe placebo group on day 7 of follow-up (p_{07}[b]= 0.82).[c]

Model	T1: Constant benefit[d]	T2: Benefit only for first three days	T3: Linear decrease in benefit	T4: Larger benefit for hospitalized patients	T6: Benefit only for last three days	T7: Linear increase in benefit
No adjustment for baseline status						
Simple Logistic	54.07	63.20	63.42	62.44	36.23	45.35
Sliding Dichotomy	47.75	58.77	56.92	55.50	32.77	39.06
Win Ratio	**75.87**	76.15	76.16	76.08	**76.09**	**75.59**
Proportional Odds	**75.90**	**76.20**	76.16	**76.13**	**76.11**	**75.63**
Longitudinal Ordinal Outcome	64.44	44.81	52.52	62.78	**65.88**	**70.26**
Cox Proportional Hazards	70.74	**99.04**	**95.81**	**80.04**	15.94	31.88
Accelerated Failure Time (Exponential)	55.18	**97.11**	**90.71**	66.84	7.51	17.25
Accelerated Failure Time (Weibull)	**75.21**	**99.11**	**96.52**	**83.48**	18.67	37.27
Adjustment for baseline status						
Simple Logistic	59.39	68.44	69.06	68.29	41.08	50.68
Sliding Dichotomy	49.97	60.64	58.93	57.89	34.40	41.08
Win Ratio[e]	73.82	72.61	72.31	73.28	**77.76**	**75.46**
Proportional Odds	**80.52**	81.15	80.80	**80.68**	**80.12**	**79.94**
Longitudinal Ordinal Outcome	45.15	6.13	14.09	40.69	**71.35**	**70.70**
Cox Proportional Hazards	**79.02**	**99.79**	**98.50**	**87.80**	17.60	37.29
Accelerated Failure Time (Exponential)	58.68	**98.82**	**94.48**	70.81	5.83	16.23
Accelerated Failure Time (Weibull)	**82.50**	**99.79**	**98.80**	**90.09**	20.03	41.97

[a]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the estimated treatment effect of the fitted model was significant at the two-sided 0.05 level. Power for the top three models for each treatment effect with or without adjustment for baseline health status is printed in bold. [b]p_{07} denotes the probability of discharge (i.e., categories 5 and 6 of the ordinal endpoint combined) by day 7 of follow-up for the placebo group.

[c]Supplementary Table 1 displays the category percentages for the FLU-IVIG placebo group on day 7 and each of the seven IVIG groups.

[d]Benefit refers to the differences in transition probabilities between randomized groups across treatment effects.

[e]Unlike all other models which include a covariate to adjust for baseline status, the win ratio stratifies by only comparing patients who started at the same baseline category between randomized groups.

Due to the skewness of the less severe placebo group distribution, a treatment effect corresponding to T5 that approximated an odds ratio of 1.77 on day 7 could not be found.

2.4 Discussion

To our knowledge, the FLU-IVIG study is the first randomized trial to use an ordinal endpoint to evaluate a novel influenza treatment. A variation of the ordinal scale is now being used by other influenza trials as a primary or secondary endpoint.[13,14] Any novel endpoint should be rigorously evaluated to address clinical and statistical concerns, especially one that makes use of a relatively uncommon data type like an ordinal scale. Thus, we considered it necessary to thoroughly examine the ordinal endpoint with respect to factors that may affect its statistical power for the trial. Additionally, a novel endpoint should be both interpretable and able to more consistently detect a treatment effect relative to other endpoints that may be derived from the same information. In that regard, we compared the power of the FLU-IVIG ordinal endpoint to a time-to-event endpoint analytically and by simulation and to five other endpoints by simulation. Selecting the most efficient endpoint for a trial includes weighing a number of clinical and statistical factors, and our case studies of the FLU-IVIG ordinal endpoint give insight on their relative importance.

Examining factors that may affect the power of the ordinal endpoint, we find that the ordinal endpoint yields higher power relative to any collapse of the ordinal endpoint into a binary endpoint. Further, the decisions about the number of categories and assumptions made about the treatment effect, distribution of the placebo group, and the amount of misclassification can have substantial consequences for power. Provided that the overall treatment effect is maintained and other factors are held constant, deviations from proportional odds marginally reduce power. We also found that, holding other factors constant, more skewed placebo group distributions, misclassification of patients among the ordinal categories, and considering fewer ordinal categories decreased power consistent with previous research.[17,26,27]

However, we find that these general conclusions must be qualified as the effect of each of these factors may be reversed when another factor is varied simultaneously. To increase power, if the proportional odds assumption does not hold, categories in which the treatment is presumed to be effective should be divided as evenly as possible; conversely, categories where the treatment is presumed to be less effective should be collapsed.

For IVIG, the treatment may be more beneficial for severe cases. Therefore, we considered an ordinal endpoint which was granular for these patients to attain sufficient power.

Comparing the ordinal endpoint to other types of endpoints, we find that there are some scenarios where a time-to-event endpoint may be better suited to detect a treatment effect. Provided that the hazard ratio remains constant over the follow-up period, we demonstrated analytically that time-to-event endpoints assessed over a longer follow-up period yield greater power than ordinal endpoints when given moderate placebo group discharge probabilities at the time the ordinal endpoint was assessed on day 7. With a high number of placebo group patients on the cusp of hospital discharge on day 7, we would expect the time-to-event endpoint to substantially improve in power with follow-up periods of longer than 7 days which provide more time for events (i.e., hospital discharge) to occur. Conversely, with high discharge probabilities on day 7 and hence more patients who have already had the event in the placebo group, we would expect longer follow-up periods to only marginally benefit the time-to-event endpoint. Similarly, with very low discharge probabilities on day 7, very few patients would be close to leaving the hospital in the placebo group. Follow-up periods beyond 14 days would then be required to raise the power of the time-to-event endpoint above the ordinal endpoint. If trial designers expect a fairly ill target population and can assume that the treatment effect will be constant over a prolonged period, time-to-event endpoints may be preferable.

Of course, a constant treatment effect may not be reasonable to assume. Our simulation study gives insight into the types of endpoints that may provide the highest power under plausible expressions of a non-constant treatment effect. Regardless of the underlying risk in the placebo group, for treatment effects T2–4, the Cox and Weibull AFT models generally returned larger power relative to the proportional odds model with or without adjustment for baseline health status. This is likely because treatment effects T2–4 assume disproportionate benefit soon after randomization or to hospitalized patients to reduce time-to-hospital discharge. Conversely, when the treatment effect was delayed as in T6 and T7, the time-to-event endpoints performed relatively poorly. This is likely because treatment effects T6 and T7 fail to discharge patients from the hospital early

on in follow-up. In terms of consistency, the proportional odds model with adjustment for baseline health status was the only model to return power at the pre-specified level or higher across all treatment effect and placebo group combinations.

Alternatively, the ordinal endpoint evaluated on day 7 could be dichotomized into a binary endpoint but with the potential caveat of a loss of information. To that end, more complex binary endpoints have been proposed. For the purposes of our study, neither the win ratio, which compares ordinal endpoint category on day 7, nor the sliding dichotomy, which compares the change in patient health status between baseline and day 7, consistently improved power relative to the ordinal endpoint. The win ratio without adjustment for baseline health status was always less than half a percentage point below the proportional odds model in power. With adjustment for baseline health status, the win ratio was often within a few percentage points of the proportional odds model. The gap may have widened because unlike all other models in this paper which include a covariate to adjust for baseline health status, the win ratio stratifies by only comparing patients who started at the same baseline category between randomized groups. This may have resulted in a number of ties in patient improvement to weaken increases in power. Overall, the win ratio closely tracked the proportional odds model in power and without adjustment for baseline health status may be preferable given its ease of interpretation relative to the odds ratio from fitting a proportional odds model.

In addition, longitudinal endpoints may perform better than endpoints assessed at a single time point. Yet, we found that the longitudinal ordinal outcome model was less powerful than evaluating the ordinal endpoint at a single time point, except when enrolling a target population with many hospitalized patients (i.e., the more severe placebo group). This may be because the longitudinal ordinal outcome model studied assumes a constant treatment effect over time, an erroneous assumption given our data generating mechanism in the simulation setting. Additionally, test statistics other than the Wald test statistic of the day by treatment group interaction may be more powerful.

In the simulation setting, our analysis found that adjusting for baseline health status universally increased power across all treatment effect and placebo group combinations for the proportional odds model, and for many of the other models. Analyses which

include an adjustment for baseline health status should be considered even if the primary endpoint is not an ordinal endpoint.

In contrast to previous research,[17,26,27] we explored deviations from proportional odds while holding the overall treatment effect constant. In addition, we studied the joint effect of multiple factors related to design decisions and assumptions about the ordinal endpoint. Previous research has primarily examined the effect of a single factor at a time. Additionally, previous research has generally compared the proportional odds model to other models by fitting them to retrospective data.[28] Our longitudinal simulation scheme provides a framework for how such data may be generated, and can be engineered to generate data that meet any odds ratio of interest.

Our case studies are mainly limited by their specificity to the six-level ordinal endpoint and the type of target population considered in the FLU-IVIG trial. Ordinal endpoints can be re-defined to better fit the target population. For example, if the FLU-IVIG ordinal endpoint were to be used for a more severely ill target population, the In ICU category could be divided according to whether mechanical ventilation is required and the two hospital discharge categories could be combined. Additionally, in an unblinded study, the discharge categories as well as the oxygen categories could be combined due to their dependence on subjective clinician/patient assessments. Though different ordinal endpoints have been used in trials of vascular disease, S. pneumoniae infection, and traumatic brain injury,[28–31] our general findings and approach to evaluating the FLU-IVIG ordinal endpoint are applicable to those ordinal endpoints and others.

When selecting an endpoint, our findings suggest that ordinal endpoints with high granularity can reliably exceed time-to-event endpoints in power for hospitalized influenza populations in which most patients will naturally progress to hospital discharge by the time of endpoint assessment, similar to the FLU-IVIG trial. Furthermore, we find that the FLU-IVIG ordinal endpoint can perform even better after adjusting for baseline health status. However, if only 20–50% of patients will be discharged from the hospital by the end of the follow-up period, leaving many on the cusp of discharge, a time-to-event endpoint with a longer follow-up period may be more efficient. Additionally, a time-to-event endpoint may be able to more reliably detect a treatment effect that is strongest in the first few days following randomization.

More broadly, when deciding between an ordinal endpoint and a time-to-event endpoint, important factors to consider for each are the granularity of the ordinal endpoint and the follow-up period of the time-to-event endpoint. Categories which contain the largest proportions of patients should be divided as evenly as possible; conversely, categories which contain the fewest patients should be collapsed into categories with greater numbers. For the time-to-event endpoint, longer follow-up periods should be weighed to allow more time for events to occur. Moreover, longer follow-up periods and trials that cannot be blinded may not be ideal for ordinal endpoints that rely on patients to report information, like the two discharged categories in the FLU-IVIG ordinal endpoint. Concentrating on these two factors — granularity and follow-up period — in addition to the target population should help trial designers choose between ordinal endpoints and time-to-event endpoints for trials in severe influenza and other disease areas.

Although the choice of an ordinal outcome as the primary endpoint may be uncommon in clinical trials, our findings from both case studies and others provide strong support for the robustness of an ordinal endpoint to detect a treatment effect in trials of severe influenza.[28–31] The data type of an ordinal scale with its multiple qualitative categories should not be discouraging to trial designers; rather, if an ordinal scale is to be considered, factors that may influence its clinical relevance and power should be thoroughly investigated.

Chapter 3

Detecting Participant Noncompliance Across Multiple Time Points by Modeling a Longitudinal Biomarker

3.1 Introduction

Participant noncompliance, in which participants do not follow their randomly assigned treatment protocol, has long complicated the interpretation and conduct of randomized clinical trials. Participant noncompliance often occurs when participants must self-administer the treatment without the supervision of study personnel.[32] Given this autonomy, participants may deviate from their treatment protocol for numerous reasons, including advent of side effects, insufficient benefit, and availability of commercial alternatives to the treatment.[33] Examples of trials with participant noncompliance include regulatory tobacco trials of very low nicotine content cigarettes, where current smokers must only smoke the study cigarettes;[34] and opioid dependence trials, where patients with substance abuse disorders must self-inject study depots.[35,36] Additionally, in longitudinal studies, multiple time points present multiple opportunities for participants

to be noncompliant.

Compared to compliant participants, noncompliant participants do not receive the full dose of treatment and as such may have poor study outcomes. This may dilute the intention-to-treat (ITT) estimate.[37] Additionally, both noncompliance and self-reported noncompliance may systematically vary according to some confounding variable.[38] This may create a difference between the Per Protocol estimate based on self-reported compliance and the treatment effect if all participants had complied (i.e., the causal effect).[39] As a motivating example, in the United States, the Family Smoking Prevention and Tobacco Control Act provides the Food and Drug Administration with the regulatory authority to reduce (but not eliminate) the nicotine content in commercial cigarettes if it would improve public health. Regulatory tobacco trials seek to evaluate the effect of such potential changes to commercial cigarettes. As these trials investigate interventions that could be mandated by federal law to force compliance, the causal effect is more relevant than the ITT estimator.[40]

Modeling longitudinal compliance is thus important for the following reasons. First, for completed trials, identifying compliant participants enables estimation of the causal effect,[41] which may be different than both the ITT and Per Protocol estimates. Methods are thus needed to identify compliant participants to properly weigh the study outcomes and adjust for confounding, but only those for a single time point have been proposed.[42,43] Second, noncompliant participants are more likely to drop out.[37] Methods are thus needed to identify noncompliant participants during the trial for remedial intervention.

However, in many therapeutic areas, no gold standard has been identified for detecting participant noncompliance. Trial designers must frequently rely on imperfect measures of participant noncompliance based on subjective or indirect information (e.g., self-reported compliance).[44,45] Moreover, detecting participant noncompliance falls into the broader category of diagnosis without a gold standard, a subject of considerable study. With respect to true (but unobserved) disease status, statistical methods have been able to both estimate summary measures (e.g., disease prevalence) and individually model disease progression.[46,47] Yet, few statistical methods exist for diagnosing patients whose disease status (in our case, compliance) may shift back and forth between two disease

states.

In some trials, participants' biomarkers may systematically change in response to the treatment or any alternatives to provide objective information about noncompliance. Consider three recently published regulatory tobacco trials studying the effect of very low nicotine content cigarettes (with 0.4mg of nicotine per gram of tobacco) versus normal nicotine content cigarettes (with 15.8mg of nicotine per gram of tobacco) on smoking behavior.[34,48,49] During the follow-up periods of 6 or 20 weeks, participants were asked to only smoke their assigned study cigarettes but could additionally smoke commercial cigarettes (i.e., non-study cigarettes) with normal nicotine content. Participants who smoked non-study cigarettes were considered to be noncompliant. Most participants self-reported compliance at each time point.

At each follow-up visit, participants were asked to give samples of various biomarkers including total nicotine equivalents, which measures most nicotine metabolites in the urine to evaluate recent nicotine exposure. Based on the findings of a previous study of participants who were sequestered in a hotel and only had access to very low nicotine content cigarettes, only 5% of participants randomized to very low nicotine cigarettes were expected to have total nicotine equivalents above 6.41 nmol/mL if they were fully compliant.[50] However, in one trial, 63% of participants who self-reported compliance at week 6 had total nicotine equivalents above 6.41 nmol/mL.[51]

In randomized trials of very low nicotine content cigarettes, biomarkers like total nicotine equivalents can suggest exposure to non-study cigarettes.[52] In the absence of a gold standard for compliance, Boatman et al. (2017) developed a method for estimating the probability of compliance at a single time point based on biomarker data collected at that time point.[42] However, this method cannot leverage biomarker data collected at previous time points, and nor can it predict compliance at a future time point.

We propose a longitudinal method that uses biomarker data collected across multiple time points to estimate longitudinal compliance when the true compliance status is not directly observed. Specifically, we model the biomarker as a mixture density whose (latent) components correspond to different compliance patterns over time, a modeling approach similar to growth mixture models.[53] Our method can estimate the probability

of compliance at: 1) a single time point of the trial; 2) all time points; and 3) a future time point. We first evaluate our method by simulation, examining two factors: 1) the effect of compliance on the biomarker; and 2) incorrect specification of the covariance structures of the models fit. Second, we apply our method to the data collected in the Center for the Evaluation of Nicotine in Cigarettes Project 1 Study 2 (CENIC-P1S2) trial.

3.2 Methods

3.2.1 Overview

For our purposes, we treat the true compliance status as unobserved and binary; that is, each participant can either fully comply or not fully comply at each time point. To derive the probability of compliance at a single time point conditional on a single biomarker measurement, Boatman et al. modeled the distribution of the biomarker as a two-component mixture density of compliance and noncompliance and then used Bayes' rule to derive the desired probability of compliance from the mixture density.[42] Our goal is to generalize this approach to multiple time points to identify if and when participants are noncompliant. In the longitudinal setting, the number of components in the mixture density exponentially increases beyond two as participants can shift between compliance and noncompliance at each time point. Additionally, multiple observations on each participant present the possibility of correlated data. In the following section, we explain how to extend the mixture density to multiple time points while allowing for correlation among repeated measurements.

3.2.2 Modeling the Mixture Density

Let C_{ij} be the true, unobserved compliance indicator (1 denotes compliance, 0 denotes noncompliance) for the ith participant at the jth time point, where $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, K\}$. Similarly, let B_{ij} be the observed biomarker of exposure for the ith participant measured at the jth time point. As participant compliance can vary over time, let c_k be the kth compliance pattern where $k \in \{1, \ldots, 2^K\}$ for K time points. For example, for $K = 2$ time points, $c_1 = \{0, 0\}$, $c_2 = \{1, 0\}$, $c_3 = \{0, 1\}$, and $c_4 = \{1, 1\}$.

Across time points, we specify the joint density of the biomarker as a mixture density with components corresponding to the 2^K compliance patterns. That is,

$$f(B_{i1},\ldots,B_{iK}) = \sum_{k=1}^{2^K} P(\{C_{i1},\ldots,C_{iK}\}=c_k) f(B_{i1},\ldots,B_{iK}|\{C_{i1},\ldots,C_{iK}\}=c_k). \tag{3.1}$$

A number of longitudinal models can be fit to $P(\{C_{i1},\ldots,C_{iK}\} = c_k)$ and $f(B_{i1},\ldots,B_{iK}|\{C_{i1},\ldots,C_{iK}\} = c_k)$. For both densities, we fit mixed effects models such that conditional on a vector of random effects, observations on the same participant are independent. For $P(\{C_{i1},\ldots,C_{iK}\}=c_k)$, we have:

$$\begin{aligned} P(\{C_{i1},\ldots,C_{iK}\}=c_k) &= \int_{-\infty}^{\infty} P(\{C_{i1},\ldots,C_{iK}\}=c_k|q_i)\psi(q_i)dq_i \\ &= \int_{-\infty}^{\infty} \prod_{j=1}^{K} P(C_{ij}=c_{ij}|q_i)\psi(q_i)dq_i, \end{aligned}$$

where q_i is the participant-specific random intercept with probability density function (PDF) $\psi(q_i)$, and we have used the conditional independence assumption in the last equality to write $P(\{C_{i1},\ldots,C_{iK}\}=c_k|q_i) = \prod_{j=1}^{K} P(C_{ij}=c_{ij}|q_i)$.

Similarly, we can use the conditional independence assumption to write $f(B_{i1},\ldots,B_{iK}|\{C_{i1},\ldots,C_{iK}\}=c_k)$ as:

$$\begin{aligned} f(B_{i1},\ldots,B_{iK}|\{C_{i1},\ldots,C_{iK}\}=c_k) &= \int_{-\infty}^{\infty} g(B_{i1},\ldots,B_{iK}|\{C_{i1},\ldots,C_{iK}\}=c_k,z_i)\omega(z_i)dz_i \\ &= \int_{-\infty}^{\infty} \prod_{j=1}^{K} g(B_{ij}|C_{ij}=c_{ij},z_i)\omega(z_i)dz_i, \end{aligned}$$

where z_i is the participant-specific random intercept with PDF $\omega(z_i)$ and is independent of compliance status.

Our model setup allows us to select link functions and densities for $P(C_{ij}=c_{ij}|q_i)$ and $g(B_{ij}|C_{ij}=c_{ij},z_i)$. Additionally, we can select distributions for the random intercepts q_i and z_i. For compliance, we assume a probit link function such that:

$$P(C_{ij}=1|q_i) = \Phi(\beta_0+q_i), \tag{3.2}$$

where Φ denotes the cumulative density function of the standard normal distribution and $q_i \sim N(0,\gamma^2)$. The inclusion of a random intercept implies a compound symmetry

covariance structure. This model assumes that the probability of compliance does not change over time. For the biomarker, we assume an identity link function such that:

$$g(B_{ij}|C_{ij} = c_{ij}, z_i) = N(\alpha_0 + \alpha_1 c_{ij} + z_i, \sigma^2), \tag{3.3}$$

where $z_i \sim N(0, \tau^2)$ which again implies compound symmetry. This model assumes that the distribution of the biomarker depends only on the current compliance status.

Our method depends on correctly specifying both the covariates and covariance structures for both the compliance and biomarker mixed effects models. As our method is fully likelihood-based, any number of likelihood-based tests and information criteria can be used to select which models to fit. In the application to CENIC-P1S2, we will use Bayesian Information Criterion (BIC) to investigate whether or not to include a fixed effect for time in the compliance mixed effects model.[54] While likelihood-based tests and information criteria can help guide model selection, the models fit may not be correctly specified. Moreover, previous research of latent models has found that model misspecification of the covariance structures can lead to bias and variability in the parameter estimates.[55,56] In the simulation, we will investigate the effect of misspecifying the covariance structures on model fit.

Let $\boldsymbol{\theta} = (\beta_0, \gamma^2, \alpha_0, \alpha_1, \tau^2, \sigma^2)^T$ be the vector of parameters to be estimated. As is often the case with mixture densities such as (3.1), it would be difficult to directly solve for the maximum likelihood estimates of $\boldsymbol{\theta}$. Instead, we can use the EM algorithm.[57] Appendix B.1 delineates the E-step and M-step.

One advantage of this approach is that we can include participants who lack biomarker data at some time points. That is, mixed effects models estimated by maximum likelihood can inherently handle data that are missing at random. Given this assumption, estimators of $\boldsymbol{\theta}$ are consistent.

3.2.3 Deriving Compliance Probabilities of Interest

We can use the mixture density to derive compliance probabilities that condition on the longitudinal biomarker data. Intuitively, these compliance probabilities should be more accurate than methods which only condition on a single biomarker measurement.

Examining all time points, we can use Bayes' rule to derive the probability of compliance at all time points conditional on the biomarker history as:

$$P(C_{i1} = 1, \ldots, C_{iK} = 1 | B_{i1}, \ldots, B_{iK}) = \frac{P(C_{i1} = 1, \ldots, C_{iK} = 1) f(B_{i1}, \ldots, B_{iK} | C_{i1} = 1, \ldots, C_{iK} = 1)}{\sum_{k=1}^{2^K} P(\{C_{i1}, \ldots, C_{iK}\} = c_k) f(B_{i1}, \ldots, B_{iK} | \{C_{i1}, \ldots, C_{iK}\} = c_k)}.$$

Note that both $P(\{C_{i1}, \ldots, C_{iK}\} = c_k)$ and $f(B_{i1}, \ldots, B_{iK} | \{C_{i1}, \ldots, C_{iK}\} = c_k)$ are already specified as part of the mixture density in (3.1) and can be estimated by plugging in the maximum likelihood estimates of $\boldsymbol{\theta}$.

Additionally, by summing over the relevant posterior compliance probabilities, we can derive the probability of compliance at the last time point conditional on the biomarker history as:

$$P(C_{iK} = 1 | B_{i1}, \ldots, B_{iK}) = \sum_{k=1}^{2^{K-1}} P(C_{iK} = 1, \{C_{i1}, \ldots, C_{iK-1}\} = c_k | B_{i1}, \ldots, B_{iK}).$$

Assuming that we are given biomarker data up to but not including the last time point, we can integrate out the future unobserved biomarker value to derive the prediction probability of compliance at the last time point as:

$$P(C_{iK} = 1 | B_{i1}, \ldots, B_{iK-1}) = \int_{-\infty}^{\infty} P(C_{iK} = 1 | B_{i1}, \ldots, B_{iK}) f(B_{iK} | B_{i1}, \ldots, B_{iK-1}) dB_{iK},$$

where:

$$f(B_{iK} | B_{i1}, \ldots, B_{iK-1}) = \frac{f(B_{i1}, \ldots, B_{iK})}{f(B_{i1}, \ldots, B_{iK-1})}.$$

Note that both the numerator and denominator of the above fraction are simply the mixture density in (3.1) given K and $K-1$ time points, respectively, and can be estimated by plugging in the maximum likelihood estimates of $\boldsymbol{\theta}$ for the available $K-1$ time points.

3.2.4 Simulation Study Design

We used Monte Carlo simulation to assess the accuracy of our method in discriminating between compliers and noncompliers as well as predicting compliance at a future time point. We generated longitudinal compliance data from the mixed effects probit model

outlined in (3.2), where there are no covariates and a single participant-specific random intercept. Specifically, we set $\beta_0 = 0$ to make it equally likely that a participant did or did not comply with their treatment assignment at each time point on average. We set $\gamma^2 = 1$ to make for equal between-subject variability and within-subject variability in compliance status.

We generated longitudinal biomarker data from the linear mixed effects model outlined in (3.3), where there is an indicator variable for compliance at the given time point and a single participant-specific random intercept. Specifically, we set the intercept of noncompliant participants to be $\alpha_0 = 4.0$. Note that for the purposes of detecting noncompliance, the exact value of α_0 does not matter. We set $\tau^2 = 0.7$ and $\sigma^2 = 0.3$ for two reasons: 1) to simulate more between-subject variability than within-subject variability in biomarker values conditional on compliance status with $\tau^2 > \sigma^2$; and 2) so that any compliance effect (α_1) is already standardized with $\sqrt{\tau^2 + \sigma^2} = 1$.

To set the effect of compliance on the mean biomarker value (i.e., α_1), we can use the binormal receiver operating characteristic curve to solve for values of α_1 that result in a desired AUC for compliance based off a single biomarker measurement.[58] As the biomarkers in the CENIC trials have been found to have a strong association with compliance,[59] we set $\alpha_1 \in \{-1.5, -1.8, -2.1\}$ which correspond to exact AUC values of $\{0.856, 0.898, 0.931\}$, respectively.

To mimic the data collected in CENIC-P1S2, we set $n = 100$ participants with $K = 6$ time points. For each simulation scenario, we generated data for 1000 trials each with no missing data. As a supplement, we ran simulations with a larger sample size to investigate a weaker compliance effect. In these simulations, $n = 500$ and $\alpha_1 \in \{-1.2, -1.5, -1.8, -2.1\}$, where $\alpha_1 = -1.2$ corresponds to an exact AUC of 0.802.

Additionally, for both simulations of $n = 100$ and $n = 500$, we investigated scenarios where the covariance structures of both the compliance and biomarker mixed effects models were incorrectly specified. In these scenarios, both variables were generated from first order autoregressive (AR(1)) covariance structures but models with compound symmetry were erroneously fit. Only the regression coefficients are consistent between the data generating mechanism and the fitted models. Appendix B.2 provides the exact

form of both covariance structures under correct and incorrect specification.

To evaluate our method, we compared the average parameter estimates to the true values, where only comparisons for $(\beta_0, \alpha_0, \alpha_1)^T$ were made for the scenario of AR(1). We derived two sets of parameter estimates: 1) those for all $K = 6$ time points; and 2) those for the first five time points which will be used for prediction of compliance at the last time point.

To assess the accuracy of our method in detecting noncompliance, we first compared the area under the curve (AUC) values of the various compliance probabilities under the true parameters to the AUC values under the estimated parameters. Moreover, to protect against model overfit, AUC values were estimated from large, fixed independent test sets (i.e., we used Monte Carlo integration) as closed-form solutions are not available.

Additionally, to measure the extent to which conditioning on longitudinal biomarker data improves detection of noncompliance, we compared the AUC values of $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i6})$ to $\hat{P}(C_{i6} = 1|B_{i6})$, where the latter conditions only on the last biomarker value. To create a comparator for $\hat{P}(C_{i1} = 1, \ldots, C_{i6} = 1|B_{i1}, \ldots, B_{i6})$, we alternatively estimated the probability of compliance at all time points as $\prod_{j=1}^{6} \hat{P}(C_{ij} = 1|B_{ij})$ under the (erroneous) assumption that observations across time points are independent. In this expression, $\hat{P}(C_{ij} = 1|B_{ij})$ is estimated six times corresponding to the six time points of j.

To assess the accuracy of $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i5})$, we plotted its calibration lines. Specifically, for each set of parameter estimates from the first five time points, we plotted the average estimated prediction probability of $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i5})$ against the observed probability by decile in the corresponding fixed independent test set. Code to implement our method in the programming language R is available as a GitHub repository (https://github.com/RPeterson4/Longitudinal_Noncompliance_Detection).

3.2.5 Application to CENIC-P1S2

We applied our method to the CENIC-P1S2 trial, which sought to evaluate the effect of very low nicotine content cigarettes with and without a nicotine patch on number of cigarettes smoked per day (NCT02301325).[49] Current smokers who had no intention of

quitting were randomized in a 2x2 factorial design, with very low nicotine content vs normal nicotine content cigarettes as the first factor and nicotine patch vs no nicotine patch as the second factor. During the follow-up period of six weeks with study visits every week, participants were asked to only smoke the study cigarettes provided by the trial but could additionally smoke commercial cigarettes.

Although previous trials of very low nicotine content cigarettes have used total nicotine equivalents as a biomarker for detecting noncompliance, nicotine patches elevate nicotine levels regardless of compliance to very low nicotine content cigarettes; therefore, we cannot use total nicotine equivalents to identify noncompliant participants in CENIC-P1S2. Instead, we used the tobacco alkaloid anatabine which has a weaker association with compliance than total nicotine equivalents.[59,60] Due to the skewness of the distribution of anatabine, and the fact that it is a concentration, we modeled it on the natural logarithm scale. Additionally, some participants had observations of anatabine that fell below the limit of quantification and were thus censored. For these participants, we modified the likelihood of the conditional biomarker to accommodate left-censored observations.[61]

A total of $n = 114$ participants from CENIC-P1S2 had anatabine data for at least one of the $K = 6$ time points. We assumed that missing anatabine data was missing at random. To improve parameter estimation for the biomarker mixed effects model, we included information from a partial gold standard of participant compliance. Specifically, we included data from the study of 23 smokers who volunteered to be sequestered in a hotel for five nights with access only to very low nicotine content cigarettes.[50] We treat these anatabine data as coming from a single time point. Given that these participants were known to be compliant, their anatabine data no longer comes from a mixture density but can still be incorporated into the biomarker log likelihood. Moreover, although these anatabine data are not longitudinal, we can include them by assuming a univariate biomarker distribution of $N(\mu = \alpha_0 + \alpha_1, \Sigma = \sigma^2 + \tau^2)$.

In the model fit, we sought to determine whether or not compliance changed across time points. We compared the BIC of the models with a single compliance intercept vs a compliance intercept for each of the six time points. Additionally, to better understand the relationship between the longitudinal anatabine values and the compliance

probabilities, we constructed a spaghetti plot of log(anatabine) shaded by the probability of compliance at all time points. Similar to the simulation study, we compared $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i6})$ to $\hat{P}(C_{i6} = 1|B_{i6})$ according to AUC. As true compliance is unknown in CENIC-P1S2, we used a model-based approach to estimate AUC. Specifically, we generated a test set based off the parameter estimates and then estimated the AUCs of both $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i6})$ and $\hat{P}(C_{i6} = 1|B_{i6})$. Relative to generating an independent test set as in the simulation, this approach only overestimates the AUC by approximately 0.005.

To assess $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i5})$, we plotted the average estimated prediction probability of $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i5})$ against the average estimated compliance probability of $\hat{P}(C_{i6} = 1|B_{i6})$, which is known to accurately estimate compliance,[59] by quintile for the 97 participants with anatabine data at the last time point. Note that to estimate $P(C_{i6} = 1|B_{i1}, \ldots, B_{i5})$ we used data from only the first five time points.

3.3 Results

3.3.1 Simulation Study

For the simulation scenarios where $n = 100$ and $\alpha_1 \in \{-1.5, -1.8, -2.1\}$, Table 3.1 displays the bias and Monte Carlo standard deviation (MC SD) of the parameter estimates of both the compliance and biomarker mixed effects models. When the covariance structures were correctly specified, the parameters estimates were approximately unbiased across $\alpha_1 \in \{-1.5, -1.8, -2.1\}$ with low MC SDs. Under incorrect specification, the biomarker coefficient parameter estimates of $(\alpha_0, \alpha_1)^T$ were marginally biased, while the compliance coefficient parameter estimates of β_0 were unbiased. Both sets of parameter estimates had greater MC SDs relative to correct specification. Similar results were observed for both $K = 6$ and $K = 5$ time points. Supplementary Table A.1 displays the results of the simulation scenarios where $n = 500$ and $\alpha_1 \in \{-1.2, -1.5, -1.8, -2.1\}$.

Across $\alpha_1 \in \{-1.5, -1.8, -2.1\}$, Table 3.2 displays the AUC values for both $P(C_{i6} = 1|B_{i1}, \ldots, B_{i6})$ and $P(C_{i1} = 1, \ldots, C_{i6} = 1|B_{i1}, \ldots, B_{i6})$ under the true parameter values and estimated parameter values. Under correct specification for both compliance probabilities, estimating the parameters resulted in a marginal loss of AUC relative to

using the true values on average. Under incorrect specification, the loss of AUC was larger at about 3-5 percentage points on average.

Comparing the compliance probabilities with those from Boatman et al.'s method, we found that $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i6})$ had uniformly higher AUC than $\hat{P}(C_{i6} = 1|B_{i6})$. The gains in AUC were moderate under correct specification and marginal under incorrect specification. Similarly, for compliance at all time points, $\hat{P}(C_{i1} = 1, \ldots, C_{i6} = 1|B_{i1}, \ldots, B_{i6})$ had uniformly higher AUC than $\prod_{j=1}^{6} \hat{P}(C_{ij} = 1|B_{ij})$. Even with incorrect specification, the gains in AUC were about 5 percentage points on average (see Table 3.2).

For prediction of compliance at the last time point, Figure 3.1 displays the (gray) calibration lines of $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i5})$ corresponding to the sets of parameter estimates across $\alpha_1 \in \{-1.5, -1.8, -2.1\}$. The black lines mark the average calibration line and the 45 degree red lines mark perfect prediction. Under correct specification, the calibration lines fairly closely follow the red lines. More variability is introduced under incorrect specification. Similar to the other compliance probabilities, the AUC values of $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i5})$ closely approximated the true AUC values under correct specification. Under incorrect specification, there was a loss of about 7 percentage points of AUC on average.

Table 3.1: Estimated bias and Monte Carlo standard deviation (MC SD) of parameter estimates for $\boldsymbol{\theta}$ where $n = 100$, the number of time points $K \in \{6, 5\}$, and the compliance effect $\alpha_1 \in \{-1.5, -1.8, -2.1\}$. For both compliance and the conditional biomarker, the models fit are either correctly specified with compound symmetric covariance structures, or incorrectly specified where compound symmetric covariance structures are erroneously fit to data that is generated as AR(1).

		Compliance (C_{ij}) Parameters				Biomarker (B_{ij}) Parameters							
		Intercept		Random Intercept SD		Intercept		Compliance Effect		Random Intercept SD		Residual SD	
	Truth	$\beta_0 = 0$		$\gamma = 1$		$\alpha_0 = 4.0$		α_1		$\tau = 0.84$[a]		$\sigma = 0.55$[b]	
		$\alpha_1 = -1.5$											
$K = 6$	Bias[c]	-0.01	0	-0.01	-	-0.01	0.10	0	-0.19	-0.01	-	0	-
	MC SD[d]	0.30	0.71	0.34	-	0.14	0.22	0.11	0.13	0.09	-	0.04	-
$K = 5$	Bias	0.02	-0.05	0.04	-	0.01	0.07	0	-0.15	-0.01	-	0	-
	MC SD	0.59	0.90	0.57	-	0.17	0.23	0.14	0.17	0.10	-	0.05	-
		$\alpha_1 = -1.8$											
$K = 6$	Bias	-0.01	-0.01	-0.02	-	-0.01	0.05	0.01	-0.11	-0.01	-	0	-
	MC SD	0.19	0.33	0.18	-	0.12	0.16	0.08	0.10	0.08	-	0.03	-
$K = 5$	Bias	0	-0.01	-0.01	-	0	0.04	0	-0.08	-0.01	-	0	-
	MC SD	0.20	0.39	0.22	-	0.13	0.16	0.09	0.12	0.09	-	0.03	-
		$\alpha_1 = -2.1$											
$K = 6$	Bias	-0.01	-0.01	-0.01	-	-0.01	0.03	0.01	-0.07	-0.01	-	0	-
	MC SD	0.15	0.23	0.14	-	0.11	0.13	0.07	0.10	0.08	-	0.02	-
$K = 5$	Bias	0	-0.01	-0.01	-	0	0.02	0	-0.04	-0.01	-	0	-
	MC SD	0.16	0.26	0.17	-	0.12	0.14	0.08	0.10	0.08	-	0.03	-

[a] $\tau^2 = 0.7$.

[b] $\sigma^2 = 0.3$.

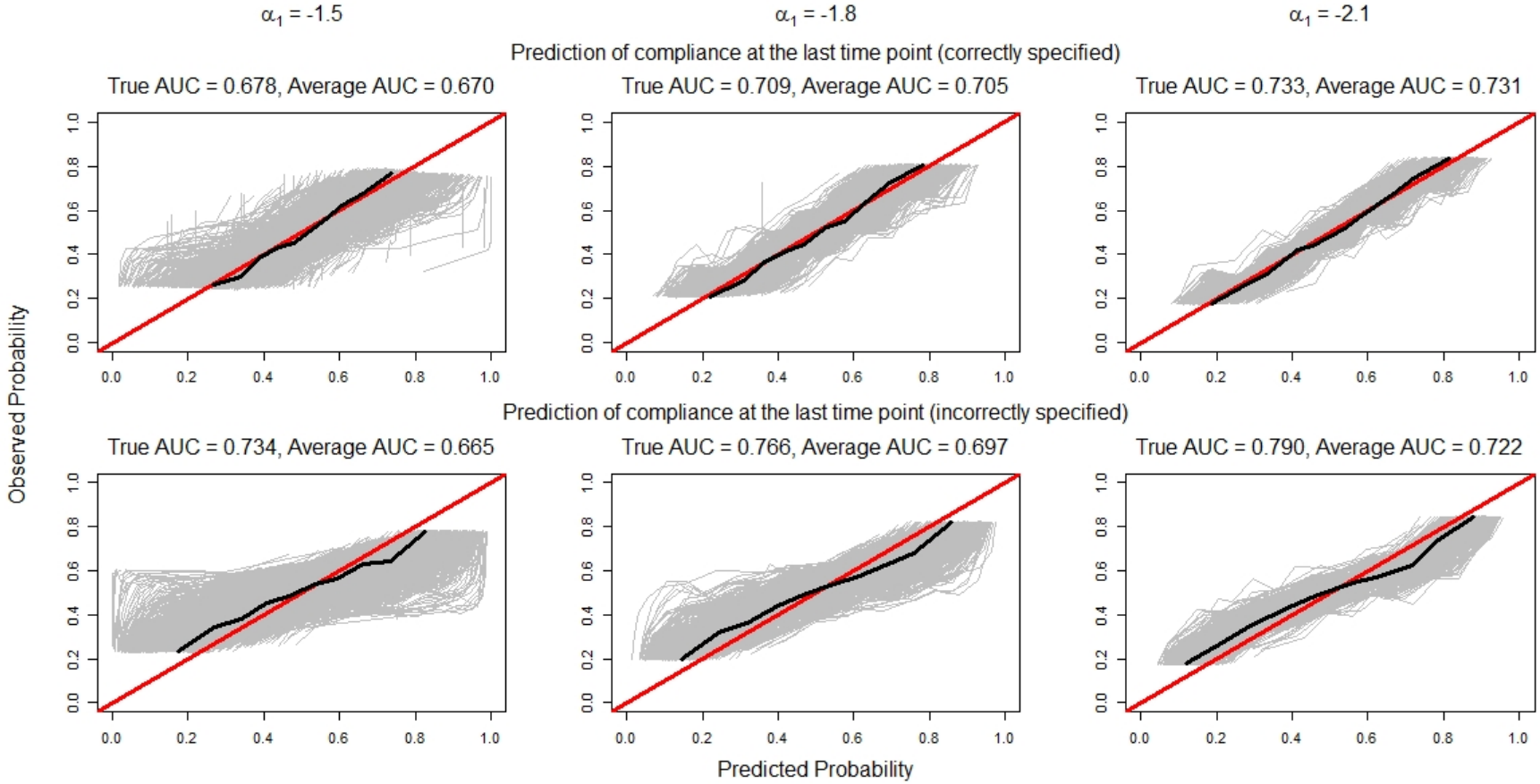
[c] The left (right) column of each parameter is the bias under the model with the correctly (incorrectly) specified covariance structure, where applicable.

[d] The left (right) column of each parameter is the MC SD under the model with the correctly (incorrectly) specified covariance structure, where applicable.

Table 3.2: True AUC values for $P(C_{i6} = 1|B_{i1}, \ldots, B_{i6})$ and $P(C_{i1} = 1, \ldots, C_{i6} = 1|B_{i1}, \ldots, B_{i6})$, and average AUC values for $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i6})$, $\hat{P}(C_{i1} = 1, \ldots, C_{i6} = 1|B_{i1}, \ldots, B_{i6})$, $\hat{P}(C_{i6} = 1|B_{i6})$, and $\prod_{j=1}^{6} \hat{P}(C_{ij} = 1|B_{ij})$, where the compliance effect $\alpha_1 \in \{-1.5, -1.8, -2.1\}$. For both compliance and the conditional biomarker, the models fit are either correctly specified with compound symmetric covariance structures, or incorrectly specified where compound symmetric covariance structures are erroneously fit to data that is generated as AR(1).

	Compliance effect α_1		
	-1.5	-1.8	-2.1
Compliance at the last time point (correctly specified)			
$P(C_{i6} = 1\|B_{i1}, \ldots, B_{i6})$	0.909	0.950	0.976
$\hat{P}(C_{i6} = 1\|B_{i1}, \ldots, B_{i6})$	0.903	0.948	0.975
$\hat{P}(C_{i6} = 1\|B_{i6})$	0.859	0.901	0.934
Compliance at the last time point (incorrectly specified)			
$P(C_{i6} = 1\|B_{i1}, \ldots, B_{i6})$	0.891	0.938	0.969
$\hat{P}(C_{i6} = 1\|B_{i1}, \ldots, B_{i6})$	0.859	0.909	0.944
$\hat{P}(C_{i6} = 1\|B_{i6})$	0.855	0.898	0.931
Compliance at all time points (correctly specified)			
$P(C_{i1} = 1, \ldots, C_{i6} = 1\|B_{i1}, \ldots, B_{i6})$	0.896	0.943	0.971
$\hat{P}(C_{i1} = 1, \ldots, C_{i6} = 1\|B_{i1}, \ldots, B_{i6})$	0.892	0.941	0.970
$\prod_{j=1}^{6} \hat{P}(C_{ij} = 1\|B_{ij})$	0.784	0.834	0.880
Compliance at all time points (incorrectly specified)			
$P(C_{i1} = 1, \ldots, C_{i6} = 1\|B_{i1}, \ldots, B_{i6})$	0.906	0.949	0.974
$\hat{P}(C_{i1} = 1, \ldots, C_{i6} = 1\|B_{i1}, \ldots, B_{i6})$	0.853	0.907	0.944
$\prod_{j=1}^{6} \hat{P}(C_{ij} = 1\|B_{ij})$	0.801	0.852	0.896

Figure 3.1: Calibration lines (in gray) for $\hat{P}(C_{i6}|B_{i1},\ldots,B_{i5})$ corresponding to the different sets of parameter estimates where the compliance effect $\alpha_1 \in \{-1.5,-1.8,-2.1\}$. The black lines mark the average calibration line and the 45 degree red lines mark perfect prediction. For both compliance and the conditional biomarker, the models fit are either correctly specified with compound symmetric covariance structures, or incorrectly specified where compound symmetric covariance structures are erroneously fit to data that is generated as AR(1).



3.3.2 Application to CENIC-P1S2

The model with a single compliance intercept for all time points returned a lower BIC (1882.35 vs 1902.72) compared to the model with a different compliance intercept for each time point; therefore, we assumed the more parsimonious model. Table 3.3 displays the parameter estimates and corresponding standard errors for $\boldsymbol{\theta}$ using all six time points and the first five time points for the $n = 114$ participants of CENIC-P1S2. For all six time points, compliance status was found to be highly autocorrelated with $\hat{\gamma} = 2.90$. We also found that compliance had a substantial standardized effect on the biomarker with $\frac{\hat{\alpha}_1}{\sqrt{\hat{\tau}^2+\hat{\sigma}^2}} = \frac{-2.36}{0.98^2+0.72^2} = -1.94$. The conditional biomarker was fairly autocorrelated with $\hat{\tau} = 0.98$.

Figure 3.2 displays the spaghetti plot of log(anatabine) for the time points when participants had data for CENIC-P1S2. For the anatabine values, we find that consistency is as important as magnitude in determining the probabilities of compliance at all time points. Our method assigned high probabilities to participants who always had relatively low anatabine values. Conversely, our method assigned low probabilities to participants who had low anatabine values at some time points but upticks at others.

For compliance at the last time point, the AUC of $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i6})$ was 0.943 while the AUC of $\hat{P}(C_{i6} = 1|B_{i6})$ was 0.916, a gain of 0.027 AUC. For prediction of compliance at the last time point, Figure 3.3 compares $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i5})$ to $\hat{P}(C_{i6} = 1|B_{i6})$. The green lines mark the bounds of the 95% prediction interval and the 45 degree red line marks perfect prediction. The line of quintiles fairly follows the red line, indicating that the probabilities of $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i5})$ somewhat accurately approximate the probabilities of $\hat{P}(C_{i6} = 1|B_{i6})$.

Table 3.3: Parameter estimates and standard errors (SE) for $\boldsymbol{\theta}$ using all six time points ($n = 114$) and the first five time points ($n = 114$) of CENIC-P1S2 for participants who had at least one anatabine value.

	Compliance (C_{ij}) Parameters				Biomarker (B_{ij}) Parameters							
	Intercept		Random Intercept SD		Intercept		Compliance Effect		Random Intercept SD		Residual SD	
	$\hat{\beta}_0$	SE($\hat{\beta}_0$)	$\hat{\gamma}$	SE($\hat{\gamma}$)	$\hat{\alpha}_0$	SE($\hat{\alpha}_0$)	$\hat{\alpha}_1$	SE($\hat{\alpha}_1$)	$\hat{\tau}$	SE($\hat{\tau}$)	$\hat{\sigma}$	SE($\hat{\sigma}$)
$K = 6$	−0.36	0.83	2.90	0.81	−3.56	0.27	−2.36	0.15	0.98	0.14	0.72	0.04
$K = 5$	0.15	0.86	3.54	1.18	−3.42	0.25	−2.34	0.17	0.88	0.13	0.70	0.05

Figure 3.2: Spaghetti plot of the longitudinal biomarker for the time points when participants had data for CENIC-P1S2. Lines are shaded by the probability of compliance at all time points. Lighter lines indicate a probability closer to 0 while darker lines indicate a probability closer to 1. The x-axis refers to the time points when participants had data. For example, a participant with data only at time points 1 and 4 would have a line from 1 to 2.

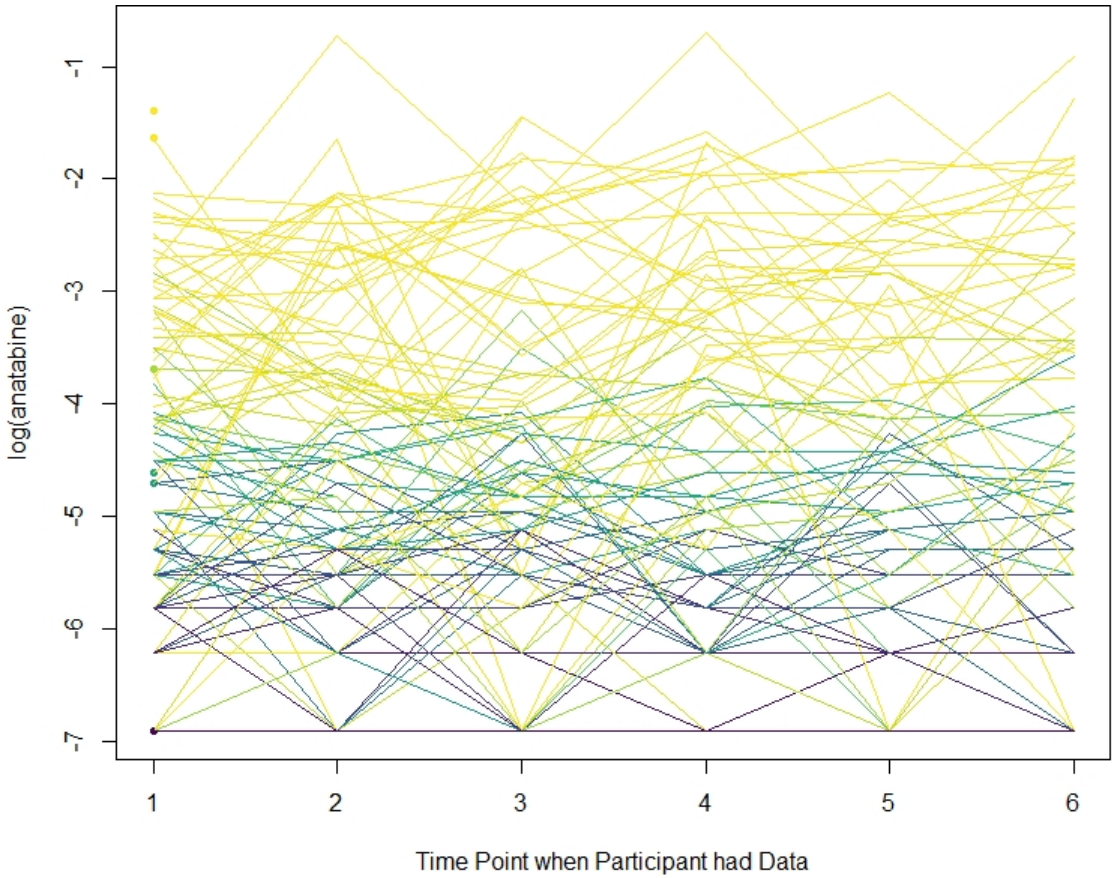
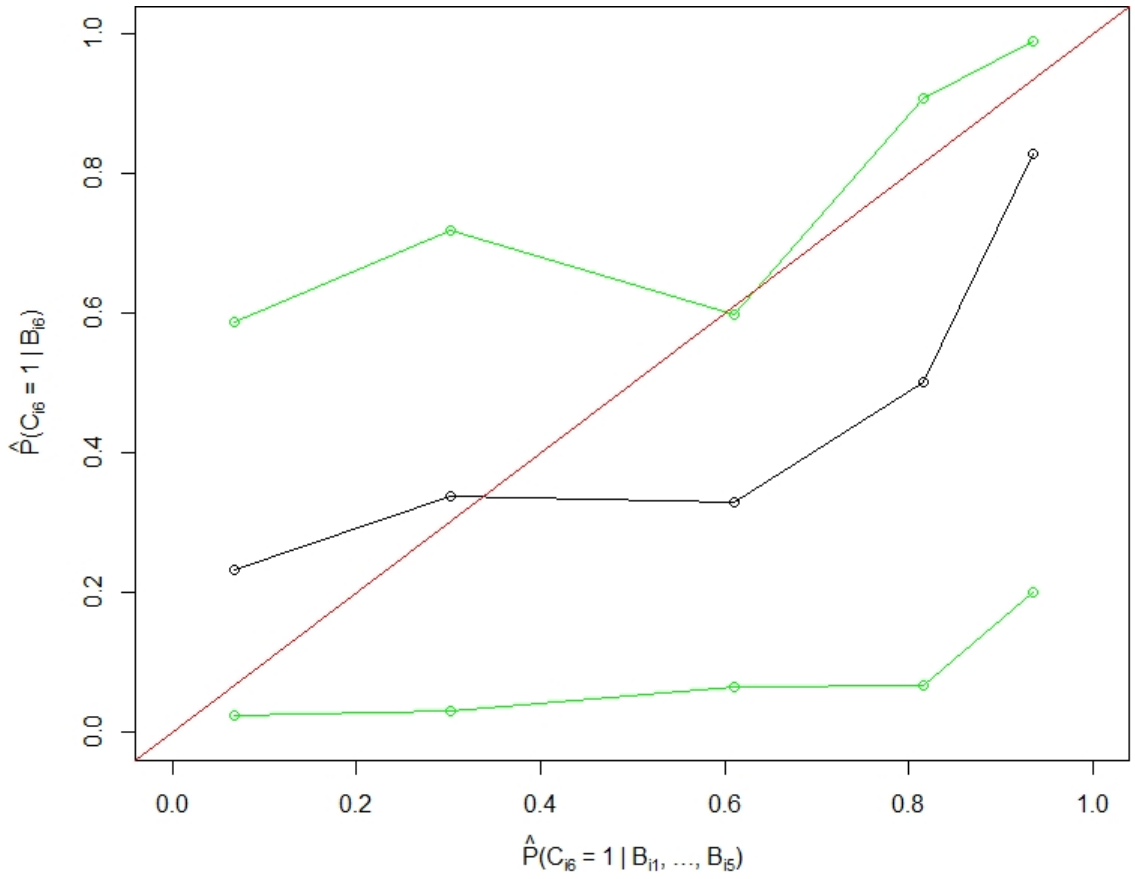


Figure 3.3: Averaged estimated prediction probability of $\hat{P}(C_{i6} = 1|B_{i1}, \ldots, B_{i5})$ plotted against the average estimated compliance probability of $\hat{P}(C_{i6} = 1|B_{i6})$ by quintile for the 97 participants of CENIC-P1S2 who had anatabine data at the last time point. The 45 degree red line marks perfect prediction and the green lines mark the bounds of the 95% prediction interval.



3.4 Discussion

For trials in which there is no known gold standard for measuring compliance, we have developed a method that can use longitudinal biomarker data to more accurately identify compliant participants, as well as predict which participants may or may not

comply in the future. Our simulation study confirms that our method, which conditions on participants' full biomarker history, is better able to discriminate between compliers and noncompliers compared to Boatman et al.'s method which conditions only on the most recent biomarker value. We also find that our prediction of compliance at a future time point conditional on biomarker data from previous time points is well-calibrated. Furthermore, when the covariance structures of both models for compliance and the biomarker are misspecified, we find that our method still has high discrimination. These results held across a range of differences in the biomarker between compliant and noncompliant participants.

In the application to the biomarker data collected from the $n = 114$ participants of CENIC-P1S2, we showed that our method was able to improve upon Boatman et al.'s method in estimating compliance at the last time point. Additionally, the prediction probability was able to somewhat accurately predict compliance at the last time point. Though we only investigated a subset of compliance patterns in this paper, our method can be used to estimate the compliance probability for any compliance pattern given observed biomarker data.

As demonstrated with the inclusion of the non-longitudinal hotel dataset, our method can include information from a partial gold standard. When compliance is known at all time points for a subset of participants, both longitudinal and non-longitudinal biomarker data can be incorporated into the biomarker log likelihood with the corresponding observed compliance patterns. Moreover, it is straightforward to adapt our method when compliance is known at some time points for a subset of participants. As only a subset of compliance patterns would be plausible, a mixture density with fewer compliance patterns can be fit for these participants.

For completed trials, our method can be used to improve estimation of causal effects. For example, the probability of compliance at all time points can be used to modify an inverse probability of compliance weight to more accurately weight the study outcomes, similar to Boatman et al.[42] For trials in progress, the prediction probability can be used to identify noncompliant participants to enable remedial intervention. This could both improve study outcomes and reduce the number of dropouts, the latter of which may also reduce the sample size at the end of the trial. That is, with fewer dropouts, fewer

additional participants may be needed to maintain statistical power.

Though our method belongs to the broader class of growth mixture models commonly found in the psychometrics literature,[62] there are some key differences that distinguish our method. Specifically, the number of latent groups we considered ($2^6 = 64$) is substantially greater than in more traditional applications where 2-4 latent groups are more common. When the number of latent groups is small, the (marginal) probability of group membership is generally treated as a free parameter and the longitudinal data trajectory within each group is generally unconstrained. Given the large number of latent groups in our study, however, we constrained the probabilities of the latent groups by assuming mixed effects models for both compliance and the biomarker. Under this relatively simple modeling structure, the number of parameters is kept small relative to the number of latent groups.

Although we considered relatively straightforward models for both compliance and the biomarker, our method can be extended to fit richer longitudinal models with more parameters. Our method can also be adapted to a measure of compliance consisting of more than two levels, but substantially more computational power would be required to fit the models. That is, with more levels of compliance the number of compliance patterns would exponentially increase.

Our method does have some limitations. First, although mixed effects models can handle data which are missing at random, this assumption may not be realistic when studying compliance. Second, we have defined compliance to be a discrete random variable. Adapting our method to a continuous measure of compliance is not straightforward. Third, the compliance effect on the biomarker must be substantial to be reliably detected and larger samples are required to detect smaller compliance effects. Fourth, although likelihood-based tests and information criteria can help identify the optimal model structure, we have not developed model diagnostics specific to our method. Finally, fitting mixed effects models to both observed and unobserved variables is computationally intensive and parameter estimates may be sensitive to initial starting values.

In future clinical trials where a biomarker is known to respond to deviations from the treatment assignment, our method can be used to identify noncompliant participants

both during and after the trial. The mode of noncompliance is immaterial; provided that noncompliance has a systematic effect on the biomarker, our method can be used for both participants who do not take any treatments and those who take alternatives. With this novel statistical method, trial designers should be better able to both adjust for noncompliance and prevent it from happening.

Chapter 4

Estimating Longitudinal Causal Effects with Unobserved Participant Noncompliance and Confounders with Densities that May Be Difficult to Parametrically Specify

4.1 Introduction

Randomized controlled trials (RCTs) are viewed as the gold standard for testing the efficacy of new treatments in target populations of interest. To establish causation due to the treatment, RCTs depend on properly randomizing participants to different treatment arms. In some trials, however, participants are not obligated to comply with their randomized treatment assignment. Noncompliance most often occurs when participants must self-administer the treatment without the supervision of study personnel or a medical practitioner.[32] Participants may then choose to either comply or deviate

from the treatment protocol for a variety of reasons (e.g., due to advent of side effects, insufficient benefit, or availability of commercial alternatives to the treatment).[33] Additionally, in longitudinal studies, multiple time points present multiple opportunities for participants to be noncompliant.

When participants do not comply, the treatment received may be different from the treatment randomly assigned. Primary analyses of such trials must then be qualified with the consideration that not all participants complied with the trial. Specifically, both the intention-to-treat (ITT) estimate and the Per Protocol estimate based on self-reported compliance may not match the treatment effect if all participants had complied (i.e., the causal effect).[39] For the ITT estimate, noncompliant participants do not receive the full dose of treatment and as such may have poor study outcomes. This may dilute the ITT estimate.[37] For the Per Protocol estimate, participants who self-reported compliance may systematically differ on some confounding variable.[38]

For treatments that could be mandated by federal law to effectively force compliance, the causal effect is more relevant than the ITT estimator.[40] As a motivating example, in the United States, the Family Smoking Prevention and Tobacco Control Act provides the Food and Drug Administration with the regulatory authority to reduce (but not eliminate) the nicotine content in commercial cigarettes if it would improve public health. Regulatory tobacco trials seek to evaluate the effect of nicotine reduction on measures of smoking behavior. In particular, we were motivated by the Center for the Evaluation of Nicotine in Cigarettes Project 2 (CENIC-P2) trial, a 20-week RCT that investigated the effect of very low nicotine content (VLNC, with 0.4mg of nicotine per gram of tobacco) cigarettes versus normal nicotine content (NNC, with 15.5mg of nicotine per gram of tobacco) cigarettes on average number of cigarettes smoked per day. During the follow-up period, participants were asked to only smoke their assigned study cigarettes but could additionally smoke commercial cigarettes (i.e., non-study cigarettes) with normal nicotine content.

In longitudinal studies, as noncompliance, the study outcome, and the confounding variables can all vary over time, there may be time-varying confounding. In the presence of

both time-varying noncompliance and time-varying confounding, standard regression-based methods cannot be used to estimate the causal effect.[41] Longitudinal causal estimators that can consistently estimate the causal effect of an intervention with time-varying noncompliance include the G-computation algorithm, inverse probability of compliance weighted (IPCW) estimators, and structural nested mean models estimated with G-estimation.[63–68]

For CENIC-P2, both IPCW estimators and G-estimation may be less than desirable. Specifically, for IPCW estimators, the denominator of the weights of compliance is a product of the compliance probabilities at each time point. As there are many time points for compliance in CENIC-P2 (i.e., 5 time points), the product of many compliance probabilities could create bias and variability in estimation of the weights and hence estimation of the causal effect.[69] Additionally, a primary advantage of G-estimation is that structural nested mean models can directly model the interactions between time-varying treatments and time-varying confounding.[70] This can help identify the optimal treatment regime; however, in CENIC-P2, we are only interested in modeling a single treatment regime (i.e., compliance across all time points of the trial). Thus, there are no time-varying treatments in the model fit for the study outcome, and hence no need to model any interactions that include them.

We thus turn our attention to the G-computation algorithm. To adjust for time-varying confounding, G-computation models both the time-varying confounders and the study outcome conditional on the treatment and confounder histories. The causal effect is then estimated by randomly sampling from the fitted densities.

Like all other longitudinal causal estimators, G-computation assumes that compliance is observed. However, in trials where participants must self-administer the treatment, compliance is not directly observed. Various methods have been devised to determine compliance when unobserved (e.g., self-reports), but most all depend on subjective or indirect information, and as such are unreliable.[44,45]

In some trials, participants' biomarkers may systematically change in response to the treatment or any alternatives to provide objective information about noncompliance. Consider three recently published regulatory tobacco clinical trials studying the effect

of VLNC cigarettes versus NNC cigarettes on smoking behavior.[34,48,49] At each study visit for each trial, investigators measured participants' biomarkers of nicotine exposure (e.g., total nicotine equivalents (TNE), which measures most nicotine metabolites in the urine) which can be used to detect noncompliance to VLNC cigarettes.[50,52,59] Without having to observe compliance, Boatman et al. (2017) used TNEs to re-weight an IPCW estimator to estimate the causal effect of VLNC cigarettes on average number of cigarettes smoked per day.[42] However, Boatman et al.'s method is limited to examining information collected at a single time point.

We will develop a G-computation estimator that uses biomarker information in place of unobserved compliance. The motivating dataset from CENIC-P2, however, presents one additional challenge for G-computation. The confounders include survey instruments whose supports are bounded with many observations near the ends of the scales. This makes it difficult to correctly specify parametric conditional densities. While we could fit fully parametric regression models, incorrectly specifying the conditional densities may lead to biased estimators. Instead, we will implement predictive mean matching, which requires that we only correctly specify the conditional mean as opposed to the full conditional distribution.

The goal of this research is thus two-fold: to develop a G-computation estimator that can 1) account for unobserved noncompliance; and 2) semiparametrically model confounders with densities that may be difficult to parametrically specify. We first evaluate our method by simulation, investigating the effects of both different levels of confounding and sample sizes on estimation of the causal mean. We also apply our method to the data collected in CENIC-P2.

4.2 Methods

4.2.1 Dataset and Target of Inference

We consider a longitudinal RCT, where $i \in \{1, \ldots, n\}$ denotes the participant, $j \in \{0, \ldots, K\}$ denotes the time point, and $A_i = a_i$ denotes the randomized treatment group of participant i. For our purposes, we treat the true compliance status as unobserved and binary; that is, each participant can either fully comply or not fully comply at

each time point. Let C_{ij} be the true, unobserved compliance indicator (1 denotes compliance, 0 denotes noncompliance) for the ith participant at the jth time point for $j \in \{1, \ldots, K\}$. As participant compliance can vary over time, let $\bar{c}_{iK}$ be one of the 2^K compliance patterns for K time points for participant i. For example, for $K = 2$ time points, the set of compliance patterns has length $2^2 = 4$ and consists of: $\{\{0,0\},\{1,0\},\{0,1\},\{1,1\}\}$. We use the overbar notation to denote history (e.g., for full compliance, $\bar{c}_{iK} = \{1, \ldots, 1\}$).

Define $Y^*_{iK}(a_i, \bar{c}_{iK})$ to be the study outcome at the end of the trial of a randomly selected participant if, possibly contrary to fact, we set $A_i = a_i$ and $\bar{C}_{iK} = \bar{c}_{iK}$. Because for each participant we do not observe $Y^*_{iK}(a_i, \bar{c}_{iK})$ for all a_i and $\bar{c}_{iK}$, $Y^*_{iK}(a_i, \bar{c}_{iK})$ is a potential outcome. The target of inference is the expected difference in the outcome among randomized treatment groups a_i and a'_i if all participants were to be compliant. That is:

$$E(Y^*_{iK}(a_i, \bar{c}_{iK} = \{1, \ldots, 1\}) - Y^*_{iK}(a'_i, \bar{c}_{iK} = \{1, \ldots, 1\})).$$

Define $L_{ij} = \{Y_{ij}, Z_{ij}\}$ as the set of time-varying confounders for $j \in \{0, \ldots, K\}$ with history $\bar{L}_{ij} = \{L_{i0} \ldots, L_{ij}\}$, where Y_{ij} is the study outcome measured at both the end of the trial and at earlier time points and Z_{ij} is an additional set of confounders. Let $W_i = \{L^*_{i1}(a_i, c_{i1}), \ldots, L^*_{iK}(a_i, \bar{c}_{iK})\}$ be the set of potential outcomes for L_{ij} for the ith participant across all a_i and $\{c_{i1}, \ldots, \bar{c}_{iK}\}$. Additionally, let B_{ij} be the biomarker for $j \in \{1, \ldots, K\}$, let X_i be the set of time-invariant confounders, and D_{ij} be the self-reported indicator of compliance (1 denotes self-reported compliance, 0 denotes self-reported noncompliance) for $j \in \{1, \ldots, K\}$.

4.2.2 Identifying Assumptions

To estimate the expected value of the potential outcome based on the observed data, we make the following identifying assumptions.[69] First, we make the no unmeasured confounders assumption, where we assume that compliance status is only confounded by L_{ij} and X_i. Given this information, compliance status at each time point does not depend on the potential outcomes (i.e., $P(C_{ij} = 1|\bar{L}_{ij}, X_i, A_i = a_i, W_i) = P(C_{ij} = 1|\bar{L}_{ij}, X_i, A_i = a_i)$). Second, we make the positivity assumption, where we assume that there is a non-zero probability of compliance at each time point across all values of the

confounders (i.e., $P(C_{ij} = 1|\bar{L}_{ij}, X_i, A_i = a_i) > 0$). Third, we make the consistency assumption, where we assume that the mode of compliance is immaterial. That is, at each time point, a participant who voluntarily complies has the same outcome relative to if he or she were forced to be compliant (i.e., if $A_i = a_i$ and $\bar{C}_{ij} = \bar{c}_{ij}$, then $Y_{ij} = Y^*_{ij}(a_i, \bar{c}_j)$).

4.2.3 The G-computation Algorithm

We delineate the G-computation algorithm as follows. Under the identifying assumptions, the joint density of both the potential outcomes under full compliance (i.e., $\{L^*_{i1}(a_i, c_{i1} = 1), \ldots, L^*_{iK}(a_i, \bar{c}_{iK} = \{1, \ldots, 1\})\}$) and X_i is equal to:

$$\begin{aligned} &f(L_{iK}|\bar{L}_{iK-1}, X_i, A_i = a_i, \bar{C}_{iK} = \{1, \ldots, 1\}) \\ &* f(L_{iK-1}|\bar{L}_{iK-2}, X_i, A_i = a_i, \bar{C}_{iK-1} = \{1, \ldots, 1\}), \ldots, f(L_{i0}|X_i)f(X_i), \end{aligned} \tag{4.1}$$

where $f(L_{iK}|\bar{L}_{iK-1}, X_i, A_i = a_i, \bar{C}_{iK} = \{1, \ldots, 1\})$ is the conditional density of L_{iK} given $\bar{L}_{iK-1}, X_i, A_i = a_i, \bar{C}_{iK} = \{1, \ldots, 1\}$. Then the density of $Y^*_{iK}(a_i, \bar{c}_{iK} = \{1, \ldots, 1\})$ is given by:

$$\begin{aligned} \int \ldots \int & f(L_{iK}|\bar{L}_{iK-1}, X_i, A_i = a_i, \bar{C}_{iK} = \{1, \ldots, 1\}) f(L_{iK-1}|\bar{L}_{iK-2}, X_i, A_i = a_i, \bar{C}_{iK-1} = \{1, \ldots, 1\}) \\ & , \ldots, f(L_{i0}|X_i)f(X_i) dX_i dL_{i0} \ldots dL_{iK-1} dZ_{iK}. \end{aligned} \tag{4.2}$$

We can estimate $E(Y^*_{iK}(a_i, \bar{c}_{iK} = \{1, \ldots, 1\}))$ with three steps:

1. We model the densities of the baseline confounders, L_{i0} and X_i, often with the empirical distributions.
2. For each subsequent time point, we model the conditional confounder density of L_{ij} given past covariate history, often with parametric models.
   - Both the estimated densities for steps 1 and 2 can then be plugged into (4.1).
3. We use Monte Carlo integration to approximate the integrals in (4.2) to estimate the distribution of $Y^*_{iK}(a_i, \bar{c}_{iK} = \{1, \ldots, 1\})$ as well as any corresponding summary statistics (e.g., the mean).

We discuss steps 2 and 3 in detail. In the setting where the compliance indicator C_{ij} is known, we could assume a parametric model for the density of L_{ij} indexed by parameter

vector β_j for time point j. We could then estimate β_j by solving the following score equations:

$$\sum_{i=1}^{n} I(\bar{C}_{ij} = \{1, \ldots, 1\}) \frac{\partial}{\partial \beta_j} \log f(L_{ij} | \bar{L}_{ij-1}, X_i, A_i = a_i, \bar{C}_{ij} = \{1, \ldots, 1\}; \beta_j) = 0. \quad (4.3)$$

In many applications, we may be willing to assume that the conditional density of L_{ij} given $\bar{L}_{ij-1}, X_i, A_i = a_i, \bar{C}_{ij} = \{1, \ldots, 1\}$ equals the conditional density of L_{ij} given $L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1$ That is, the conditional density only depends on the most recent L_{ij-1} and compliance at the current time point, in addition to the time-invariant confounders and treatment arm. In that case, the score equations in (4.3) simplify to:

$$\sum_{i=1}^{n} I(C_{ij} = 1) \frac{\partial}{\partial \beta_j} \log f(L_{ij} | L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1; \beta_j) = 0.$$

Note that in some contexts we may be willing to assume that $\beta_1 = \ldots = \beta_K = \beta$ (i.e., the parameters are shared across time points). We could then estimate β by solving the following estimating equations:

$$\sum_{j=1}^{K} \sum_{i=1}^{n} I(C_{ij} = 1) \frac{\partial}{\partial \beta} \log f(L_{ij} | L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1; \beta) = 0. \quad (4.4)$$

As $L_{ij} = \{Y_{ij}, Z_{ij}\}$, it may be difficult to specify a multidimensional joint density. Instead, we can model the conditional density of L_{ij} by factorizing $f(L_{ij} | L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1; \beta)$ into individual components for Z_{ij} given $L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1$ and Y_{ij} given $Z_{ij}, L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1$. Then:

$$\begin{aligned} &f(L_{ij} | L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1; \beta) \\ &= h(Y_{ij} | Z_{ij}, L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1; \beta_Y) s(Z_{ij} | L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1; \beta_Z), \end{aligned} \quad (4.5)$$

where parameter vectors β_Y and β_Z correspond to the subsets of β for the conditional densities of Y_{ij} and Z_{ij}, respectively. We can further factorize the conditional density of Z_{ij} when Z_{ij} is multivariate.

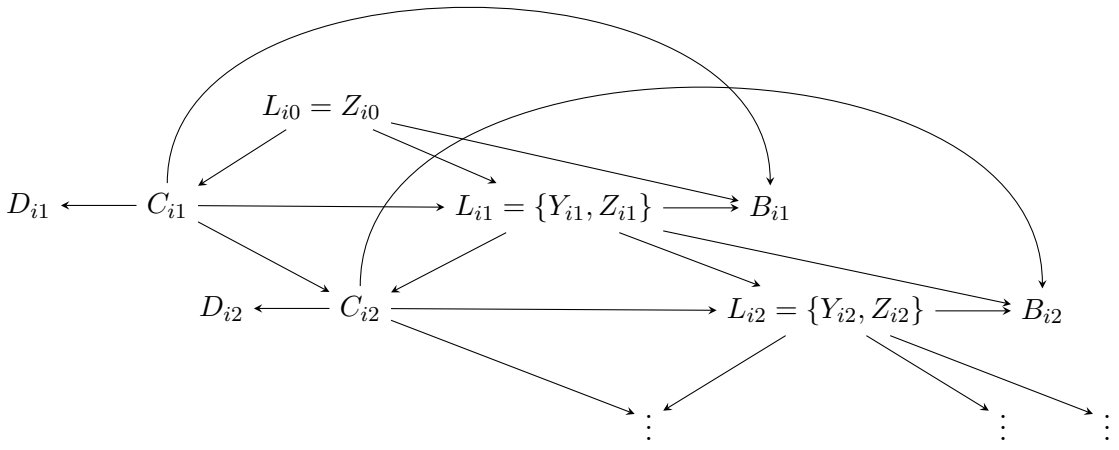
To implement Monte Carlo integration, we use the following steps for random samples $r \in \{1, \ldots, R\}$. First, we randomly sample both $l_{i0,r}$ and $x_{i,r}$ from the empirical distributions of L_{i0} and X_i. We then randomly sample $l_{i1,r}$ from the fitted density

$f(L_{i1}|L_{i0} = l_{i0,r}, X_i = x_{i,r}, A_i = a_i, C_{i1} = 1; \hat{\beta})$. For each subsequent time point, we randomly sample from $f(L_{ij}|L_{ij-1} = l_{ij-1,r}, X_i = x_{i,r}, A_i = a_i, C_{i1} = 1; \hat{\beta})$. The average of the randomly sampled $y_{iK,r}$ at the last time point is the causal estimator:

$$\hat{E}(Y^*_{iK}(a_i, \bar{c}_{iK} = \{1, \ldots, 1\})) = \frac{1}{R}\sum_{r=1}^{R} y_{iK,r}.$$

Given that we assume that observations only depend on data collected at the current and previous time points, Figure 4.1 details one plausible directed acyclic graph (DAG) for how the data might have been generated in CENIC-P2. The DAG assumes that self-reported compliance D_{ij} is conditionally independent of all other variables given compliance C_{ij}.

Figure 4.1: One plausible directed acyclic graph for CENIC-P2. The time-invariant confounders X_i affect all L_{ij}, C_{ij} and B_{ij}, but not D_{ij}.



4.2.4 The G-computation Algorithm with Unobserved Participant Noncompliance

CENIC-P2 presents two challenges for G-computation: 1) compliance is unknown; and 2) the confounders come from densities that may be difficult to parametrically specify. To resolve 1), we can use both the observed biomarker B_{ij} that systematically responds

to compliance C_{ij} and self-reported compliance D_{ij}. We insert probabilities of compliance conditional on the biomarker, self-reported compliance, time-varying confounders (measured at the current and previous time points), time-invariant confounders, and treatment arm in place of the indicator of compliance in (4.4):

$$\sum_{j=1}^{K}\sum_{i=1}^{n} P(C_{ij}=1|L_{ij},L_{ij-1},B_{ij},X_i,D_{ij},A_i=a_i)\log\frac{\partial}{\partial\beta}f(L_{ij}|L_{ij-1},X_i,A_i=a_i,C_{ij}=1;\beta)=0. \tag{4.6}$$

Moreover, by iterated expectation, the expected value of the estimating function in (4.4) is equal to the expected value of (4.6) (see Appendix C.1). Thus, (4.6) is also a mean-zero estimating function. Therefore, under suitable regularity conditions, estimators of β from solving (4.6) are consistent and asymptotically known.

Additionally, in most trials there is usually no incentive to report noncompliance. Therefore, we typically assume that self-reports of noncompliance are accurate, where $D_{ij}=0$ implies $C_{ij}=0$ (i.e., $P(C_{ij}=1|L_{ij},L_{ij-1},B_{ij},X_i,D_{ij}=0,A_i=a_i)=0$). Observations with $D_{ij}=0$ thus do not contribute to parameter estimation.

Although the probability of compliance (i.e., $P(C_{ij}=1|L_{ij},L_{ij-1},B_{ij},X_i,D_{ij},A_i=a_i)$) is unknown and cannot be directly estimated from the data (e.g., by specifying a logistic regression model with C_{ij} as the response variable), we can derive a consistent estimator as a function of directly estimable densities. By Bayes' rule, we can write the probability of compliance as a function of the probability without the biomarker and the biomarker conditional on compliance. Specifically, $P(C_{ij}=1|L_{ij}=l_{ij},L_{ij-1}=l_{ij-1},B_{ij}=b_{ij},X_i=x_i,D_{ij},A_i=a_i)$ can be expressed as:

$$\frac{\rho(l_{ij},l_{ij-1},x_i,a_i,d_{ij};\alpha)g(b_{ij}|l_{ij},l_{ij-1},x_i,a_i,d_{ij},c_{ij}=1;\xi)}{\psi(b_{ij}|l_{ij},l_{ij-1},x_i,a_i,d_{ij};\alpha,\xi)},$$

where $\rho(l_{ij},l_{ij-1},x_i,a_i,d_{ij};\alpha)=P(C_{ij}=1|L_{ij}=l_{ij},L_{ij-1}=l_{ij-1},X_i=x_i,A_i=a_i,D_{ij};\alpha)$ indexed by parameter vector α, g is the conditional density of B_{ij} indexed by parameter vector ξ, and the denominator is equal to:

$$\begin{aligned}&\psi(b_{ij}|l_{ij},l_{ij-1},x_i,a_i,d_{ij},\alpha,\xi)\\&=\rho(l_{ij},l_{ij-1},x_i,a_i,d_{ij};\alpha)g(b_{ij}|l_{ij},l_{ij-1},x_i,a_i,d_{ij},c_{ij}=1,\xi)\\&+(1-\rho(l_{ij},l_{ij-1},x_i,a_i,d_{ij};\alpha))g(b_{ij}|l_{ij},l_{ij-1},x_i,a_i,d_{ij},c_{ij}=0;\xi).\end{aligned} \tag{4.7}$$

As all variables but C_{ij} are observed, we can directly estimate the components of $P(C_{ij}=1|L_{ij},L_{ij-1},B_{ij},X_i,A_i=a_i,D_{ij})$ from the mixture density in (4.7). As is

often the case with mixture densities such as (4.7), it would be difficult to directly solve for the maximum likelihood estimates. Instead, we can use the EM algorithm.[57] Appendix C.2 delineates the E-step and M-step.

4.2.5 The G-computation Algorithm with Predictive Mean Matching

As it may not be reasonable to fit parametric models to the densities of the confounders in (4.5), we instead fit semiparametric models using predictive mean matching.[71,72] For now, assume that the conditional density of Y_{ij} can be correctly specified using parametric models while the conditional density of Z_{ij} may not be correctly specified.

While we may not be willing to assume a fully parametric density (e.g., a normal distribution) for $s(Z_{ij}|L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1; \beta_Z)$, we can instead assume that only the conditional mean (i.e., $E(Z_{ij}|L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1; \beta_Z)$) can be parametrically specified where β_Z solely consists of regression coefficients. We can consistently estimate β_Z by solving a mean-zero estimating function such as the following:

$$\sum_{j=1}^{K}\sum_{i=1}^{n} P(C_{ij} = 1|L_{ij}, L_{ij-1}, B_{ij}, X_i, A_i = a_i, D_{ij}) \\ * \{Z_{ij} - E(Z_{ij}|L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1; \beta_Z)\}\frac{\partial}{\partial \beta_Z} E(Z_{ij}|L_{ij-1}, X_i, A_i = a_i, C_{ij} = 1; \beta_Z) = 0.$$

The residual distribution of Z_{ij}, however, may not be well-approximated by a parametric distribution. Specifically, for CENIC-P2 Z_{ij} comprises survey data with bounded supports. This poses a problem when we have to randomly sample from the conditional Z_{ij} distribution for Monte Carlo integration, as the generated Z_{ij} must fall within the survey bounds. Predictive mean matching can bring the data generation closer to randomly sampling from the true density.

We delineate predictive mean matching as follows. At each time point, we first calculate the predicted values of the generated Z_{ij} data and match them with the predicted values of the observed Z_{ij} data across all time points. Then, we impute one of the corresponding observed Z_{ij}. Specifically, let the dot symbol $\dot{}$ denote the predicted value for the observed data and the hat symbol $\hat{}$ denote the predicted value for the generated data. For random samples $r \in \{1, \ldots, R\}$ with time points $j \in \{0, \ldots, K\}$, we can generate confounder data with the following steps:

1. Randomly sample baseline confounders $l_{i0,r}$ and $x_{i,r}$ from the empirical distributions of L_{i0} and X_i, respectively.
2. For each subsequent time point, calculate $\hat{Z}_{ij,r}$ from the fitted $\hat{E}(Z_{ij}|L_{ij-1} = l_{ij-1,r}, X_i = x_{i,r}, A_i = a_i, C_{ij} = 1; \hat{\beta}_Z)$.
3. Find the top five candidates of $\dot{Z}_i$ across all time points for which $|\dot{Z}_i - \hat{Z}_{ij,r}|$ is minimal.
4. Randomly select one $\dot{Z}_i$ and impute its corresponding observed z_i for $z_{ij,r}$.
5. Randomly sample $y_{ij,r}$ from the fitted $h(Y_{ij}|Z_{ij} = z_{ij,r}, L_{ij-1} = l_{ij-1,r}, X_i = x_{i,r}, A_i = a_i, C_{ij} = 1; \hat{\beta}_1)$.
6. This generates $l_{ij,r} = \{y_{ij,r}, z_{ij,r}\}$.

This approach can easily be generalized to the case where Z_{ij} contains multiple confounders and the density of Y_{ij} may not be correctly specified.

Randomly sampling from some number of candidates according to proximity of predictive means is known as predictive mean random hot deck.[73] The selection criteria for matching is flexible; instead of the top five candidates, candidates could be selected based on some fixed threshold η for $|\dot{Z}_i - \hat{Z}_{ij,r}| < \eta$ or the observed value for the top candidate could automatically be imputed. All of these approaches have been found to be asymptotically unbiased estimators of the true density.[74] We chose the top five candidates for two reasons: 1) relative to selecting candidates based on a fixed threshold η, the top five candidates ensure that there are always a limited number of candidates to draw from to expedite computation; and 2) relative to only selecting the top candidate, the top five candidates should provide more variability to protect against overfitting the empirical distributions.

4.3 Simulation Study Design

4.3.1 Data Generation

We used Monte Carlo simulation to assess the small sample properties of our estimator assuming a single treatment arm a_i. Across different simulation scenarios, we varied

both the sample size and the effect size of confounding on the study outcome Y_{ij}.

Across all scenarios, we assumed there to be single confounders for both Z_{ij} and X_i. We generated longitudinal data in accordance with the DAG in Figure 4.1, where at each time point data is generated from other data collected at the current or previous time points. At baseline, X_i is generated from a normal distribution while the study outcome $Y_{i0} = 0$. To mimic bounded confounder data, we generated Z_{ij} from a zero-inflated poisson distribution. We then used linear regression to estimate the coefficients for the conditional mean of Z_{ij}, after which we used predictive mean matching to randomly sample the confounders in G-computation. Compliance C_{ij} is generated from a bernoulli distribution with a logit link — however, we were not able to generate compliance data in congruence with the density of compliance in (4.7). The fitted logistic model is thus misspecified. We assumed that compliance varies across time points with a different intercept at each time point. Both Y_{ij} and B_{ij} were generated from normal distributions which can be correctly specified.

For the indicator of self-reported compliance at each time point, D_{ij}, we assumed that compliers honestly reported compliance and that noncompliers dishonestly reported compliance $\frac{2}{3}$ of the time (i.e., $D_{ij} = T^{1-C_{ij}}$, where $T \sim Bern(\frac{2}{3})$). Observations with $D_{ij} = 0$ did not contribute to parameter estimation.

We set the number of time points at $K = 6$ where $j \in \{0, \ldots, 5\}$, analogous to CENIC-P2. We ran simulations for two sample sizes: $n = 500$ and $n = 1000$. Across all scenarios, data were generated such that over all participants and time points, approximately 40% of observations were compliant while 60% were noncompliant. After filtering out observations with self-reported noncompliance, approximately 80% of observations equally distributed between compliance and noncompliance contributed to parameter estimation. Data were generated such that the percentage of compliance marginally increased with each time point.

The following statistics for the simulation only relate to self-reported compliers who we intend to analyze. To investigate the effect of confounding on estimation of the causal mean, we varied the size of R^2 in the model for the study outcome Y_{ij} in (4.5). We simulated three values of R^2: 0.3, 0.5, and 0.7. Note that as confounding increases

for Y_{ij}, the discrimination of the compliance probability in (4.6) increases. For $R^2 \in \{0.3, 0.5, 0.7\}$, the areas under the curve (AUC) of the compliance probabilities were $\{0.927, 0.952, 0.980\}$, respectively. The AUC values are within the range of those found in previous CENIC trials.[59] For Z_{ij} over all time points, 8% of observations were 0's while the mean of the poisson component was 33.

To evaluate each estimator, we calculated the empirical bias, Monte Carlo standard deviation (MC SD), and mean-squared error (MSE) across the estimates of the causal mean. Code to implement our method in the programming language R is available as a GitHub repository (https://github.com/RPeterson4/Modified_G_computation).

4.3.2 Proposed Estimators

In both the simulation and application, we sought to compare a number of estimators of $E(Y^*_{iK}(a_i, \bar{c}_{iK} = \{1, \ldots, 1\}))$. First, we have estimators already established in the literature, including the conventional ITT and Per Protocol estimators, the latter of which is based off of self-reported compliance. We also consider a regression-based estimator from Boatman et. al (2018), the EM-REG estimator.[43] Instead of the mixture density of the biomarker in (4.7), we could model the mixture density of the joint biomarker and study outcome by factorization:

$$\begin{aligned}
&\lambda(b_{ij}, y_{ij} | z_{ij}, l_{ij-1}, x_i, a_i, d_{ij}; \delta, \xi, \beta_Y) \\
&= \omega(z_{ij}, l_{ij-1}, x_i, a_i, d_{ij}; \delta) g(b_{ij} | y_{ij}, z_{ij}, l_{ij-1}, x_i, a_i, d_{ij}, c_{ij} = 1; \xi) h(y_{ij} | z_{ij}, l_{ij-1}, x_i, a_i, d_{ij}, c_{ij} = 1; \beta_Y) \\
&+ (1 - \omega(z_{ij}, l_{ij-1}, x_i, a_i, d_{ij}; \delta)) g(b_{ij} | y_{ij}, z_{ij}, l_{ij-1}, x_i, a_i, d_{ij}, c_{ij} = 0; \xi) h(y_{ij} | z_{ij}, l_{ij-1}, x_i, a_i, d_{ij}, c_{ij} = 0; \beta_Y),
\end{aligned}$$

where $\omega(z_{ij}, l_{ij-1}, x_i, a_i, d_{ij}; \delta) = P(C_{ij} = 1 | Z_{ij} = Z_{ij}, L_{ij-1} = l_{ij-1}, X_i = x_i, A_i = a_i, D_{ij}; \delta)$ indexed by parameter vector δ. Note that both $g(b_{ij} | y_{ij}, z_{ij}, l_{ij-1}, x_i, a_i, d_{ij}, c_{ij} = 1; \xi)$ and $h(y_{ij} | z_{ij}, l_{ij-1}, x_i, a_i, d_{ij}, c_{ij} = 1; \beta_Y)$ are the same densities as in (4.7) and (4.5), respectively. We assume the same link functions for the models of C_{ij}, B_{ij}, and Y_{ij} as before with G-computation. Again, we use the EM algorithm to derive the maximum likelihood estimates of the mixture density. We can estimate $E(Y^*_{iK}(a_i, \bar{c}_{iK} = \{1, \ldots, 1\}))$ as the estimated conditional expected value of Y_{iK} at the last time point:

$$\frac{1}{n} \sum_{i=1}^{n} \hat{E}(Y_{iK} = y_{ik} | Z_{iK} = z_{iK}, L_{iK-1} = l_{iK-1}, X_i = x_i, A_i = a_i, C_{iK} = 1; \beta_Y).$$

We now introduce causal estimators based off of our proposed G-computation algorithm. To illustrate the efficiency gained from implementing the full algorithm, we derive truncated G-computation estimators that ignore the two modifications we made to the algorithm. First, we have an estimator that does not use predictive mean matching and instead directly samples confounders from normal densities erroneously fit to (4.5). Second, we have an estimator that treats self-reported compliance as accurate by substituting in $I(D_{ij} = 1)$ for $I(C_{ij} = 1)$ in (4.4). Additionally, we consider a G-computation estimator with predictive mean matching where true compliance is known as in (4.4). Though this estimator cannot be fit in CENIC-P2, we include it for comparison to illustrate the costs of relying on estimated probabilities of compliance. We thus have 7 estimators of $E(Y^*_{iK}(a_i, \bar{c}_{iK} = \{1, \ldots, 1\}))$. The ITT estimator is only included in CENIC-P2 for comparison:

1. The ITT estimator (only in CENIC-P2).
2. The Per Protocol estimator based off of self-reported compliance.
3. EM-REG from Boatman et. al.
4. G-computation without predictive mean matching.
5. G-computation with self-reported compliance.
6. G-computation with true compliance (only in simulation).
7. G-computation with compliance unknown and predictive mean matching (i.e., the full algorithm).

4.3.3 Simulation Results

For sample sizes $n = 500$ and $n = 1000$ across the three levels of R^2 for the confounding in the model of the study outcome, Tables (4.1–4.3) display the bias, MC SD, and MSE for each causal estimator. As R^2 increases across both sample sizes, the MSE of each estimator decreases largely due to the drop in MC SD. A similar observation can be made when the sample size increases across R^2.

Examining each estimator, the full G-computation estimator consistently returned the lowest MSE among estimators fitted to the observed data, and only had marginally

higher MSE relative to G-computation with true compliance. Without predictive mean matching, G-computation returned higher MSE, though the gap closed as R^2 increased. Relative to the other estimators, the EM-REG estimator performed best when R^2 was low at 0.3. Per Protocol and G-computation with self-reported compliance had the highest MSEs, due to having the highest biases.

Table 4.1: Simulation results with $R^2 = 0.3$ for the confounding in the model of the study outcome across time points $j \in \{0, \ldots, 5\}$. For both $n = 500$ and $n = 1000$, approximately 80% of observations across simulations had self-reported compliance and thus contributed to parameter estimation.

n	Estimator	Bias	MC SD	MSE
500	Per Protocol	1.049	0.123	1.115
	EM-REG	0.238	0.169	0.085
	G-computation without predictive mean matching	0.187	0.396	0.191
	G-computation with self-reported compliance	1.153	0.090	1.337
	G-computation with true compliance	0.028	0.100	0.011
	G-computation with compliance unknown and predictive mean matching	0.034	0.210	0.045
1000	Per Protocol	1.038	0.085	1.085
	EM-REG	0.248	0.158	0.086
	G-computation without predictive mean matching	0.140	0.313	0.117
	G-computation with self-reported compliance	1.148	0.061	1.322
	G-computation with true compliance	0.027	0.068	0.005
	G-computation with compliance unknown and predictive mean matching	0.019	0.148	0.022

Table 4.2: Simulation results with $R^2 = 0.5$ for the confounding in the model of the study outcome across time points $j \in \{0, \ldots, 5\}$. For both $n = 500$ and $n = 1000$, approximately 80% of observations across simulations had self-reported compliance and thus contributed to parameter estimation.

n	Estimator	Bias	MC SD	MSE
500	Per Protocol	1.032	0.091	1.073
	EM-REG	0.198	0.106	0.050
	G-computation without predictive mean matching	0.055	0.166	0.030
	G-computation with self-reported compliance	1.129	0.069	1.279
	G-computation with true compliance	0.023	0.065	0.005
	G-computation with compliance unknown and predictive mean matching	−0.021	0.097	0.010
1000	Per Protocol	1.037	0.065	1.079
	EM-REG	0.189	0.077	0.042
	G-computation without predictive mean matching	0.042	0.110	0.014
	G-computation with self-reported compliance	1.131	0.050	1.282
	G-computation with true compliance	0.026	0.048	0.003
	G-computation with compliance unknown and predictive mean matching	−0.022	0.069	0.005

Table 4.3: Simulation results with $R^2 = 0.7$ for the confounding in the model of the study outcome across time points $j \in \{0, \ldots, 5\}$. For both $n = 500$ and $n = 1000$, approximately 80% of observations across simulations had self-reported compliance and thus contributed to parameter estimation.

n	Estimator	Bias	MC SD	MSE
500	Per Protocol	1.036	0.075	1.079
	EM-REG	0.177	0.053	0.034
	G-computation without predictive mean matching	0.030	0.069	0.006
	G-computation with self-reported compliance	1.125	0.056	1.268
	G-computation with true compliance	0.028	0.051	0.003
	G-computation with compliance unknown and predictive mean matching	−0.030	0.059	0.004
1000	Per Protocol	1.031	0.053	1.066
	EM-REG	0.177	0.037	0.033
	G-computation without predictive mean matching	0.028	0.048	0.003
	G-computation with self-reported compliance	1.123	0.043	1.264
	G-computation with true compliance	0.025	0.038	0.002
	G-computation with compliance unknown and predictive mean matching	−0.028	0.042	0.003

4.4 Application to CENIC-P2

4.4.1 Trial Design

We applied our method to the data collected in CENIC-P2, which sought to evaluate the effect of VLNC cigarettes on average number of cigarettes smoked per day and other

measures of smoking behaviors.[48] Current smokers who had no intention of quitting were randomized to one of three treatment arms: VLNC cigarettes only, NNC cigarettes only, and gradual reduction from NNC cigarettes to VLNC cigarettes. During the follow-up period of 20 weeks with study visits every four weeks, participants were asked to only smoke the study cigarettes provided by the trial but could additionally smoke commercial cigarettes.

We are interested in estimating the causal effect of VLNC versus NNC cigarettes (where a_i = VLNC and a_i = NNC, respectively) on average number of cigarettes smoked per day (Y_{ij}) at the end of the trial. As the NNC cigarettes have the same nicotine content as commercial cigarettes and because nicotine is thought to be primarily responsible for smoking behavior, we assumed that compliance in the NNC arm was 100%. For time points $j \in \{0, \ldots, 5\}$ corresponding to study visits at weeks $\{0, 4, \ldots, 20\}$, we are interested in estimating the causal effect for VLNC cigarettes at week 20: $E(Y^*_{i5}(a_i = \text{VLNC}, \bar{c}_{i5} = \{1, \ldots, 1\})$.

Given that the VLNC study cigarettes were not commercially available, and that CENIC-P2 enrolled current smokers who had no intention of quitting, participants that missed visits were likely smoking non-study cigarettes with normal nicotine content. Thus, we assumed that missing data indicated that participants were noncompliant.

Based on the findings of various regulatory tobacco trials,[34,48,49] for the VLNC arm we assumed that missing data indicated that participants were noncompliant. As we are only interested in estimating the causal mean among compliant participants, we can exclude missing data without introducing bias in our causal estimators.

At each study visit, participants could report any use of non-study cigarettes. We assumed that self-reports of noncompliance were accurate; therefore, observations with $D_{ij} = 0$ did not contribute to parameter estimation. For the VLNC arm of CENIC-P1S2, we thus have 287 of 411 (69.8%) participants who self-reported compliance for at least one time point after baseline, where $K = 6$ time points and $j \in \{0, \ldots, 5\}$. We selected TNE as the biomarker to model, as previous research has identified TNE as

a reliable biomarker for detecting noncompliance.[59] Due to the skewness of the distribution of TNE, and the fact that it is a concentration, we modeled it on the natural logarithm scale.

We adjusted for a number of time-varying confounders for the study outcome, Y_{ij}. These included both measurements of Y_{ij} at earlier time points and the results of several surveys of smoking behavior. For the additional set of confounders Z_{ij}, we have: the Minnesota Nicotine Withdrawal Scale, the Fagerström Test for Nicotine Dependence, and the Cigarette Evaluation Scale for Satisfaction. We also adjusted for a number of time-invariant confounders. For X_i, we have: age (by year), sex (male/female), race (Caucasian/African-American/Other), and education level (less than a high school degree, high school degree only, and more than a high school degree).

Given that each of the six models (i.e., for compliance, the biomarker, the study outcome, and the results of the three surveys) we intend to fit conditions on numerous variables collected at either baseline, the current time point, or the previous time point, we performed variable selection on the six models. First, we defined an indicator of compliance based off the value of TNE, where the cut-off of 6.41 was determined from prior research. Specifically, based on the findings of a previous study of participants who were sequestered in a hotel and only had access to VLNC cigarettes, only 5% of participants randomized to VLNC cigarettes are expected to have TNE above 6.41 if they were fully compliant.[50] For each participant at each time point, we derived $C_{ij} = I(\text{TNE}_{ij} < 6.41)$. With compliance assumed to be known, we used the least absolute shrinkage and selection operator (LASSO) to identify variables to include for each model.[75,76] For the compliance model, the LASSO did not identify a coefficient for time. Therefore, unlike the simulation, we assumed the same intercept across time points for the compliance model.

Note that in contrast to the simulation, all time-varying confounders are bounded with densities skewed toward the bounds. Y_{ij} is bounded below at 0 while each survey item is bounded both above and below. We thus used predictive mean matching to generate data for all time-varying confounders.

To derive 95% confidence intervals and standard errors for each of the estimators, we

used the bootstrap.[77] For each bootstrap sample, we resampled the same number of participants as in CENIC-P2 (i.e., 411). For each resampled participant, we included their full set of observations across time points from the observed dataset. We generated 1,000 bootstrap samples, each with $R = 10,000$ samples for estimating the causal effect in the G-computation algorithm.

4.4.2 CENIC-P2 Results

Table 4.4 displays the different causal estimators for the effect of VLNC vs. NNC cigarettes on average number of cigarettes smoked per day. Relative to ITT, each causal estimator returned a smaller effect of VLNC cigarettes with higher standard error. The 95% confidence intervals for each causal estimator did not include 0. Examining each estimator, the EM-REG estimator gave the weakest effect at almost three cigarettes less than ITT. Predictive mean matching mitigated the causal effect estimate by over one cigarette per day. The full G-computation estimator returned nearly identical results to both Per Protocol and G-computation with self-reported compliance.

Table 4.4: CENIC-P2 results. For the ITT and Per Protocol estimators, the sample size comprises 297 and 195 participants, respectively. For all other causal estimators, the sample size comprises the 287 participants who self-reported compliance for at least one time point after baseline, where time points $j \in \{0, \ldots, 5\}$.

Estimator	$\hat{\mu}$(NNC)	$\hat{\mu}$(VLNC)	$\hat{\mu}$(NNC) − $\hat{\mu}$(VLNC)	SE	95% CI
ITT	21.48	13.27	8.21	0.73	(6.78, 9.60)
Causal Estimators					
Per Protocol	21.48	14.91	6.57	0.96	(4.61, 8.36)
EM-REG	21.48	16.08	5.40	0.96	(3.40, 7.17)
G-computation without predictive mean matching	21.48	13.71	7.77	0.95	(5.63, 9.42)
G-computation with self-reported compliance	21.48	14.92	6.56	0.93	(4.47, 8.31)
G-computation with compliance unknown and predictive mean matching	21.48	14.97	6.51	0.92	(4.72, 8.50)

4.5 Discussion

For longitudinal trials with unobserved compliance, causal effects estimation can be difficult due to the potentially high number of compliance patterns. Additionally, the densities of the confounders may be difficult to parametrically model across time points. We have developed a G-computation estimator that addresses both problems. First, we simplified the longitudinal modeling structure which allowed us to calculate probabilities of compliance over time. Then, we used predictive mean matching to generate random samples of the time-varying confounders that more closely matched the empirical distributions.

Across both different levels of confounding and sample sizes, our simulation study confirms that the full G-computation estimator performs better than existing causal estimators. Moreover, both the probabilities of compliance and predictive mean matching remain important to minimizing MSE. Even when there was a single confounder fit with a semiparametric model which was not the study outcome, predictive mean matching improved estimation of the causal mean. Relative to the other estimators, full G-computation provided the largest gain in efficiency when confounding was low but not insignificant.

In our application to the data collected in CENIC-P2, each causal estimator returned a significant if reduced effect of VLNC cigarettes relative to ITT. These findings confirm the original conclusion of the trial, which is that VLNC cigarettes can reduce cigarette consumption. Interestingly, the full G-computation estimator returned similar results to G-computation with self-reported compliance, implying that self-reports may have been accurate for CENIC-P2. Additionally, both estimators returned similar results to Per Protocol, implying that there may not have been much confounding for the study outcome. Predictive mean matching, however, was necessary for properly modeling the confounders.

The primary contribution of our method is that it provides a framework for modeling multiple longitudinal variables which standard regression-based methods may be ill-suited to handle. Provided that there is a biomarker strongly associated with compliance, probabilities of compliance can be inserted into the G-computation algorithm

when compliance is unknown. When randomly sampling in G-computation, predictive mean matching can loosen the parametric assumptions of the confounder densities. As demonstrated in the simulation, the causal estimator with these two modifications is asymptotically unbiased.

The limitations of our method mainly relate to the complexities of dealing with longitudinal data and numerous confounders. To simplify the longitudinal modeling structure, we assumed that observations only depended on data collected at the previous time point or baseline. This assumption may not hold when there is autocorrelation between observations across time points. In such a scenario, the score equations in (4.3) may be more appropriate. Longitudinal compliance probabilities could be substituted for the longitudinal indicator of compliance in (4.3). However, this would require the fit of a longitudinal mixture density (e.g, with mixed effects models), which would be computationally intensive with many covariates.

Additionally, while predictive mean matching can improve randomly sampling from the confounder densities, errors may accumulate with large numbers of confounders. Variable selection may also be difficult without a biomarker that has a strong association with compliance. In the application to CENIC-P2, we were able to implement the LASSO by defining a compliance indicator based off of TNE, which has an AUC of 0.99 with compliance.[59] With a biomarker that has a weaker association with compliance, the LASSO may not be as reliable.

In secondary analyses of longitudinal trials, the G-computation algorithm can estimate the causal effect for potentially large numbers of treatment combinations. When compliance and hence the treatment received are unknown, we were able to simplify the G-computation algorithm and insert probabilities of compliance to estimate the causal effect for a single compliance pattern of interest, namely, full compliance with the trial. Furthermore, the G-computation algorithm depends on correctly specifying the models of the confounders, of which there may be many. Randomly sampling from multiple misspecified models can lead to bias and variability in estimates of the causal effect. Our analysis confirms that predictive mean matching can generate confounder data that more closely mirror the true densities of the confounders. Overall, we have demonstrated that under certain assumptions, the G-computation algorithm can be

implemented when compliance is unknown. Predictive mean matching should be considered when the densities of the confounders substantially deviate from parametric densities.

Chapter 5

Conclusion

The scientific accuracy of any clinical trial depends in part on the validity of the statistical assumptions made in the trial design phase. When statistical assumptions may be violated in the trial, it is imperative to both measure the potential loss of efficiency and seek to correct for such loss in estimation of the treatment effect. Moreover, errors made in each of the three phases of clinical trials — trial design, data monitoring, and statistical analysis — can accumulate. Statistical methods are thus needed to both investigate and mitigate misspecification in each phase. This serves as the motivation for my dissertation.

For trial design, a novel ordinal endpoint was recently proposed for evaluating new treatments for patients hospitalized by influenza. Ideally, any novel endpoint should be both robust to model misspecification and more powerful than other endpoints derived from the same information. In my first case study of the ordinal endpoint, given in Chapter 2, we examined factors that may affect its statistical power under a proportional odds model for the FLU-IVIG trial. Provided that the overall treatment effect is maintained, we find that deviations from proportional odds can reduce power at most from 80% to 77.3%. Other factors including the distribution of the placebo group, misclassification, and number of categories can both reduce power as main effects and increase power when paired with treatment effects that violate proportional odds. Overall, we find that the proportional odds assumption on its own is of small concern for power, but

should be carefully scrutinized when other factors may be misspecified.

In my second case study, also given in Chapter 2, we compared the ordinal endpoint to alternative endpoints that could have been used in FLU-IVIG. In the analytic setting, we find that ordinal endpoints of high granularity can yield greater power than time-to-event (i.e., time-to-hospital discharge) endpoints when most patients in the placebo group have already been discharged by day 7. This finding held even when the follow-up period for the time-to-event endpoint was extended beyond 7 days. In the simulation setting, we find that across different placebo group distributions, only the time-to-event endpoint yielded higher power than the ordinal endpoint for treatment effects that led to early hospital discharge. Relative to time-to-event endpoints for influenza trials, ordinals endpoints may be preferable when many patients naturally progress to hospital discharge and treatment benefit manifests late in the study.

For data monitoring, in some trials participants are not obligated to comply with the treatment protocol, but biomarkers can suggest exposure to non-study treatments. In Chapter 3, I proposed and validated a novel method that can convert information from a longitudinal biomarker into various probabilities of compliance. Specifically, we assumed that the longitudinal biomarker can be modeled as a longitudinal mixture density, from which compliance probabilities can be derived. Our simulation study confirms that our method can reliably detect noncompliance across a range of differences in the biomarker between compliant and noncompliant participants. We also find that our method is robust to misspecification of the correlation in the data.

In the application to the results of the CENIC-P1S2 trial, we find that our method assigns high probabilities of compliance to participants whose biomarkers stay relatively low. In both the simulation and application, we find that our method can more accurately identify noncompliant participants than Boatman et al.'s method. More broadly, our method may be useful in future clinical trials. Specifically, the prediction probability of compliance can be used to identify noncompliant participants during the trial for remedial intervention. Additionally, the probabilities of compliance at the last time point and at all time points can be used to properly weigh the study outcomes for causal effects analyses.

For statistical analysis, the G-computation algorithm can adjust for confounding to estimate the causal effect in longitudinal trials, but typically assumes that both 1) compliance is known; and 2) that the densities of the confounders can be parametrically modeled. In the longitudinal setting when both assumptions are violated, I proposed a modified G-computation algorithm, given in Chapter 4. Across both different levels of confounding and sample sizes, our simulation study confirms that both probabilities of compliance and predictive mean matching can address both 1) and 2), respectively, to improve estimation of the causal effect. Moreover, our method performs best when there is both high confounding for the study outcome and a large sample size. In our application to CENIC-P2, our method returned similar results to the Per Protocol estimate. Interestingly, for CENIC-P2, this implies that both self-reports of compliance were accurate and that the Per Protocol estimate was relatively unconfounded by the survey data.

The statistician George Box once famously said that "all models are wrong, but some are useful". The same could be said of statistical assumptions for clinical trials in practice. No statistical assumption is perfect. As such, it remains important to both quantify the extent to which statistical assumptions may be violated and develop methods that seek to contain any errors that may arise. For both influenza trials and regulatory tobacco trials, my dissertation should help trial designers better understand how the results may deviate from what was pre-specified. Moreover, as with all fields of science, better understanding of errors should lead to better solutions, a few of which I have proposed for the aforementioned trials. With each dissertation in clinical trials comes opportunities to both anticipate and investigate the results of trials in progress and completed trials, respectively. The goal is, and always has been, the design and execution of a more perfect trial. My dissertation serves to carry the torch of that pursuit.

Chapter 6

Bibliography

[1] Ross L Peterson et al. "Analysis of an ordinal endpoint for use in evaluating treatments for severe influenza requiring hospitalization". In: *Clinical Trials* 14.3 (2017), pp. 264–276.

[2] Ross L Peterson et al. "Comparison of an ordinal endpoint to time-to-event, longitudinal, and binary endpoints for use in evaluating treatments for severe influenza requiring hospitalization". In: *Contemporary Clinical Trials Communications* 15 (2019).

[3] *National Institute of Allergy and Infectious Diseases. Evaluating the safety and efficacy of anti-influenza intravenous hyperimmune immunoglobulin (IVIG) in adults hospitalized with influenza.* `https://clinicaltrials.gov/ct2/show/NCT02287467`.

[4] *World Health Organization. Influenza.* `http://www.who.int/topics/influenza/en`.

[5] *Food and Drug Administation. Guidance for industry influenza: developing drugs for treatment and/or prophylaxis.* Report, Silver Spring, MD. 2011.

[6] Michael G Ison et al. "Intravenous peramivir for treatment of influenza in hospitalized patients". In: *Antivir Ther* 19.4 (2014), pp. 349–361.

[7] Michael G Ison et al. "A clinical trial of intravenous peramivir compared with oral oseltamivir for the treatment of seasonal influenza in hospitalized adults". In: *Antivir Ther* 18.5 (2013), pp. 651–661.

[8] Michael G Ison et al. "End points for testing influenza antiviral treatments for patients at high risk of severe and life-threatening disease". In: *The Journal of Infectious Diseases* 201.11 (2010), pp. 1654–1662.

[9] Menno D De Jong et al. "Evaluation of intravenous peramivir for treatment of influenza in hospitalized patients". In: *Clinical Infectious Diseases* 59.12 (2014), e172–e185.

[10] South East Asia Infectious Disease Clinical Research Network et al. "Effect of double dose oseltamivir on clinical and virological outcomes in children and adults admitted to hospital with severe influenza: double blind randomised controlled trial". In: *BMJ* 346 (2013), pp. 1–16.

[11] INSIGHT FLU005 IVIG Pilot Study Group et al. "INSIGHT FLU005: An Anti–Influenza Virus Hyperimmune Intravenous Immunoglobulin Pilot Study". In: *The Journal of Infectious Diseases* 213.4 (2015), pp. 574–578.

[12] James C King et al. "Clinical development of therapeutic agents for hospitalized patients with influenza: challenges and innovations". In: *Open Forum Infectious Diseases* 6.4 (2019), pp. 1–9.

[13] *National Institute of Allergy and Infectious Diseases. Comparing the efficacy and safety of high-titer versus low-titer anti-influenza immune plasma for the treatment of severe influenza A.* `https://clinicaltrials.gov/ct2/show/NCT02572817`.

[14] *National Institute of Allergy and Infectious Diseases. A study to evaluate the efficacy and safety of pimodivir in combination with the standard-of-care treatment in adolescent, adult, and elderly hospitalized participants with influenza A infection.* `https://clinicaltrials.gov/ct2/show/NCT03376321`.

[15] *Royal College of Physicians of London. National Early Warning Score (NEWS).* `https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news`.

[16] Peter McCullagh. "Regression models for ordinal data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2 (1980), pp. 109–127.

[17] John Whitehead. "Sample size calculations for ordered categorical data". In: *Statistics in Medicine* 12.24 (1993), pp. 2257–2271.

[18] Ruth Lynfield et al. "Outcomes of influenza A (H1N1) pdm09 virus infection: results from two international cohort studies". In: *PLoS One* 9.7 (2014), pp. 1–15.

[19] Dominic E Dwyer, INSIGHT Influenza Study Group, et al. "Surveillance of illness associated with pandemic (H1N1) 2009 virus infection among adults using a global clinical site network approach: the INSIGHT FLU 002 and FLU 003 studies". In: *Vaccine* 29 (2011), B56–B62.

[20] Lee-Jen Wei. "The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis". In: *Statistics in Medicine* 11.14-15 (1992), pp. 1871–1879.

[21] David R Cox. "The regression analysis of binary sequences". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2 (1958), pp. 215–232.

[22] Stuart J Pocock et al. "The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities". In: *European Heart Journal* 33.2 (2011), pp. 176–182.

[23] Duolao Wang and Stuart Pocock. "A win ratio approach to comparing continuous non-normal outcomes in clinical trials". In: *Pharmaceutical Statistics* 15.3 (2016), pp. 238–245.

[24] Alan Agresti and Maria Kateri. *Categorical data analysis*. Springer, 2011.

[25] David R Cox. "Regression models and life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202.

[26] Jong-Hyeon Jeong. "A note on asymptotic efficiency of a regression coefficient parameter under ordinal logistic regression model". In: *Communications in Statistics - Theory and Methods* 30.7 (2001), pp. 1257–1269.

[27] Ulf Stromberg. "Collapsing ordered outcome categories: a note of concern". In: *American Journal of Epidemiology* 144.4 (1996), pp. 421–424.

[28] Philip MW Bath et al. "Use of ordinal outcomes in vascular prevention trials: comparison with binary outcomes in published trials". In: *Stroke* 39.10 (2008), pp. 2817–2823.

[29] Bob Roozenbeek et al. "The added value of ordinal analysis in clinical trials: an example in traumatic brain injury". In: *Critical Care* 15.3 (2011), pp. 1–7.

[30] Gaelle Pedrono et al. “A new endpoint definition improved clinical relevance and statistical power in a vaccine trial”. In: *Journal of Clinical Epidemiology* 62.10 (2009), pp. 1054–1061.

[31] Gillian S McHugh et al. “A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project”. In: *Clinical Trials* 7.1 (2010), pp. 44–57.

[32] Barbara J Stephenson et al. “Is this patient taking the treatment as prescribed?” In: *JAMA* 269.21 (1993), pp. 2779–2781.

[33] Ramon A Boza et al. “Patient noncompliance and overcompliance Behavior patterns underlying a patient’s failure to ’follow doctor’s orders’”. In: *Postgraduate Medicine* 81.4 (1987), pp. 163–170.

[34] Eric C Donny et al. “Randomized trial of reduced-nicotine standards for cigarettes”. In: *New England Journal of Medicine* 373.14 (2015), pp. 1340–1349.

[35] Center for Drug Evaluation and Research. *Joint Meeting of Psychopharmacologic Drugs Advisory Committee and Drug Safety and Risk Management Advisory Committee*. Tech. rep. Food and Drug Administration, 2017.

[36] Center for Drug Evaluation and Research. *Joint Meeting of Anesthetic and Analgesic Drug Products Advisory Committee and Drug Safety and Risk Management Advisory Committee*. Tech. rep. Food and Drug Administration, 2018.

[37] Sandra van Dulmen et al. “Patient adherence to medical treatment: a review of reviews”. In: *BMC health services research* 7.1 (2007), p. 55.

[38] Jing Jin et al. “Factors affecting therapeutic compliance: A review from the patient’s perspective”. In: *Therapeutics and Clinical Risk Management* 4.1 (2008), pp. 269–286.

[39] Scarlett L Bellamy, Julia Y Lin, and Thomas R Ten Have. “An introduction to causal modeling in clinical trials”. In: *Clinical Trials* 4 (2007), pp. 58–73.

[40] Joseph S Koopmeiners et al. “The importance of estimating causal effects for evaluating a nicotine standard for cigarettes”. In: *Nicotine and Tobacco Research* 21.Supplement_1 (2019), S22–S25.

[41] Paul R Rosenbaum and Donald B Rubin. “The central role of the propensity score in observational studies for causal effects”. In: *Biometrika* 70.1 (1983), pp. 41–55.

[42] Jeffrey A Boatman et al. "Estimating causal effects from a randomized clinical trial when noncompliance is measured with error". In: *Biostatistics* 19.1 (2017), pp. 103–118.

[43] Jeffrey A Boatman, David M Vock, and Joseph S Koopmeiners. "Efficiency and robustness of causal effect estimators when noncompliance is measured with error". In: *Statistics in medicine* 37.28 (2018), pp. 4126–4141.

[44] Kevin C Farmer. "Methods for measuring and monitoring medication regimen adherence in clinical trials and clinical practice". In: *Clinical Therapeutics* 21.6 (1999), pp. 1074–1090.

[45] M Robin DiMatteo. "Variations in patients' adherence to medical recommendations: a quantitative review of 50 years of research". In: *Medical Care* (2004), pp. 200–209.

[46] Lawrence Joseph, Theresa W Gyorkos, and Louis Coupal. "Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard". In: *American Journal of Epidemiology* 141.3 (1995), pp. 263–272.

[47] Richard J Cook and Jerald F Lawless. *Multistate models for the analysis of life history data*. CRC Press, 2018.

[48] Dorothy K Hatsukami et al. "Effect of immediate vs gradual reduction in nicotine content of cigarettes on biomarkers of smoke exposure: a randomized clinical trial". In: *JAMA* 320.9 (2018), pp. 880–891.

[49] Tracy T Smith et al. "Randomized Trial of Low-Nicotine Cigarettes and Transdermal Nicotine". In: *American Journal of Preventive Medicine* 57.4 (2019), pp. 515–524.

[50] Rachel L Denlinger et al. "Nicotine and anatabine exposure from very low nicotine content cigarettes". In: *Tobacco Regulatory Science* 2.2 (2016), pp. 186–203.

[51] Natalie Nardone et al. "Estimations and predictors of non-compliance in switchers to reduced nicotine content cigarettes". In: *Addiction* 111.12 (2016), pp. 2208–2216.

[52] Neal L Benowitz et al. "Biochemical estimation of noncompliance with smoking of very low nicotine content cigarettes". In: *Cancer Epidemiology and Prevention Biomarkers* 24.2 (2015), pp. 331–335.

[53] Terry E Duncan, Susan C Duncan, and Lisa A Strycker. *An introduction to latent variable growth curve modeling: concepts, issues, and application*. Routledge, 2013.

[54] Gideon Schwarz et al. "Estimating the dimension of a model". In: *The annals of statistics* 6.2 (1978), pp. 461–464.

[55] MS Pepe and TA Alonzo. "Reply to letter to editor". In: *Statistics in Medicine* 20 (2001), pp. 656–660.

[56] Paul S Albert and Lori E Dodd. "A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard". In: *Biometrics* 60.2 (2004), pp. 427–435.

[57] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[58] Margaret Sullivan Pepe et al. "The receiver operating characteristic curve". In: *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press, 2003, pp. 66–85.

[59] Jeffrey A Boatman et al. "Classification Accuracy of Biomarkers of Compliance". In: *Tobacco Regulatory Science* 5.3 (2019), pp. 301–319.

[60] Linda B Von Weymarn et al. "Quantitation of the minor tobacco alkaloids nornicotine, anatabine, and anabasine in smokers' urine by high throughput liquid chromatography–mass spectrometry". In: *Chemical research in toxicology* 29.3 (2016), pp. 390–397.

[61] David M Vock et al. "Mixed model analysis of censored longitudinal data with flexible random-effects density". In: *Biostatistics* 13.1 (2011), pp. 61–73.

[62] Carina Coulacoglou and Donald H Saklofske. *Psychometrics and psychological assessment: principles and applications*. Academic Press, 2017.

[63] James Robins. "A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect". In: *Mathematical modelling* 7.9-12 (1986), pp. 1393–1512.

[64] James M Robins. "Causal inference from complex longitudinal data". In: *Latent variable modeling and applications to causality*. Springer, 1997, pp. 69–117.

[65] Miguel A Hernán and James M Robins. "Estimating causal effects from epidemiological data". In: *Journal of Epidemiology & Community Health* 60.7 (2006), pp. 578–586.

[66] Lauren E Cain and Stephen R Cole. "Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death". In: *Statistics in medicine* 28.12 (2009), pp. 1725–1738.

[67] James M Robins. "The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies". In: *Health service research methodology: a focus on AIDS* (1989), pp. 113–159.

[68] James M Robins et al. "G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients". In: *Epidemiology* (1992), pp. 319–336.

[69] James M Robins and Miguel A Hernán. "Estimation of the causal effects of time-varying exposures". In: *Longitudinal data analysis*. Chapman and Hall/CRC Press New York, NY, 2008. Chap. 23, pp. 553–599.

[70] Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. "An introduction to g methods". In: *International journal of epidemiology* 46.2 (2017), pp. 756–762.

[71] Donald B Rubin. "Statistical matching using file concatenation with adjusted weights and multiple imputations". In: *Journal of Business & Economic Statistics* 4.1 (1986), pp. 87–94.

[72] Roderick JA Little. "Missing-data adjustments in large surveys". In: *Journal of Business & Economic Statistics* 6.3 (1988), pp. 287–296.

[73] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

[74] Rebecca R Andridge and Roderick JA Little. "A review of hot deck imputation for survey non-response". In: *International statistical review* 78.1 (2010), pp. 40–64.

[75] Fadil Santosa and William W Symes. "Linear inversion of band-limited reflection seismograms". In: *SIAM Journal on Scientific and Statistical Computing* 7.4 (1986), pp. 1307–1330.

[76] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[77] B Efron. “Bootstrap Methods: Another Look at the Bootstrap”. In: *The Annals of Statistics* 7 (1977), pp. 1–26.

Appendix A

Chapter 2 Appendix

A.1 Deviations from Proportional Odds while Maintaining the Same Overall Treatment Effect

For the ith patient, assume that we have a 3-level ordinal endpoint Y_i and define:

- p_1, p_2, p_3 are the true probabilities in the first, second, and third levels of the ordinal endpoint for the placebo group, respectively.
- q_1, q_2, q_3 are the corresponding true probabilities in the treatment group.
- A_i is an indicator variable for whether or not the ith patient is randomized to the treatment group.
- Z_i, V_i are indicator variables for whether or not $Y_i = 1$ and $Y_i = 2$, respectively, for the ith patient.

Note that $p_3 = 1 - p_1 - p_2$ and $q_3 = 1 - q_1 - q_2$; therefore, the distribution of the ordinal endpoint in the placebo and treatment groups is uniquely determined by the four parameters p_1, p_2, q_1, and q_2. If we assume a proportional odds model, we can express p_1, p_2, q_1, and q_2 in terms of three model parameters. Let α_1 and α_2 represent the log odds of being in level 1 and in level 1 or level 2, respectively, for subjects randomized to the placebo group and let β represent the log odds ratio of the treatment group to the

control group. Assuming proportional odds, $\alpha_1 = \log\left(\frac{p_1}{1-p_1}\right), \alpha_2 = \log\left(\frac{p_1+p_2}{1-p_1-p_2}\right)$ and $\beta = \log\left(\frac{q_1*(1-p_1)}{p_1*(1-q_1)}\right) = \log\left(\frac{(q_1+q_2)*(1-p_1-p_2)}{(p_1+p_2)*(1-q_1-q_2)}\right)$.

Under this model, the log likelihood for $\alpha_1, \alpha_2, \beta$ is given by:

$$\begin{aligned}\log(L(\alpha_1,\alpha_2,\beta)) = &\sum_{i=1}^{n}\{(1-A_i)[Z_i \log\left(\frac{e^{\alpha_1}}{e^{\alpha_1}+1}\right) + V_i \log\left(\frac{e^{\alpha_2}-e^{\alpha_1}}{(e^{\alpha_1}+1)(e^{\alpha_2+1})}\right)\\ &+ (1-Z_i-V_i)\log\left(\frac{e^{\alpha_1}+1}{(e^{\alpha_1}+1)(e^{\alpha_2+1})}\right)]\}\\ &+ \sum_{i=1}^{n}\{A_i[Z_i \log\left(\frac{e^{\alpha_1+\beta}}{e^{\alpha_1+\beta}+1}\right) + V_i \log\left(\frac{e^{\alpha_2}-e^{\alpha_1}}{(e^{\alpha_1+\beta})(e^{\alpha_2}+e^{-\beta})}\right)\\ &+ (1-Z_i-V_i)\log\left(\frac{e^{\alpha_1}+e^{-\beta}}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2}+e^{-\beta})}\right)]\}.\end{aligned}$$

Regardless of whether or not the proportional odds model is correctly specified, we can obtain maximum likelihood estimates for α_1, α_2, and β (i.e., $\hat{\alpha_1}$, $\hat{\alpha_2}$, and $\hat{\beta}$, the values of α_1, α_2, and β which maximize the log likelihood above). If the proportional odds assumption is not correct, $\hat{\beta}$ is still an estimate of the treatment effect across all levels of the ordinal endpoint but cannot be interpreted as the constant log odds ratio of the treatment group to the placebo group across all binary divisions of the ordinal scale.

As noted in the main text, we sought to derive distributions of the treatment group that deviated from proportional odds while maintaining the same overall treatment effect specified in the design of FLU-IVIG. By overall treatment effect, we mean the average (across repeated experimentation) estimated log odds ratio for the effect of the intervention relative to placebo from fitting a proportional odds cumulative logistic regression model to the data (i.e., $E(\hat{\beta})$).

Let $\alpha_{10} = E(\hat{\alpha_1}), \alpha_{20} = E(\hat{\alpha_2})$ and $\beta_0 = E(\hat{\beta})$ represent the average estimated cumulative log odds ratio and log odds ratio. Asymptotically, α_{10}, α_{10} and β_0 are the values of $\alpha_1, \alpha_2, \beta$ for which the expected score is equal to zero. That is, for a fixed sample size, α_{10}, α_{10} and β_0 are (approximately) the values which solve the following system of equations:

1. $E\left[\frac{d}{d\alpha_1}\log(L(\alpha_1,\alpha_2,\beta))\right] = 0$;

2. $E\big[\frac{d}{d\alpha_2}\log(L(\alpha_1,\alpha_2,\beta))\big]=0$;

3. $E\big[\frac{d}{d\beta}\log(L(\alpha_1,\alpha_2,\beta))\big]=0$.

Note that:

$$\begin{aligned}
&E\big[\frac{d}{d\beta}\log(L(\alpha_1,\alpha_2,\beta))\big]\\
&=E\big\{A_i\big[Z_i\frac{e^{\alpha_1+\alpha_2+2\beta}+2e^{\alpha_2+\beta}+1}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2+\beta}+1)}+V_i\frac{1}{e^{\alpha_1+\beta}+1}-\frac{e^{\alpha_2+\beta}}{e^{\alpha_2+\beta}+1}\big]\big\}\\
&=E\big[E\big\{A_i\big[Z_i\frac{e^{\alpha_1+\alpha_2+2\beta}+2e^{\alpha_2+\beta}+1}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2+\beta}+1)}+V_i\frac{1}{e^{\alpha_1+\beta}+1}-\frac{e^{\alpha_2+\beta}}{e^{\alpha_2+\beta}+1}\big]\big\}A_i\big]\\
&=E\big\{A_i\big[E(Z_i|A_i)\frac{e^{\alpha_1+\alpha_2+2\beta}+2e^{\alpha_2+\beta}+1}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2+\beta}+1)}+E(V_i|A_i)\frac{1}{e^{\alpha_1+\beta}+1}-\frac{e^{\alpha_2+\beta}}{e^{\alpha_2+\beta}+1}\big]\big\}.
\end{aligned}$$

Note that $E(Z_i|A_i)=A_iq_1+(1-A_i)p_1$ and $E(V_i|A_i)=A_iq_2+(1-A_i)p_2$. We then have:

$$E\big[\frac{d}{d\beta}\log(L(\alpha_1,\alpha_2,\beta))\big]=E\big\{A_iq_1\frac{e^{\alpha_1+\alpha_2+2\beta}+2e^{\alpha_2+\beta}+1}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2+\beta}+1)}+A_iq_2\frac{1}{e^{\alpha_1+\beta}+1}-\frac{e^{\alpha_2+\beta}}{e^{\alpha_2+\beta}+1}\big\}.$$

Because A_i is the only random variable in the above equation with $E(A_i)=0.5$ (due to the 1:1 allocation ratio between the randomized groups), we have:

$$E\big[\frac{d}{d\beta}\log(L(\alpha_1,\alpha_2,\beta))\big]=q_1\frac{e^{\alpha_1+\alpha_2+2\beta}+2e^{\alpha_2+\beta}+1}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2+\beta}+1)}+q_2\frac{1}{e^{\alpha_1+\beta}+1}-\frac{e^{\alpha_2+\beta}}{e^{\alpha_2+\beta}+1}=0.$$

A similar analysis can be used to simplify $E\big[\frac{d}{d\alpha_1}\log(L(\alpha_1,\alpha_2,\beta))\big]$ and $E\big[\frac{d}{d\alpha_2}\log(L(\alpha_1,\alpha_2,\beta))\big]$ which will be functions of α_1,α_2,β and the true probabilities in each level of the treatment group (q_1 and q_2) and control group (p_1 and p_2).

To derive distributions of the treatment group that deviated from proportional odds while maintaining the same overall treatment effect, we can fix β (the overall treatment effect), p_1 and p_2 (the probabilities in the first two categories of the control group), and q_1 (the probability in the first category of the treatment group) to solve the system of three (nonlinear) equations for q_2,α_1 and α_2. This approach generalizes to ordinal endpoints with any number of outcome levels.

A.2 Derivation of the Cumulative Log Odds for the Distribution of the Placebo Group

For an ordinal endpoint Y, the cumulative log odds C_j for level j out of J total levels in the distribution of the placebo group is:

$$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = C_j \text{ for } j = 1, 2, \ldots, J-1.$$

To return to the cumulative probability, $P(Y \leq j)$, we can use the expit function on C_j:

$$\frac{e^{C_j}}{1+e^{C_j}} * 100 = P(Y \leq j).$$

The probability that Y assumes level j, $P(Y = j)$, is then derived as:

$$P(Y = j) = P(Y \leq j) - P(Y \leq j-1).$$

A demonstration using patients in Death or In ICU for P0:

$$\log\left(\frac{1.2+5.3}{100-(1.2+5.3)}\right) = -2.67$$
$$\frac{e^{-2.67}}{1+e^{-2.67}} * 100 = 6.5 = 1.2 + 5.3.$$

In the second analysis of the ordinal endpoint, we applied this method to the rows of the transition matrices of the FLU-IVIG placebo group to simulate the other placebo group transition matrices, as well as the IVIG transition matrices.

A.3 Misclassification Among Categories of the Ordinal Endpoint

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
M0: No misclassification	% Placebo[a]	1.2	5.3	16.2	14.4	36.4	26.5
	% IVIG[b]	0.7	3.1	10.5	10.8	36.0	39.0

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

Misclassification is added between the oxygen and discharged categories by exchanging fixed percentages of patients between the respective categories for each pair for both randomized groups. An example is on the next page for M1, which adds 20% misclassification between the non-ICU hospitalized categories and the discharged categories.

M1 Placebo Hospitalized, not in ICU, on oxygen $= 16.2 * 0.8 + 14.4 * 0.2 = 15.8$
M1 Placebo Hospitalized, not in ICU, not on oxygen $= 16.2 * 0.2 + 14.4 * 0.8 = 14.8$
M1 Placebo Discharged, not back to normal activities $= 36.4 * 0.8 + 26.5 * 0.2 = 34.4$
M1 Placebo Discharged, back to normal activities $= 36.4 * 0.2 + 26.5 * 0.8 = 28.5$
M1 IVIG Hospitalized, not in ICU, on oxygen $= 10.5 * 0.8 + 10.8 * 0.2 = 10.6$

M1 IVIG Hospitalized, not in ICU, not on oxygen $= 10.5 * 0.2 + 10.8 * 0.8 = 10.7$
M1 IVIG Discharged, not back to normal activities $= 36.0 * 0.8 + 39.0 * 0.2 = 36.6$
M1 IVIG Discharged, back to normal activities $= 36.0 * 0.2 + 39.0 * 0.8 = 38.4$
This yields the placebo and IVIG group distributions for M1:

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
M1: 20% misclassification between the non-ICU hospitalized categories and discharged categories	% Placebo[a]	1.2	5.3	15.8	14.8	34.4	28.5
	% IVIG[b]	0.7	3.1	10.6	10.7	36.6	38.4

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

A.4 Derivation of the Non-Centrality Parameter for the Accelerated Failure Time Model

For the ith patient, assume that we have an endpoint based on the event of time to hospital discharge by day 7 and define:

- T_i as the time to hospital discharge with $T_i p\delta$ and $T_i t\delta$ as the time to hospital discharge in the placebo and treatment groups, respectively.
- A_i as randomized group status with $A_i = 0$ indicating placebo group and $A_i = 1$ indicating treatment group.
- δ_p and δ_t as the number of events (i.e., discharges) in the placebo and treatment groups, respectively.
- c_p and c_t as the number of censored observations in the placebo and treatment groups, respectively.

The parametric data distribution is then:

$$T_i|A_i \sim \text{Exp}(e^{\gamma_0+\gamma_1} * e^{-e^{\gamma_0+\gamma_1 A_i T_i}})^{\delta_i}(e^{-e^{\gamma_0+\gamma_1 A_i T_i}})^{1-\delta_i}.$$

Note that the time to censoring is fixed at 7 days. Then the log likelihood is:

$$\log L(\gamma_0, \gamma_1|T_1, A_1, \ldots, T_n, A_n) = \gamma_0\delta_p - e^{\gamma_0}\sum_{i=1}^{\delta_p} T_{ip\delta} + (\gamma_0+\gamma_1)\delta_t - e^{\gamma_0+\gamma_1}\sum_{i=1}^{\delta_t} T_{it\delta} - e^{\gamma_0}7c_p - e^{\gamma_0+\gamma_1}7c_t.$$

The Hessian matrix of the log likelihood with respect to γ_0 and γ_1 is derived as:

$$\begin{pmatrix} \frac{d^2}{d\gamma_0^2} & \frac{d^2}{d\gamma_0 d\gamma_1} \\ \frac{d^2}{d\gamma_0\gamma_1} & \frac{d^2}{d\gamma_1^2} \end{pmatrix} = \begin{pmatrix} -e^{\gamma_0}\sum_{i=1}^{\delta_p} T_{ip\delta} - e^{\gamma_0+\gamma_1}\sum_{i=1}^{\delta_t} T_{it\delta} - e^{\gamma_0}7c_p - e^{\gamma_0+\gamma_1}7c_t & -e^{\gamma_0+\gamma_1}\sum_{i=1}^{\delta_t} T_{it\delta} - e^{\gamma_0+\gamma_1}7c_t \\ -e^{\gamma_0+\gamma_1}\sum_{i=1}^{\delta_t} T_{it\delta} - e^{\gamma_0+\gamma_1}7c_t & -e^{\gamma_0+\gamma_1}\sum_{i=1}^{\delta_t} T_{it\delta} - e^{\gamma_0+\gamma_1}7c_t \end{pmatrix}.$$

Focusing on the non-upper left terms of the Hessian, iterated expectation conditional on δ_t yields:

$$E(-e^{\gamma_0+\gamma_1}\sum_{i=1}^{\delta_t}T_{it\delta}-e^{\gamma_0+\gamma_1}7c_t)=E(E(-e^{\gamma_0+\gamma_1}\sum_{i=1}^{\delta_t}T_{it\delta}-e^{\gamma_0+\gamma_1}7c_t|\delta_t))=-e^{\gamma_0+\gamma_1}E(E(\sum_{i=1}^{\delta_t}T_{it\delta}|\delta_t))-e^{\gamma_0+\gamma_1}7E(c_t)$$
$$=-e^{\gamma_0+\gamma_1}E(\delta_tE(T_{it\delta}))-e^{\gamma_0+\gamma_1}7E(c_t)=\frac{-e^{\gamma_0+\gamma_1}np_1E(T_{it\delta})-e^{\gamma_0+\gamma_1}7n(1-p_1)}{2}.$$

where we have additionally made use of the fact that $\delta_t = n - c_t$, thus $E(\delta_t) = n - E(c_t)$. A similar argument can be used to simplify the upper left term of the Hessian.

The negative expected value of the Hessian is then:

$$-E\begin{pmatrix}\frac{d^2}{d\gamma_0^2} & \frac{d^2}{d\gamma_0 d\gamma_1}\\ \frac{d^2}{d\gamma_0\gamma_1} & \frac{d^2}{d\gamma_1^2}\end{pmatrix}=\begin{pmatrix}\frac{e^{\gamma_0}np_0E(T_{ip\delta})+e^{\gamma_0+\gamma_1}np_1E(T_{it\delta})+e^{\gamma_0}7n(1-p_0)+e^{\gamma_0+\gamma_1}7n(1-p_1)}{2} & \frac{e^{\gamma_0+\gamma_1}np_1E(T_{it\delta})e^{\gamma_0+\gamma_1}7n(1-p_1)+e^{\gamma_0+\gamma_1}}{2}\\ \frac{e^{\gamma_0+\gamma_1}np_1E(T_{it\delta})e^{\gamma_0+\gamma_1}7n(1-p_1)}{2} & \frac{e^{\gamma_0+\gamma_1}np_1E(T_{it\delta})e^{\gamma_0+\gamma_1}7n(1-p_1)}{2}\end{pmatrix}$$
$$=\frac{ne^{\gamma_0}}{2}\begin{pmatrix}p_0E(T_{ip\delta})+e^{\gamma_1}p_1E(T_{it\delta})+7(1-p_0)+e^{\gamma_1}7(1-p_1) & e^{\gamma_1}p_1E(T_{it\delta})+e^{\gamma_1}7(1-p_1)\\ e^{\gamma_1}p_1E(T_{it\delta})+e^{\gamma_1}7(1-p_1) & e^{\gamma_1}p_1E(T_{it\delta})+e^{\gamma_1}7(1-p_1)\end{pmatrix}.$$

For the analytic calculation of power, we are only interested in the lower right element of the inverse of the Fisher's information, derived as:

$$\frac{2}{ne^{\gamma_0}}\left(\frac{p_0E(T_{ip\delta})+e^{\gamma_1}p_1E(T_{it\delta})+7(1-p_0)+e^{\gamma_1}7(1-p_1)}{e^{\gamma_1}(E(T_{ip\delta})p_0-7p_0+7)(E(T_{ip\delta})p_1-7p_1+7)}\right)=\frac{2}{ne^{\gamma_0}}\left(\frac{1}{E(T_{ip\delta})p_0-7p_0+7}+\frac{1}{e^{\gamma_1}(E(T_{it\delta})p_1-7p_1+7)}\right).$$

Note that:

$$E(T_{ip\delta})=E(T_i|T_i<7,A_i=0)=\frac{E(T_i|A_i=0)-E(T_i|T_i>7,A_i=0)P(T_i<7|A_i=0)}{P(T_i<7|A_i=0)}=\frac{e^{-\gamma_0}-(e^{-\gamma_0}+7)(1-p_0)}{p_0}.$$

Similarly:

$$E(T_{it\delta})=\frac{e^{-(\gamma_0+\gamma_1)}-(e^{-(\gamma_0+\gamma_1)}+7)(1-p_1)}{p_1}.$$

Then:

$$\frac{2}{ne^{\gamma_0}}\left(\frac{1}{E(T_{ip\delta})p_0 - 7p_0 + 7} + \frac{1}{e^{\gamma_1}(E(T_{it\delta})p_1 - 7p_1 + 7)}\right) = \frac{2}{ne^{\gamma_0}}\left(\frac{e^{\gamma_0}}{p_0} + \frac{e^{\gamma_1}}{p_1}\right) = \frac{2}{n}\left(\frac{1}{p_0} + \frac{1}{p_1}\right).$$

Yielding the non-centrality parameter of:

$$c_a = \sqrt{n} * \log\left[\frac{\log(1-p_1)}{\log(1-p_0)}\right] * \left[\frac{1}{\sqrt{2}}\left(\frac{1}{p_0} + \frac{1}{p_1}\right)^{-1/2}\right].$$

A.5 In-depth Explanation of the Longitudinal Data Generating Process

We used data from the cohort study to estimate the distribution of the placebo group. Data has been updated since the start of the FLU-IVIG trial. Supplementary Tables A.7 to A.9 display the distributions of the placebo and IVIG groups at day 7 across the different treatment effects and populations for the simulation setting. We used the placebo group cohort data to construct corresponding transition matrices between categories across days (see Supplementary Tables A.13 to A.19). Our choice of transition matrices makes use of the Markov property, in which each transition probability only depends on the category of the given day.

We used these transition matrices to generate IVIG longitudinal data. Each row of transition probabilities can be uniquely determined by its cumulative log odds of being in a more versus less severe category. Adding (subtracting) a constant to the cumulative log odds shifts the transition probabilities to favor more (less) severe categories. Thus, we subtracted constants from the cumulative log odds of each row of each transition matrix from the cohort study to construct potential transition matrices for the IVIG group. We used these transition matrices to analytically derive corresponding IVIG group distributions on day 7. A method devised by Peterson et al.[1] can determine the corresponding odds ratio for the treatment effect.

Given that the FLU-IVIG trial pre-specified an odds ratio of 1.77, we found subtractive constants that yielded IVIG group distributions on day 7 with odds ratios within one one thousandth of 1.77. These subtractive constants made for the treatment effects examined in our paper. Details about the subtractive constants are provided in Supplementary Table A.20.

To vary the placebo group distribution across days, we added or subtracted a constant to the cumulative log odds of the day 0 placebo group distribution as well as the rows of each transition matrix. Adding (subtracting) a constant yields analytic placebo group distributions on day 7 with more (fewer) severe cases. See Supplementary Tables A.3 and A.21 for respective examples of how to derive the cumulative log odds and the analytic placebo group distributions.

To create corresponding treatment effects for the more and less severe placebo groups, we subtracted constants from their respective transition matrices on top of what had been added or subtracted already. However, due to the skewness of the transition probabilities under P3, we were unable to find a subtractive constant for T5 that returned an odds ratio of approximately 1.77.

Note that for the FLU-IVIG placebo group, we were unable to find a subtractive constant that yielded an IVIG group distribution on day 7 that approximated proportional odds, a pre-specification of FLU-IVIG. However, our treatment effects for each placebo group only mildly deviate from proportional odds. For the FLU-IVIG placebo group, Table 2.3 displays the IVIG group distribution assuming that proportional odds holds.

Supplementary Table A.1: Power to detect a significant treatment effect.

logOR[a]	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Power (%)[b]	2.5	7.2	16.6	31.8	50.8	69.7	84.4	93.4	97.7	99.4	99.9

Power is computed at the 0.05 (2-sided) level with a sample size of 320 as a function of the log odds ratio assuming the treatment effect follows proportion odds, the distribution of the placebo group is as specified in the FLU-IVIG design, no misclassification, and the full 6-level ordinal endpoint is used.

[a]logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model assuming that proportional odds holds.

[b]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

Supplementary Table A.2: Power for detecting a significant treatment effect for all possible ways of dividing the ordinal endpoint into a binary endpoint.

Category	Death versus In ICU or better	In ICU or worse versus Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, on oxygen or worse versus Hospitalized, not in ICU, not on oxygen or better	Hospitalized, not in ICU, not on oxygen or worse versus Discharged, not back to normal activities or better	Discharged, not back to normal activities or worse versus Discharged, back to normal activities
Power (%)[a]	0.59	12.8	48.9	63.9	66.2

Power is computed at the 0.05 (2-sided) level with a log odds ratio of 0.57 and sample size of 320 assuming the treatment effect follows proportion odds, the distribution of the placebo group is as specified in the FLU-IVIG design, no misclassification, and the full 6-level ordinal endpoint is used.

[a]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

Supplementary Table A.3: Altering the distribution of the placebo group by changing its cumulative log odds.

Placebo Group		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
P0: The placebo group from the FLU-IVIG design	% Placebo[a]	1.2	5.3	16.2	14.4	36.4	26.5
	Cumulative log odds[b]	-4.41	-2.67	-1.23	-0.53	1.02	NA
P1: Less skewed placebo group distribution	% Placebo	2.0	8.3	22.4	16.6	32.7	18.0
	Cumulative log odds	-3.91	-2.17	-0.73	-0.03	1.52	NA
P2: Even less skewed placebo group distribution	% Placebo	3.2	12.7	28.5	17.1	26.7	11.7
	Cumulative log odds	-3.41	-1.67	-0.23	0.47	2.02	NA
P3: More skewed placebo group distribution	% Placebo	0.7	3.3	11.1	11.2	36.3	37.3
	Cumulative log odds	-4.91	-3.17	-1.73	-1.03	0.52	NA
P4: Even more skewed placebo group distribution	% Placebo	0.4	2.0	7.3	8.1	32.6	49.5
	Cumulative log odds	-5.41	-3.67	-2.23	-1.53	0.02	NA

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]Cumulative log odds: (natural) logarithm of the odds of the given category or more severe versus less severe.

Supplementary Table A.4: Effect of all possible interactions of the treatment effect (factor 1) and placebo group distribution (factor 2) on power.

		T1: Treatment effect constantly weakens	T2: Treatment effect limited to the hospitalization or death categories	T3: Treatment effect limited to the discharged categories	T4: Smaller treatment effect limited to the hospitalization or death categories	T5: Smaller treatment effect limited to the discharged categories
P1: Less skewed placebo group distribution	Power (%)[a]	95.8	96.7	47.2	52.8	12.4
	Avg. logOR[b]	0.76	79.2	0.38	0.42	0.16
P2: Even less skewed placebo group distribution	Power (%)	**99.6**	**99.7**	**19.6**	67.6	6.5
	Avg. logOR	**0.94**	**0.96**	**0.22**	0.48	0.09
P3: More skewed placebo group distribution	Power (%)	54.5	40.8	93.9	16.8	39.9
	Avg. logOR	0.43	0.37	0.77	0.21	0.35
P4: Even more skewed placebo group distribution	Power (%)	**33.1**	**16.8**	**97.4**	8.4	51.2
	Avg. logOR	**0.33**	**0.21**	**0.91**	0.13	0.44

Bolded interactions were chosen for mention in the text of the paper.

[a]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[b]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Supplementary Table A.5: Effect of all possible interactions of the treatment effect (factor 1) and misclassification (factor 3) on power.

		T1: Treatment effect constantly weakens	T2: Treatment effect limited to the hospitalization or death categories	T3: Treatment effect limited to the discharged categories	T4: Smaller treatment effect limited to the hospitalization or death categories	T5: Smaller treatment effect limited to the discharged categories
M1: 20% misclassification between the non-ICU hospitalized categories and discharged categories	Power (%[a] Avg. logOR[b]	**79.1** **0.57**	**86.2** **0.64**	**38.0** **0.33**	40.4 0.35	11.3 0.15
M2: 40% misclassification between the non-ICU hospitalized categories and discharged categories	Power (%) Avg. logOR	**79.6** **0.57**	**92.7** **0.71**	**7.5** **0.24**	48.0 0.39	4.4 0.49
M3: 20% misclassification between the non-ICU hospitalized categories	Power (%) Avg. logOR	76.2 0.56	76.9 0.57	79.1 0.58	33.1 0.31	23.6 0.25
M4: 20% misclassification between the discharged categories	Power (%) Avg. logOR	82.4 0.60	86.8 0.64	38.0 0.34	40.9 0.36	11.2 0.15

Bolded interactions were chosen for mention in the text of the paper.

[a]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[b]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Supplementary Table A.6: Effect of all possible interactions of the treatment effect (factor 1) and number of categories (factor 4) on power.

		T1: Treatment effect constantly weakens	T2: Treatment effect limited to the hospitalization or death categories	T3: Treatment effect limited to the discharged categories	T4: Smaller treatment effect limited to the hospitalization or death categories	T5: Smaller treatment effect limited to the discharged categories
C1: Collapse the non-ICU hospitalized categories and discharged categories	Power (%)	92.3	99.4	0.05	65.4	0.05
	Avg. logOR	0.85	1.18	0	0.57	0
C2: Collapse the non-ICU hospitalized categories	Power (%)	72.4	76.4	80.0	32.4	24.2
	Avg. logOR	0.53	0.57	0.59	0.31	0.26
C3: Collapse the discharged categories	Power (%)	**95.8**	**99.5**	0.05	**65.6**	0.05
	Avg. logOR	**0.89**	**1.16**	0	**0.57**	0
C4: Collapse the four most severe categories	Power (%)	**67.0**	**76.0**	80.5	**32.0**	25.7
	Avg. logOR	**0.50**	**0.56**	0.60	**0.31**	0.27
C5: Collapse the four most severe categories and discharged categories (binary endpoint)	Power (%)	88.9	99.4	0.05	64.5	0.05
	Avg. logOR	0.81	1.18	0	0.57	0

Bolded interactions were chosen for mention in the text of the paper.

[a]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[b]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Supplementary Table A.7: Day 7 distributions of the ordinal endpoint for the FLU-IVIG placebo group (P1) and its corresponding treatment effects.

Scenario	Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities	Analytic odds ratio
P1: FLU-IVIG placebo group (%)[a]	1.0	4.9	16.3	14.5	36.2	27.1	
T1: Constant benefit (%)	0.6	3.6	10.3	9.2	37.0	39.3	1.770
T2: Benefit only for first three days (%)	0.7	3.1	9.3	10.0	38.8	38.1	1.771
T3: Linear decrease in benefit (%)	0.6	3.3	9.5	9.6	38.7	38.2	1.770
T4: Larger benefit for hospitalized patients (%)	0.6	3.5	9.9	9.0	38.4	38.6	1.770
T5: Benefit only for In ICU and non-ICU, on oxygen patients (%)	0.7	1.9	4.0	15.4	44.7	33.3	1.769
T6: Benefit only for last three days (%)	0.7	3.9	11.6	8.8	33.3	41.6	1.770
T7: Linear increase in benefit (%)	0.7	3.8	10.9	8.9	35.2	40.5	1.771

[a](%) percentage of patients in the placebo group for the given ordinal endpoint category.

The placebo group distribution in our study has been updated from the FLU-IVIG protocol using data from the cohort study as of 2016. For reference, the category percentages in the FLU-IVIG protocol are 1.8, 3.6, 15.6, 14.1, 39.0, and 25.8% for death through discharged, back to normal activities, respectively.

Supplementary Table A.8: Day 7 distributions of the ordinal endpoint for the more severe placebo group (P2) and its corresponding treatment effects.

Scenario	Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities	Analytic odds ratio
P1: More severe placebo Group (%)[a]	2.0	9.3	28.0	19.2	27.2	14.3	
T1: Constant benefit (%)	1.2	7.3	19.6	14.4	33.6	23.9	1.769
T2: Benefit only for first three days (%)	1.5	6.5	18.8	16.2	34.7	22.4	1.770
T3: Linear decrease in benefit (%)	1.3	6.9	18.9	15.5	34.8	22.6	1.770
T4: Larger benefit for hospitalized patients (%)	1.2	7.2	19.4	14.4	34.7	23.2	1.770
T5: Benefit only for In ICU and non-ICU, on oxygen patients (%)	1.3	6.1	14.1	23.7	36.2	18.5	1.770
T6: Benefit only for last three days (%)	1.3	7.7	20.9	12.7	30.4	26.9	1.770
T7: Linear increase in benefit (%)	1.3	7.6	20.1	13.4	32.2	25.4	1.770

[a](%) percentage of patients in the placebo group for the given ordinal endpoint category.

Supplementary Table A.9: Day 7 distributions of the ordinal endpoint for the less severe placebo group (P3) and its corresponding treatment effects.

Scenario	Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities	Analytic odds ratio
P1: Less severe placebo Group (%)	0.5	2.2	7.1	7.7	37.4	45.0	
T1: Constant benefit (%)	0.3	1.5	3.9	4.0	31.7	58.6	1.770
T2: Benefit only for first three days (%)	0.3	1.2	3.3	4.1	33.1	58.0	1.770
T3: Linear decrease in benefit (%)	0.3	1.3	3.4	4.0	33.1	58.0	1.769
T4: Larger benefit for hospitalized patients (%)	0.3	1.4	3.6	3.7	32.9	58.1	1.770
T6: Benefit only for last three days (%)	0.3	1.7	4.8	4.4	29.2	59.6	1.770
T7: Linear increase in benefit (%)	0.3	1.6	4.4	4.2	30.4	59.1	1.770

[a](%) percentage of patients in the placebo group for the given ordinal endpoint category.

Due to the skewness of the P3 placebo group distribution, a treatment effect corresponding to T5 that approximated an odds ratio of 1.77 on day 7 could not be found.

Supplementary Table A.10: Average coefficient values across the 10,000 trials for the FLU-IVIG placebo group on day 7 of follow-up (p_{07}[a]$= 0.63$).[b]

Model	T1: Constant benefit[c]	T2: Benefit only for first three days	T3: Linear decrease in benefit	T4: Larger benefit for hospitalized patients	T5: Benefit only for ICU and non-ICU, on oxygen patients	T6: Benefit only for last three days	T7: Linear increase in benefit
No adjustment for baseline status							
Simple Logistic	1.88	1.93	1.94	1.96	2.08	1.74	1.82
Sliding Dichotomy	1.98	2.14	2.09	2.07	2.68	1.81	1.88
Win Ratio	1.64	1.64	1.64	1.65	1.65	1.64	1.65
Proportional Odds	1.77	1.77	1.77	1.78	1.78	1.78	1.78
Longitudinal Ordinal Outcome	1.10	1.08	1.09	1.10	1.09	1.09	1.10
Cox Proportional Hazards	1.43	1.65	1.57	1.47	1.51	1.24	1.32
Accelerated Failure Time (Exponential)	0.32	0.45	0.41	0.34	0.36	0.19	0.24
Accelerated Failure Time (Weibull)	0.24	0.37	0.32	0.26	0.27	0.13	0.17
Adjustment for baseline status							
Simple Logistic	2.12	2.18	2.20	2.23	2.24	1.95	2.05
Sliding Dichotomy	2.08	2.24	2.20	2.18	2.74	1.90	1.98
Win Ratio[d]	1.78	1.76	1.76	1.78	1.63	1.81	1.80
Proportional Odds	1.91	1.91	1.91	1.91	1.93	1.92	1.91
Longitudinal Ordinal Outcome	1.11	1.05	1.07	1.11	1.09	1.13	1.13
Cox Proportional Hazards	1.49	1.77	1.67	1.53	1.47	1.27	1.35
Accelerated Failure Time (Exponential)	0.34	0.49	0.44	0.36	0.34	0.20	0.25
Accelerated Failure Time (Weibull)	0.25	0.39	0.34	0.27	0.24	0.13	0.17

See Table 2.1 for how to interpret each coefficient.

[a]p_{07} denotes the probability of discharge (i.e., categories 5 and 6 of the ordinal endpoint combined) by day 7 of follow-up for the placebo group.

[b]Supplementary Table 1 displays the category percentages for the FLU-IVIG placebo group on day 7 and each of the seven IVIG groups.

[c]Benefit refers to the differences in transition probabilities between randomized groups across treatment effects.

[d]Unlike all other models which include a covariate to adjust for baseline status, the win ratio stratifies by only comparing patients who started at the same baseline category between randomized groups.

Supplementary Table A.11: Average coefficient values across the 10,000 trials for the more severe placebo group on day 7 of follow-up (p_{07}[a]$= 0.42$).[b]

Model	T1: Constant benefit[c]	T2: Benefit only for first three days	T3: Linear decrease in benefit	T4: Larger benefit for hospitalized patients	T5: Benefit only for ICU and non-ICU, on oxygen patients	T6: Benefit only for last three days	T7: Linear increase in benefit
No adjustment for baseline status							
Simple Logistic	1.91	1.88	1.90	1.94	1.70	1.89	1.92
Sliding Dichotomy	1.96	2.05	2.01	2.00	2.16	1.91	1.92
Win Ratio	1.61	1.61	1.61	1.61	1.61	1.61	1.61
Proportional Odds	1.77	1.78	1.77	1.78	1.77	1.78	1.78
Longitudinal Ordinal Outcome	1.11	1.10	1.10	1.11	1.10	1.10	1.11
Cox Proportional Hazards	1.58	1.71	1.66	1.60	1.48	1.44	1.50
Accelerated Failure Time (Exponential)	0.43	0.51	0.48	0.44	0.37	0.34	0.38
Accelerated Failure Time (Weibull)	0.30	0.41	0.37	0.31	0.26	0.21	0.24
Adjustment for baseline status							
Simple Logistic	2.16	2.11	2.14	2.20	1.81	2.15	2.18
Sliding Dichotomy	2.05	2.14	2.10	2.09	2.22	2.00	2.02
Win Ratio[d]	1.85	1.84	1.84	1.85	1.82	1.86	1.86
Proportional Odds	1.96	1.97	1.96	1.97	1.95	1.99	1.97
Longitudinal Ordinal Outcome	1.12	1.07	1.10	1.12	1.12	1.14	1.14
Cox Proportional Hazards	1.65	1.83	1.76	1.68	1.46	1.48	1.55
Accelerated Failure Time (Exponential)	0.45	0.56	0.52	0.47	0.36	0.34	0.39
Accelerated Failure Time (Weibull)	0.31	0.43	0.39	0.32	0.24	0.21	0.24

See Table 2.1 for how to interpret each coefficient.

[a]p_{07} denotes the probability of discharge (i.e., categories 5 and 6 of the ordinal endpoint combined) by day 7 of follow-up for the placebo group.

[b]Supplementary Table 1 displays the category percentages for the FLU-IVIG placebo group on day 7 and each of the seven IVIG groups.

[c]Benefit refers to the differences in transition probabilities between randomized groups across treatment effects.

[d]Unlike all other models which include a covariate to adjust for baseline status, the win ratio stratifies by only comparing patients who started at the same baseline category between randomized groups.

Supplementary Table A.12: Average coefficient values across the 10,000 trials for the less severe placebo group on day 7 of follow-up (p_{07}[a]$= 0.82$).[b]

Model	T1: Constant benefit[c]	T2: Benefit only for first three days	T3: Linear decrease in benefit	T4: Larger benefit for hospitalized patients	T6: Benefit only for last three days	T7: Linear increase in benefit
No adjustment for baseline status						
Simple Logistic	2.02	2.22	2.23	2.20	1.71	1.85
Sliding Dichotomy	2.20	2.52	2.47	2.41	1.83	1.98
Win Ratio	1.71	1.72	1.72	1.72	1.71	1.71
Proportional Odds	1.78	1.78	1.78	1.78	1.78	1.78
Longitudinal Ordinal Outcome	1.08	1.07	1.07	1.08	1.08	1.09
Cox Proportional Hazards	1.35	1.67	1.56	1.40	1.12	1.20
Accelerated Failure Time (Exponential)	0.25	0.43	0.37	0.28	0.10	0.15
Accelerated Failure Time (Weibull)	0.21	0.36	0.31	0.23	0.07	0.12
Adjustment for baseline status						
Simple Logistic	2.29	2.52	2.55	2.52	1.88	2.07
Sliding Dichotomy	2.32	2.67	2.62	2.55	1.91	2.08
Win Ratio[d]	1.79	1.77	1.77	1.77	1.83	1.81
Proportional Odds	1.87	1.89	1.88	1.87	1.88	1.87
Longitudinal Ordinal Outcome	1.09	1.01	1.04	1.08	1.12	1.12
Cox Proportional Hazards	1.40	1.79	1.65	1.46	1.13	1.22
Accelerated Failure Time (Exponential)	0.26	0.45	0.39	0.29	0.10	0.15
Accelerated Failure Time (Weibull)	0.21	0.38	0.33	0.24	0.07	0.12

See Table 2.1 for how to interpret each coefficient.

[a]p_{07} denotes the probability of discharge (i.e., categories 5 and 6 of the ordinal endpoint combined) by day 7 of follow-up for the placebo group.

[b]Supplementary Table 1 displays the category percentages for the FLU-IVIG placebo group on day 7 and each of the seven IVIG groups.

[c]Benefit refers to the differences in transition probabilities between randomized groups across treatment effects.

[d]Unlike all other models which include a covariate to adjust for baseline status, the win ratio stratifies by only comparing patients who started at the same baseline category between randomized groups.

Due to the skewness of the P3 placebo group distribution, a treatment effect corresponding to T5 that approximated an odds ratio of 1.77 on day 7 could not be found.

Supplementary Table A.13: FLU-IVIG placebo group transition matrix from day 0 to day 1.

	Category on day 1					
Category on day 0	Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
In ICU	0.032	0.839	0.065	0.065	0	0
Hospitalized, not in ICU, on oxygen	0	0	0.872	0.075	0.053	0
Hospitalized, not in ICU, not on oxygen	0	0	0.005	0.846	0.122	0.027

Each cell presents the probability of transitioning from the given row category on day 0 to the column category on day 1.

Supplementary Table A.14: FLU-IVIG placebo group transition matrix from day 1 to day 2.

	Category on day 2					
Category on day 1	Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
Death	1	0	0	0	0	0
In ICU	0	0.962	0.038	0	0	0
Hospitalized, not in ICU, on oxygen	0	0.006	0.807	0.096	0.084	0.006
Hospitalized, not in ICU, not on oxygen	0	0.006	0.011	0.783	0.183	0.017
Discharged, not back to normal activities	0	0	0	0.030	0.879	0.091
Discharged, back to normal activities	0	0	0	0	0	1

Each cell presents the probability of transitioning from the given row category on day 1 to the column category on day 2.

Supplementary Table A.15: FLU-IVIG placebo group transition matrix from day 2 to day 3.

	Category on day 3					
Category on day 2	Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
Death	1	0	0	0	0	0
In ICU	0	0.889	0.074	0.037	0	0
Hospitalized, not in ICU, on oxygen	0	0.007	0.839	0.117	0.29	0.007
Hospitalized, not in ICU, not on oxygen	0	0.013	0	0.760	0.130	0.097
Discharged, not back to normal activities	0	0	0	0	0.933	0.067
Discharged, back to normal activities	0	0	0	0	0	1

Each cell presents the probability of transitioning from the given row category on day 2 to the column category on day 3.

Supplementary Table A.16: FLU-IVIG placebo group transition matrix from day 3 to day 4.

	Category on day 4					
Category on day 3	Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
Death	1	0	0	0	0	0
In ICU	0	0.926	0	0.074	0	0
Hospitalized, not in ICU, on oxygen	0	0.009	0.726	0.111	0.137	0.017
Hospitalized, not in ICU, not on oxygen	0.007	0	0.007	0.694	0.187	0.104
Discharged, not back to normal activities	0	0	0	0.011	0.936	0.053
Discharged, back to normal activities	0	0	0	0	0	1

Each cell presents the probability of transitioning from the given row category on day 3 to the column category on day 4.

Supplementary Table A.17: FLU-IVIG placebo group transition matrix from day 4 to day 5.

	Category on day 5					
Category on day 4	Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
Death	1	0	0	0	0	0
In ICU	0	0.923	0.077	0	0	0
Hospitalized, not in ICU, on oxygen	0	0	0.884	0.035	0.070	0.012
Hospitalized, not in ICU, not on oxygen	0	0	0	0.761	0.119	0.119
Discharged, not back to normal activities	0	0	0	0.016	0.899	0.085
Discharged, back to normal activities	0	0	0	0.019	0	0.981

Each cell presents the probability of transitioning from the given row category on day 4 to the column category on day 5.

Supplementary Table A.18: FLU-IVIG placebo group transition matrix from day 5 to day 6.

	Category on day 6					
Category on day 5	Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
Death	1	0	0	0	0	0
In ICU	0.042	0.833	0.083	0.042	0	0
Hospitalized, not in ICU, on oxygen	0	0	0.885	0.090	0.013	0.013
Hospitalized, not in ICU, not on oxygen	0	0	0.011	0.753	0.191	0.045
Discharged, not back to normal activities	0.007	0	0	0	0.911	0.081
Discharged, back to normal activities	0	0	0	0	0	1

Each cell presents the probability of transitioning from the given row category on day 5 to the column category on day 6.

Supplementary Table A.19: FLU-IVIG placebo group transition matrix from day 6 to day 7.

	Category on day 7					
Category on day 6	Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
Death	1	0	0	0	0	0
In ICU	0	1	0	0	0	0
Hospitalized, not in ICU, on oxygen	0	0	0.917	0.056	0.014	0.014
Hospitalized, not in ICU, not on oxygen	0	0	0	0.733	0.213	0.053
Discharged, not back to normal activities	0	0	0	0	0.922	0.078
Discharged, back to normal activities	0	0	0	0	0	1

Each cell presents the probability of transitioning from the given row category on day 6 to the column category on day 7.

Supplementary Table A.20: Mathematical descriptions of treatment effects designed to simulate potential IVIG data.

Treatment effect	Descriptions of constants subtracted from the cumulative log odds of the rows of the placebo group transition matrices
T1: The treatment effect remains constant for all categories across each day of follow-up.	Subtract the same constant from each row of all transition matrices
T2: The treatment effect only benefits patients for the first three days following randomization for all categories.	Subtract the same constant from each row of the first three transition matrices
T3: The treatment effect constantly decreases with each day with no additional benefit on day 7 for all categories.	Linear decrease in subtractive constant to all rows of each transition matrix. First day subtracts a constant, C, second days subtracts 5/6 C, third day subtracts 4/6 C, etc. No constant subtracted from the last transition matrix.
T4: The treatment effect remains constant across each day of follow-up but is 33.3% more effective for hospitalized patients on the log odds ratio scale.	For all transition matrices, subtract a constant, C, from each row corresponding to hospitalized patients. Subtract 3/4 C from each row corresponding to discharged, not back to normal patients.
T5: The treatment effect only benefits patients in the "In ICU" and "non-ICU, on oxygen categories" across each day of follow-up.	Subtract the same constant from the two rows corresponding to In ICU and non-ICU, on oxygen patients for all transition matrices.
T6: The treatment effect only benefits patients for the last three days of follow-up for all categories.	Subtract the same constant from each row of the last three transition matrices.
T7: The treatment effect constantly increases with each day, with benefit starting on day 2 for all categories.	Linear increase in subtractive constant to all rows of each transition matrix. Second day subtracts 1/6 of a constant C, third day subtracts 2/6 C, fourth day subtracts 3/6 C, etc. No constant subtracted from the first transition matrix.

Each treatment effect was derived for each of P1–P3.

All treatment effects return analytic IVIG distributions on day 7 that yield analytic odds ratios within one one thousandth of 1.77, the pre-specified value of FLU-IVIG.

Due to the skewness of the P3 placebo group distribution, a treatment effect corresponding to T5 that approximated an odds ratio of 1.77 on day 7 could not be found.

Supplementary Table A.21: FLU-IVIG placebo group transition matrix from day 0 to day 1 additionally including distributions on both days.

	Category on day 1						
Category on day 0	Death	ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities	Day 0 (%)[a]
ICU	0.032	0.839	0.065	0.065	0	0	7.6
Hospitalized, not in ICU, on oxygen	0	0	0.872	0.075	0.053	0	46.1
Hospitalized, not in ICU, not on oxygen	0	0	0.005	0.846	0.122	0.027	46.3
Day 1 (%)	0.2	6.4	40.9	43.1	8.1	1.2	

[a](%) percentage of patients in the placebo group for the given ordinal endpoint category.

Deriving analytic distributions of the ordinal endpoint for the placebo and IVIG groups

Given the distribution of the ordinal endpoint on one day, we derived the analytic distribution of the ordinal endpoint on the next day by use of the corresponding transition matrix. By the law of total probability, each next day category percentage can be calculated as the sum of the current day category percentages times the corresponding transition probabilities. For example, for the FLU-IVIG placebo group moving from day 0 to day 1:

- Day 1 Death: $7.6 * 0.032 = 0.2$
- Day 1 ICU: $7.6 * 0.839 = 6.4$
- Day 1 Hospitalized, not in ICU, on oxygen: $7.6 * 0.065 + 46.1 * 0.872 + 46.3 * 0.005 = 40.9$
- Day 1 Hospitalized, not in ICU, not on oxygen: $7.6 * 0.065 + 46.1 * 0.075 + 0.463 * 0.846 = 43.1$
- Day 1 Discharged, not back to normal activities: $46.1 * 0.053 + 46.3 * 0.122 = 8.1$
- Day 1 Discharged, back to normal activities: $46.3 * 0.027 = 1.2$

We cyclically applied this method to derive the analytic distribution of the ordinal endpoint on each following day. We also used this method to derive analytic distributions for the other placebo groups as well as the IVIG groups.

Appendix B

Chapter 3 Appendix

B.1 EM Algorithm Step Updates for $\boldsymbol{\theta}$

We delineate the EM algorithm step updates to derive the maximum likelihood estimates of the vector of parameters, $\boldsymbol{\theta} = (\beta_0, \gamma^2, \alpha_0, \alpha_1, \tau^2, \sigma^2)^T$, for the mixture density below:

$$f(B_{i1}, \ldots, B_{iK}) = \sum_{k=1}^{2^K} P(\{C_{i1}, \ldots, C_{iK}\} = c_k) f(B_{i1}, \ldots, B_{iK} | \{C_{i1}, \ldots, C_{iK}\} = c_k).$$

The subscripts denote the ith participant at the jth time point for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, K\}$. Random vectors are denoted with bold font, where the subscript denotes the time point. For example, for time point 1, the set of all biomarker values is denoted as $\boldsymbol{B_1} = (B_{11}, \ldots, B_{n1})^T$. The complete data log likelihood for $\boldsymbol{\theta}$ is:

$$\begin{aligned}
&\log(\boldsymbol{\theta} | \boldsymbol{B_1}, \ldots, \boldsymbol{B_K}, \boldsymbol{C_1}, \ldots, \boldsymbol{C_K}) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{2^K} I(\{C_{i1}, \ldots, C_{iK}\} = c_k) \log f(B_{i1}, \ldots, B_{iK} | \{C_{i1}, \ldots, C_{iK}\} = c_k) \\
&\qquad + I(\{C_{i1}, \ldots, C_{iK}\} = c_k) \log P(\{C_{i1}, \ldots, C_{iK}\} = c_k),
\end{aligned}$$

where log denotes the natural logarithm and I denotes the indicator function. The conditional expectation of the E-step is given by:

$$\begin{aligned}
&Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(v)}, \boldsymbol{B_1}, \ldots, \boldsymbol{B_K}, \boldsymbol{C_1}, \ldots, \boldsymbol{C_K}) \\
&= E_{\boldsymbol{\theta}^{(v)}}[\log(\boldsymbol{\theta}|\boldsymbol{B_1}, \ldots, \boldsymbol{B_K}, \boldsymbol{C_1}, \ldots, \boldsymbol{C_K})|\boldsymbol{B_1}, \ldots, \boldsymbol{B_K}] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{2^K} w_k^{(v)} \log f(B_{i1}, \ldots, B_{iK}|\{C_{i1}, \ldots, C_{iK}\} = c_k) \\
&\qquad + w_k^{(v)} \log P(\{C_{i1}, \ldots, C_{iK}\} = c_k),
\end{aligned}$$

where:

$$\begin{aligned}
w_k^{(v)} &= E_{\boldsymbol{\theta}^{(v)}}(I(\{C_{i1}, \ldots, C_{iK}\} = c_k)|B_{i1}, \ldots, B_{iK}) \\
&= \frac{P(\{C_{i1}, \ldots, C_{iK}\} = c_k)^{(v)} f(B_{i1}, \ldots, B_{iK}|\{C_{i1}, \ldots, C_{iK}\} = c_k)^{(v)}}{\sum_{m=1}^{2^K} P(\{C_{i1}, \ldots, C_{iK}\} = c_l)^{(v)} f(B_{i1}, \ldots, B_{iK}|\{C_{i1}, \ldots, C_{iK}\} = c_l)^{(v)}}.
\end{aligned}$$

In weight $w_k^{(v)}$, the k corresponds to compliance pattern c_k, the (v) denotes the vth iteration of the EM algorithm, and the l subscript in c_l is used to distinguish between the compliance patterns in the denominator and those in the numerator (i.e., c_k). The M-step update is provided by the fit of the probit and linear mixed effects models with weights $w_k^{(v)}$ added to the respective log likelihoods over all of the compliance patterns.

Note that both $P(\{C_{i1}, \ldots, C_{iK}\} = c_k)$ and $f(B_{i1}, \ldots, B_{iK}|\{C_{i1}, \ldots, C_{iK}\} = c_k)$ are available in closed-form. For $P(\{C_{i1}, \ldots, C_{iK}\} = c_k)$, we have:

$$P(\{C_{i1}, \ldots, C_{iK}\} = c_k) = \int_{x_{i1}}^{y_{i1}} \ldots \int_{x_{iK}}^{y_{iK}} f(s_1, \ldots, s_K) ds_1 \ldots ds_K,$$

where $x_{ij} = 0$ and $y_{ij} = \infty$ if $C_{ij} = 0$. If $C_{ij} = 1$, then $x_{ij} = -\infty$ and $y_{ij} = 0$. The function $f(s_1, \ldots, s_K)$ is defined as:

$$f(s_1, \ldots, s_K) = N_K \left(\mu = \begin{bmatrix} -\beta_0 \\ \vdots \\ -\beta_0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1+\gamma^2 & \gamma^2 & \ldots & \gamma^2 \\ \gamma^2 & 1+\gamma^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma^2 \\ \gamma^2 & \ldots & \gamma^2 & 1+\gamma^2 \end{bmatrix} \right),$$

where N_K denotes the PDF of the multivariate normal distribution for K time points with mean vector μ and covariance structure Σ, which is specified as compound symmetry. Essentially, we have defined $P(\{C_{i1}, \ldots, C_{iK}\} = c_k)$ to be the cumulative density

function of $N_K(\mu, \Sigma)$, where the integral bounds are defined by longitudinal compliance status.

For $f(B_{i1}, \ldots, B_{iK} | \{C_{i1}, \ldots, C_{iK}\} = c_k)$, we have:

$$
\begin{aligned}
& f(B_{i1}, \ldots, B_{iK} | \{C_{i1}, \ldots, C_{iK}\} = c_k) \\
& = N_k \left(\mu = \begin{bmatrix} \alpha_0 + \alpha_1 c_{i1} \\ \vdots \\ \alpha_0 + \alpha_1 c_{iK} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma^2 + \tau^2 & \tau^2 & \ldots & \tau^2 \\ \tau^2 & \sigma^2 + \tau^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \tau^2 \\ \tau^2 & \ldots & \tau^2 & \sigma^2 + \tau^2 \end{bmatrix} \right),
\end{aligned}
$$

where N_K denotes the PDF of the multivariate normal distribution for K time points with mean vector μ and covariance structure Σ, which is specified as compound symmetry.

B.2 Misspecification of the Covariance Structure

When the covariance structure Σ is misspecified for both compliance and the biomarker, we assume that both random variables are generated from models with AR(1) covariance structures. Models with compound symmetry are erroneously fit. Given AR(1) for $P(\{C_{i1}, \ldots, C_{iK}\} = c_k)$, $f(s_1, \ldots, s_K)$ is defined as:

$$
f(s_1, \ldots, s_K) = N_K \left(\mu = \begin{bmatrix} -\beta_0 \\ \vdots \\ -\beta_0 \end{bmatrix}, \quad \Sigma = \lambda^2 \begin{bmatrix} 1 & \rho & \ldots & \rho^{K-1} \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho^{K-1} & \ldots & \rho & 1 \end{bmatrix} \right),
$$

where λ^2 is the variance parameter and ρ is the autocovariance parameter. Note that in the simulation, $\lambda^2 = 1 + \gamma^2$; thus, the diagonal variances are set equal to those of compound symmetry.

Given AR(1) for $f(B_{i1}, \ldots, B_{iK} | \{C_{i1}, \ldots, C_{iK}\} = c_k)$, we have:

$$
\begin{aligned}
& f(B_{i1}, \ldots, B_{iK} | \{C_{i1}, \ldots, C_{iK}\} = c_k) \\
& = N_k \left(\mu = \begin{bmatrix} \alpha_0 + \alpha_1 c_{i1} \\ \vdots \\ \alpha_0 + \alpha_1 c_{iK} \end{bmatrix}, \quad \Sigma = \xi^2 \begin{bmatrix} 1 & \rho & \ldots & \rho^{K-1} \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho^{K-1} & \ldots & \rho & 1 \end{bmatrix} \right),
\end{aligned}
$$

where ξ^2 is the variance parameter and ρ is the autocovariance parameter. Note that in the simulation, $\xi^2 = \sigma^2 + \tau^2$; thus, the diagonal variances are set equal to those of compound symmetry.

Supplementary Table A.1: Estimated bias and Monte Carlo standard deviation (MC SD) of parameter estimates for $\boldsymbol{\theta}$ where $n = 500$, the number of time points $K \in \{6, 5\}$, and the compliance effect $\alpha_1 \in \{-1.2, -1.5, -1.8, -2.1\}$. For both compliance and the conditional biomarker, the models fit are either correctly specified with compound symmetric covariance structures, or incorrectly specified where compound symmetric covariance structures are erroneously fit to data that is generated as AR(1).

		Compliance (C_{ij}) Parameters				Biomarker (B_{ij}) Parameters							
		Intercept		Random Intercept SD		Intercept		Compliance Effect		Random Intercept SD		Residual SD	
	Truth	$\beta_0 = 0$		$\gamma = 1$		$\alpha_0 = 4.0$		α_1		$\tau = 0.84$[a]		$\sigma = 0.55$[b]	
		$\alpha_1 = -1.2$											
$K = 6$	Bias[c]	0.01	0.01	0.01	-	0	0.14	0	−0.28	0	-	0	-
	MC SD[d]	0.20	0.52	0.32	-	0.07	0.14	0.06	0.07	0.04	-	0.02	-
$K = 5$	Bias	0.01	0.02	0.02	-	0	0.12	0	−0.23	0	-	0	-
	MC SD	0.41	0.57	0.35	-	0.10	0.13	0.08	0.07	0.04	-	0.03	-
		$\alpha_1 = -1.5$											
$K = 6$	Bias[c]	0.01	0.01	0	-	0	0.08	0	−0.16	0	-	0	-
	MC SD[d]	0.11	0.21	0.11	-	0.06	0.08	0.04	0.05	0.04	-	0.02	-
$K = 5$	Bias	0	0.01	0	-	0	0.07	0	−0.13	0	-	0	-
	MC SD	0.12	0.22	0.13	-	0.06	0.08	0.05	0.05	0.04	-	0.02	-
		$\alpha_1 = -1.8$											
$K = 6$	Bias	0.01	0.01	0	-	0	0.05	0	−0.10	0	-	0	-
	MC SD	0.09	0.13	0.08	-	0.05	0.07	0.03	0.05	0.03	-	0.01	-
$K = 5$	Bias	0	−0.01	0	-	0	0.03	0	−0.07	0	-	0	-
	MC SD	0.09	0.14	0.09	-	0.06	0.07	0.04	0.05	0.04	-	0.01	-
		$\alpha_1 = -2.1$											
$K = 6$	Bias	0	0.01	0	-	0	0.03	0	−0.06	0	-	0	-
	MC SD	0.07	0.10	0.07	-	0.05	0.06	0.03	0.04	0.03	-	0.01	-
$K = 5$	Bias	0	−0.01	0	-	0	0.02	0	−0.04	0	-	0	-
	MC SD	0.07	0.11	0.08	-	0.05	0.06	0.03	0.05	0.03	-	0.01	-

[a] $\tau^2 = 0.7$.

[b] $\sigma^2 = 0.3$.

[c] The left (right) column of each parameter is the bias under the model with the correctly (incorrectly) specified covariance structure, where applicable.

[d] The left (right) column of each parameter is the MC SD under the model with the correctly (incorrectly) specified covariance structure, where applicable.

Appendix C

Chapter 4 Appendix

C.1 Proof that the Expected Value of (4.4) is Equal to the Expected Value of (4.6)

The expected value of (4.4) is given as:

$$\sum_{j=1}^{K}\sum_{i=1}^{n} E[I(C_{ij}=1)\frac{\partial}{\partial\beta}\log f(L_{ij}|L_{ij-1},X_i,A_i=a_i,C_{ij}=1;\beta)].$$

By iterated expectation:

$$= \sum_{j=1}^{K}\sum_{i=1}^{n} E[E[I(C_{ij}=1)\frac{\partial}{\partial\beta}\log f(L_{ij}|L_{ij-1},X_i,A_i=a_i,C_{ij}=1;\beta)|L_{ij},L_{ij-1},B_{ij},X_i,D_{ij},A_i=a_i]]$$
$$= \sum_{j=1}^{K}\sum_{i=1}^{n} E[E[I(C_{ij}=1)|L_{ij},L_{ij-1},B_{ij},X_i,D_{ij},A_i=a_i]\frac{\partial}{\partial\beta}\log f(L_{ij}|L_{ij-1},X_i,A_i=a_i,C_{ij}=1;\beta)]$$
$$= \sum_{j=1}^{K}\sum_{i=1}^{n} E[P(C_{ij}=1|L_{ij},L_{ij-1},B_{ij},X_i,D_{ij},A_i=a_i)\log\frac{\partial}{\partial\beta}f(L_{ij}|L_{ij-1},X_i,A_i=a_i,C_{ij}=1;\beta)],$$

where the last expression is the expected value of (4.6).

C.2 EM Algorithm Step Updates

We delineate the EM algorithm step updates to derive the maximum likelihood estimates of the vector of parameters for the mixture density below:

$$\begin{aligned}
&\psi(b_{ij}|l_{ij}, l_{ij-1}, x_i, a_i, d_{ij}, \alpha, \xi) \\
&= \rho(l_{ij}, l_{ij-1}, x_i, a_i, d_{ij}; \alpha) g(b_{ij}|l_{ij}, l_{ij-1}, x_i, a_i, d_{ij}, c_{ij} = 1, \xi) \\
&+ (1 - \rho(l_{ij}, l_{ij-1}, x_i, a_i, d_{ij}; \alpha)) g(b_{ij}|l_{ij}, l_{ij-1}, x_i, a_i, d_{ij}, c_{ij} = 0; \xi).
\end{aligned}$$

where $\rho(l_{ij}, l_{ij-1}, x_i, a_i, d_{ij}; \alpha) = P(C_{ij} = 1 | L_{ij} = l_{ij}, L_{ij-1} = l_{ij-1}, X_i = x_i, A_i = a_i, D_{ij}; \alpha)$ indexed by parameter vector α and g is the conditional density of B_{ij} indexed by parameter vector ξ. The subscripts denote the ith participant at the jth time point for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, K\}$.

Let $\boldsymbol{\theta} = \{\alpha, \xi\}$ denote the vector of parameters. Random vectors are denoted with bold font, where the subscript denotes the time point. For example, for time point 1, the set of all biomarker values is denoted as $\boldsymbol{B_1} = (B_{11}, \ldots, B_{n1})^T$. The complete data log likelihood for $\boldsymbol{\theta}$ is:

$$\begin{aligned}
&\log(\boldsymbol{\theta}|\boldsymbol{B_1}, \ldots, \boldsymbol{B_K}, \boldsymbol{C_1}, \ldots, \boldsymbol{C_K}, \boldsymbol{L_0}, \ldots, \boldsymbol{L_K}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{D_1}, \ldots, \boldsymbol{D_K}) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K} I(C_{ij} = 1) \log g(b_{ij}|L_{ij}, L_{ij-1}, X_i, A_i, D_{ij}, C_{ij} = 1; \xi) + (1 - I(C_{ij} = 1)) \log g(b_{ij}|L_{ij}, L_{ij-1}, X_i, A_i, D_{ij}, C_{ij} = 0; \xi) \\
&\qquad + I(C_{ij} = 1) \log \rho(L_{ij}, L_{ij-1}, X_i, A_i, D_{ij}; \alpha) + (1 - I(C_{ij} = 1)) \log(1 - \rho(L_{ij}, L_{ij-1}, X_i, A_i, D_{ij}; \alpha)),
\end{aligned}$$

where log denotes the natural logarithm and I denotes the indicator function. The conditional expectation of the E-step

is given by:

$$
\begin{aligned}
&Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(v)}, \boldsymbol{B_1}, \ldots, \boldsymbol{B_K}, \boldsymbol{C_1}, \ldots, \boldsymbol{C_K}, \boldsymbol{L_0}, \ldots, \boldsymbol{L_K}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{D_1}, \ldots, \boldsymbol{D_K}) \\
&= E_{\boldsymbol{\theta}^{(v)}}[\log(\boldsymbol{\theta}|\boldsymbol{B_1}, \ldots, \boldsymbol{B_K}, \boldsymbol{C_1}, \ldots, \boldsymbol{C_K}, \boldsymbol{L_0}, \ldots, \boldsymbol{L_K}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{D_1}, \ldots, \boldsymbol{D_K})|\boldsymbol{B_1}, \ldots, \boldsymbol{B_K}] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{K} w_{ij}^{(v)} \log g(b_{ij}|L_{ij}, L_{ij-1}, X_i, A_i, D_{ij}, C_{ij}=1;\xi) + (1-w_{ij}^{(v)}) \log g(b_{ij}|L_{ij}, L_{ij-1}, X_i, A_i, D_{ij}, C_{ij}=0;\xi) \\
&\qquad + w_{ij}^{(v)} \log \rho(L_{ij}, L_{ij-1}, X_i, A_i, D_{ij};\alpha) + (1-w_{ij}^{(v)}) \log(1-\rho(L_{ij}, L_{ij-1}, X_i, A_i, D_{ij};\alpha)),
\end{aligned}
$$

where:

$$
\begin{aligned}
&w_{ij}^{(v)} = E_{\boldsymbol{\theta}^{(v)}}(I(C_{ij}=1)|L_{ij}, L_{ij-1}, X_i, A_i, D_{ij}) \\
&= \frac{\rho(L_{ij}, L_{ij-1}, X_i, A_i, D_{ij};\alpha^{(v)}) g(b_{ij}|L_{ij}, L_{ij-1}, X_i, A_i, D_{ij}, C_{ij}=1;\xi^{(v)})}{\rho(L_{ij}, L_{ij-1}, X_i, A_i, D_{ij};\alpha^{(v)}) g(b_{ij}|L_{ij}, L_{ij-1}, X_i, A_i, D_{ij}, C_{ij}=1;\xi^{(v)}) + (1-\rho(L_{ij}, L_{ij-1}, X_i, A_i, D_{ij};\alpha^{(v)})) g(b_{ij}|L_{ij}, L_{ij-1}, X_i, A_i, D_{ij}, C_{ij}=0;\xi^{(v)})}.
\end{aligned}
$$

In weight $w_{ij}^{(v)}$, the (v) denotes the vth iteration of the EM algorithm. The M-step update is provided by the fit of the logistic and linear regression models with weights $w_{ij}^{(v)}$ added to the respective log likelihoods.